SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of

Psychology

PhD program in **Psychology, Linguistics and Cognitive Neuroscience**

Cycle **XXXII**

Curriculum in **Mind, Brain and Behavior**

If We Have Data, Let Them Talk: The Use of Big Data and Data Mining in Psychology

Surname: **Vezzoli**

Name: **Michela**

Registration number: **727779**

Tutor: **Prof. ZOGMAISTER Cristina**

Coordinator: **Prof. PERUGINI Marco**

ANNO ACCADEMICO / ACADEMIC YEAR 2019 / 2020

TABLE OF CONTENTS

ENGLISH SUMMARY

In this dissertation, I study how Big data and data mining can be used as an ensemble of methods for understanding the psychological underpinnings of people's behavior, specifically of customer behavior. Major advances in computing technology, combined with the human tendencies to collect and store information, have brought us to the Big Data era. Big Data refers to datasets that are not only big but also high in variety and velocity. The potential for Big Data to provide value for psychology is considerable: Big Data has the potential to overcome some of the issues of psychological studies and provide a set of methods, tools, and techniques (i.e., data mining) that can positively contribute to the advancements of the field.

In Chapter II, "A guide for conducting psychological research with Big Data", I aim to provide technical knowledge to those psychologists who want to conduct Big Data projects. By taking the Knowledge Discovery from Database steps as the fil rouge, I show where it is possible to find data suitable for psychological investigations, describe the methods for preprocessing these data, enlist some techniques to analyze data, and programming languages (R and Python) through which all these steps can be realized. Throughout the chapter, I also discuss some methodological issues and highlight some related pitfalls that need to be considered when applying data mining and machine learning techniques.

This dissertation aims to implement data mining methods to extract psychological knowledge from customers' data. Thus, it is essential to know how data mining has been used in the past CRM literature, which is the fields where customer data has been extensively used to create and maintain profitable relationships throughout the customer lifecycle. In Chapter III, "The use of data mining in customer relationship management:

a review on 18 years of publications (2000 - 2018)", I provide a comprehensive literature review of the published works on the use of data mining on customer acquisition, cross-sell, customer churn, and customer win-back CRM processes. The review examines journal articles and conference proceedings published between 2000 and 2018, retrieved from three academic databases. Nearly five hundred works were selected for the analyses. The selected works were analyzed according to several dimensions, such as the type of data mining techniques used, the characteristics of works (e.g., type of study, study aims, factors that predict customer behavior). Findings of this review indicate that customer churn (i.e., customers who decide to leave the company) is the area that received the most attention in the last eighteen years and that, in the same period, data mining techniques have never been applied to customer win-back. Classification models are the most used to predict customers' behaviors and ensemble algorithms (e.g., random forest) are the most used techniques. The findings of this review provide useful guidelines to direct future research and facilitate knowledge creation on the application of data mining in CRM.

In Chapter IV, "Will they stay or will they go? The prediction and understanding of customer churn behavior through data mining techniques", I develop a churn prediction model using data mining and machine learning techniques in order to shed light on the psychological underpinnings of customers' churn behavior. To build the model, I use customers' data from an energy retailer. I build several predictive models using decision trees, random forest, and logistic regression. Finally, I address *a posteriori* psychological explanations for the predictive relationships that emerged from the model to shed some light on the psychology behind churn behavior.

Keywords:

Big Data; Data Mining; Psychology; Churn Prediction; Customer Relationship Management; Literature Review

RIASSUNTO IN ITALIANO

Lo scopo della presente tesi è indagare come i Big data e il data mining possano essere utilizzati come strumenti per comprendere i fondamenti psicologici del comportamento umano, in particolare del comportamento dei consumatori. I progressi nella tecnologia informatica, in combinazione con la tendenza umana a raccogliere informazioni, hanno portato all'era dei Big Data. Il termine Big Data si riferisce ad un insieme di dati caratterizzati da grande volume, varietà e velocità. Il valore che i Big Data potrebbero fornire alla psicologia è considerevole. Infatti, i Big Data hanno la capacità di superare alcuni limiti delle ricerche psicologiche tradizionali e forniscono un insieme di metodi e di tecniche (es. data mining) che possono contribuire positivamente ai progressi del settore.

Il capitolo II, "A guide for conducting psychological research with Big Data", mira a fornire delle conoscenze tecniche a quegli psicologi intendono condurre progetti con i Big Data. Tenendo le fasi del Knowledge Discovery from Database come *fil rouge*, descrivo dove è possibile trovare dati passibili di indagini psicologiche, i metodi di pre-elaborazione dei dati, le tecniche analitiche del data mining e i linguaggi di programmazione (R e Python) attraverso cui tutti questi passi possono essere eseguiti. Inoltre, vengono discusse alcune questioni metodologiche che devono essere prese in considerazione quando si applicano le tecniche di data mining.

Poiché questa tesi ha lo scopo di implementare metodi di data mining per estrarre conoscenze psicologiche dai dati dei clienti, è essenziale sapere come il data mining è stato utilizzato in passato nella letteratura CRM, un settore in cui i dati dei clienti sono stati ampiamente utilizzati per creare e mantenere relazioni proficue con i clienti. Il Capitolo III, "The use of data mining in customer relationship management: a literature

review on 18 years of publications (2000 - 2018)", fornisce una revisione della letteratura volta alla comprensione dell'utilizzo del data mining nei seguenti processi di CRM: acquisizione clienti, cross-selling, customer churn e customer win-back. Sono stati presi in considerazione articoli di riviste e atti di convegni pubblicati tra il 2000 e il 2018, provenienti da tre banche dati accademiche. Sono stati selezionati quasi cinquecento studi che sono stati analizzati in base a diverse dimensioni, come il tipo di tecniche di data mining utilizzate, le caratteristiche dei lavori (ad esempio, il tipo di studio, gli obiettivi dello studio, le variabili che predicono il comportamento dei clienti). I risultati indicano che il churn (cioè, l'abbandono dell'azienda da parte del cliente) è il processo che ha ricevuto maggiore attenzione negli ultimi diciotto anni e che, nello stesso periodo, le tecniche di data mining non sono mai state applicate ai processi di customer win-back. I modelli di classificazione sono quelli più utilizzati per predire il comportamento dei consumatori, e che gli algoritmi ensemble (ad esempio, Random Forest) sono le tecniche più utilizzate. I risultati di questa revisione forniscono delle linee guida utili per orientare la ricerca futura e facilitare la creazione di conoscenze sull'applicazione del data mining nel CRM.

Nel Capitolo IV, " Will they stay or will they go? The prediction and understanding of customer churn behavior through data mining techniques", è stato sviluppato un modello predittivo del comportamento di churn utilizzando alcune tecniche di data mining e di machine learning al fine di far luce sulle motivazioni psicologiche di tale comportamento. Per costruire il modello, sono stati utilizzati i dati dei clienti di un rivenditore di energia. I modelli predittivi sono stati creati utilizzando le tecniche di Alberi Decisionali, Random Forest e Regressione Logistica. Infine, le predizioni del modello maggiormente predittivo sono state interpretate a posteriori per far luce sulle motivazioni psicologiche sottostanti al comportamento di churn.

Parole Chiave:

Big Data; Data Mining; Psicologia; Customer Churn; Customer Relationship Management; Revisione della Letteratura

CHAPTER I

INTRODUCTION

1. Introduction

In this dissertation, I study how Big data and data mining can be used as an ensemble of methods for understanding the psychological underpinnings of people's behavior, specifically of customer behavior.

Data is everywhere. Since I started to write this sentence, probably millions of data of different nature have been registered and stored somewhere. A phrase that is often used in relation to this phenomenon is "Big Data". Generally, Big Data is defined as observational records that may be exceptionally numerous (Volume), heterogeneous (Variety), and generated, stored, and aggregated at a high rate (Velocity). Also, Big Data is often portrayed as the intersection of new generations of technology (e.g., computational power and pervasiveness), and breakthrough analytical techniques (e.g., data mining and machine learning) (Boyd & Crawford, 2012).

The Big Data era is providing psychologists new means to tap into psychological constructs in fresh and new ways. Indeed, a further understanding of people's psychological processes and behaviors may lie at the tips of our fingers on the keyboards of our computers or the touchscreens of our smartphones. Every time we interact with an informatic device some kind of data is registered somewhere: From a text message on our mobile phone, to an Amazon order, to the use of a loyalty card in a supermarket. Each of these data can tell something on ourselves, on our behavior and, perhaps, why we behave in the way we do. Everything we do can leave a mark of our personality, often unintentionally. These are our digital footprints, traces of behaviors we leave while making our way through cyberspace, and

they may be useful instruments to answering many of the questions that psychologists have had for centuries.

For example, one interesting topic for psychology is the relationship between personality and word use. The investigation, until recently, was limited to resources that often were artificial samples of texts produced in the laboratory on specific topics and were limited in size to no more than a few thousand words per speaker, uttered or written in a small number of occasions. These constraints had an important impact on the types of discoveries that could be made. In particular, the relatively small corpora of words per participant imposed an important limitation to analyses that could be conducted, as words had to be grouped in categories to allow reliable estimates. To overcome these limitations, Yarkoni (2010) investigated the use of language in internet blogs. Blogs are easily accessible and large writing samples on various topics of choice of the bloggers. Demand characteristics do not influence the bloggers' writing, because they are unaware that, later on, a researcher will ask them to use their materials for a study on personality. Using a corpus of writing samples from 694 blogs, containing on average 115,423 words per blog, Yarkoni replicated previous results (hence showing evidence for the validity of this strategy of analysis) and found novel results, that could not have emerged with the traditional techniques. For example, while previous research had showed that individuals with high scores in neuroticism used more negative emotion words, Yarkoni's study indicated that they specifically showed an increased used of adjectival words to describe events negatively (e.g., 'stressful', 'depressing').

As mentioned before, many human behaviors now produce digital traces and, therefore, the sources of data can be diverse: Mas and Moretti (2009) for instance, to investigate the impacts of the productivity of cash clerks in supermarkets of a large grocery chain on the productivity of co-workers, used data from the scanners at the checkers. They investigated all the transactions that took place in six stores for two years, observing a total of 370 cashiers. Interesting results emerged, which could hardly have been found with more traditional

techniques, such as surveys or interviews. They found evidence of positive productivity spillovers and could further investigate the optimal conditions to cause them (e.g., presence of low productivity and high workers together in the same shift). Given the rich dataset, they could conduct further analyses to collect information on the mechanism that generated the spillover effect: They examined the spatial arrangement to test whether workers' ability to observe the performance of the colleagues had an impact, and investigated if having previously worked together increased the effect.

With Big Data analyses psychologists can access to thousands, if not millions of pieces of people's information, allowing to operate at a scale barely imaginable only a few years ago. Yet, this is not just about sample size. Indeed, Big Data allow to get insight into people's behavior with better accuracy and detail. Overall, the use of unobtrusive methods to collect data offers the chance to detect subtle changes, to make much more accurate estimations of the frequency of events, and to obtain data points more often than with traditional forms of assessment. It allows to get insight into users' feelings and spontaneous actions in the natural context where they occur, allowing the detection of rare behaviors and characterizing processes in detail as they unfold over time. Nowadays, a researcher could investigate in great detail how we react, interact with others, and make our decisions in specific life situations. For instance, in their quest to make games as life-like as possible, the gaming industry has incidentally built the psychologist's dream. Complex behaviors and environments have been converted into digital representations, pre-populated by human gamers having virtual but still emotional and immersive experiences. Naturalistic observation, virtual laboratories, experiments, and even psychometric assessments can be carried out in a virtual environment that is ready-made, pre-distributed to millions of people who comfortably and naturally interact in a quantified reality. For example, the observation of how online players interact with the Church of the Holy Light, the World of Warcraft in-game religion (a game with 12 million of active players), may give researchers an insight into how young adults think and

relate to off-line religious concepts. In such a game, players re-enact versions of their own selves at different times in their spiritual journeys of becoming an atheist, playing a pagan character while they are themselves religious, or on the contrary, experimented with religious worldviews through priest avatars while not believing in a deity.

In sum, what Big Data represent for psychologists goes far beyond the mere access to larger datasets. Instead, it holds the key to a qualitative and quantitative revolution in psychological science. Recording the behaviors of thousands of people could arise the opportunity to validate or refine previous theoretical models. Even more excitingly, we could have the chance to detect completely new psychological and social phenomena until now invisible to the human eye. Unlike psychologists, algorithms do not need to have a preconception of how the human psyche could work. Thus, they possibly reveal patterns that no one would have thought to explore before.

In the next sections, I will explain what is meant by Big Data, what are their characteristics, and I will also discuss some of their advantages and disadvantages.

2. Big Data: Characteristics, Advantages and Disadvantages

The term Big Data has been used both in research and non-research papers for quite a long time. However, there is no universally recognized definition of Big Data. Some suffice with "Big Data are big," though leaving the limits of "big" vague. Others may believe that this nomenclature should be related to the mode of analysis: According to this view, Big Data are those data exceeding the availability of the physical memory of the computer, and therefore requiring a distributed data analysis based on architectures like Apache Hadoop or Spark. Others consider Big Data in a broader sense and define it as a nascent paradigm that connects scholars, practitioners, and policymakers from across disciplines on the basis of techniques, beliefs, and practices that underlie new types of data-intensive research, insights, and practices (Wenzel & Van Quaquebeke, 2018).

A useful perspective for conceptualizing Big Data was provided by Laney (2001). According to the author, Big Data are characterized by three essential dimensions: volume, variety, and velocity. Later, other dimensions of Big Data entered the stage (e.g., veracity, value), but most of them rely on the three dimensions identified by Laney (Ylijoki & Porras, 2016). I now describe the dimensions and outline their theoretical and practical advantages and disadvantages in psychological research.

2.1 Volume

Volume is the dimension most often associated with Big Data. It refers to the quantity of data.

The big volume of data may overcome some issues of psychological research. First of all, a large portion of psychological studies is based on small samples (Bakker, Van Dijk, & Wicherts, 2012; Fraley and Marks, 2007). The extensive reliance on small samples is far from adequate, given the typical sizes of the effects observed in psychology. Indeed, if the sizes of the samples on which our research is based are too small, the statistical power is low, and Confidence Intervals are too wide (Cohen, 1992; Maxwell, 2004). Thus, the small number of participants limits the size of effects researchers can study due to low statistical power (Cohen, 1992). For example, the average effect size of a correlation in social psychology is r = 0.21 (Richard, Bond, & Zoota, 2003). To have an 80% chance of detecting an effect with such an effect size with the standard false positive rate of 5%, researchers would need at least 194 participants in each study. However, it has been found that the median sample sizes were 80 total participants for regular articles and 52 participants for brief articles (Bertamini & Munafò, 2012). Underpowered studies are problematic because they may lead to biased conclusions (Christley, 2010; Turner et al., 2013; Kühberger et al., 2014) which negatively affect replicability and, ultimately, the trust we can put in research results (see Asendorpf et al., 2013). The reason behind these biased conclusions is that underpowered studies yield excessively wide sampling distributions for the sample estimates. This means that all

parameters computed from the sample (e.g., effect sizes) can differ considerably from the population value, and also over replications (Crutzen & Peters, 2017). This partly explains why many replication studies did not reproduce the original results (Open Science Collaboration, 2015; Peters and Crutzen, 2017). The issue of small data presents a severe limitation in psychology, especially when researchers seek to model complex relationships between multiple factors and their interactions (Murphy & Russell, 2016; Scherbaum & Ferreter, 2009). Creating robust models of multilevel phenomena requires large samples, or even multiple observations of the same event, which in aggregate outperform any single observation (Epstein, 1979; Fishbein & Ajzen, 1974).

Among the advantages of Big Data is that their use in certain circumstances may allow psychologists to completely overcome the perils of statistical inferences, by examining the full populations. For example, if we measure the height of all Italians and all Swedes, we do not need to conduct a statistical analysis for testing the hypothesis that the average height of Swedes is higher than the average height of Italians. In fact, since we have access to information about the population, we do not have to make any inference, because it is sufficient to describe the averages of the two populations. When reaching the entire population is not feasible, accessing a high volume of cases can hugely increase the sensitivity of the analyses and the accuracy of the estimates.

Another advantage that comes from Big Data is the possibility to reach populations different from WEIRD samples (i.e., Western, educated, industrialized, rich, and democratic; Henrich, Heine, & Norenzayan, 2010) on which a considerable portion of psychological studies is conducted. It may be problematic to rely on a limited population because it may entail a threat to the external validity of research findings. As stated by Henrich and colleagues (2010), WEIRD samples are not just a restricted sample of humanity; they are frequently distinct outliers vis-à-vis other global samples. Thus, they may represent the unsuitable population on which to base our understanding of humans. In this sense, the use of

internet technologies and Big Data can provide information on people of various populations. A research conducted by Gosling and colleagues (2010) demonstrated that Internet samples are not as dominated by WEIRD participants as more traditional samples. Moreover, using Big Data samples permits to collect information generated directly by the individual who behaves in a given context. Consequently, the advent of Big Data allows researchers to understand behavior at an unprecedented scale and in settings of high ecological validity.

Moreover, Big Data allows small but relevant phenomena to become the subject of quantitative analysis. It allows studying phenomena that otherwise could not be studied with small samples because of their low incidence. For example, Welles (2014) conducted a research on female online gamers. In one instance, she was interested in examining a very specific group of women over the age of 50, who were high-frequency gamers. In a database of 10 million Second Life users, only 1500 women met the selection criteria for the population of interest. Thus, the author could collect an extremely large dataset to make a detailed examination of the gaming behaviors of an extreme statistical minority with a meaningful sample size. Thus, these extensive data collections may provide a way to understand psychological constructs and processes that have been impractical, if not impossible, to study thus far. In psychological research, atypical data and individuals are often considered a nuisance, that can distort then analyses. If we have enough observations, however, they can become a target of research in themselves. In this way, atypical observations, as for instance, individuals with atypical characteristics or constellations of characteristics (which are called outliers and kept under control with various strategies in traditional research) can be transformed from a burden to an opportunity for new discoveries.

Despite the great advantages of accessing large samples, there are also some drawbacks. One of these consists in the risk of overlooking the issue of representativeness, due to overconfidence caused by the sheer size of the dataset (i.e., data hubris; Lazer, et al., 2014). Large samples may not necessarily be representative of the full population. An example from

the past, which remains relevant, is the case of the 1936 election poll conducted by the Literary Digest and by George Gallup. From a sample of 2.4 million participants, Literary Digest predicted that London would win the election. Contrary to this prediction, based on a much smaller sample of 50 thousand people Gallup predicted that Roosevelt would become the next US President. Roosevelt won the election. Although the Literary Digest sample was far larger than the one used by Gallup, the Literary Digest sample was not representative of the reference population for two reasons: the selection bias (they had reached a convenience sample of middle- and upper-class voters from telephone directories, club membership lists, and lists of magazine subscribers) and the non-response bias (only 2.4 million out of 10 million people invited to participate had responded the survey). This example illustrates how even very large datasets may be subject to sampling problems, which have the dangerous potential of causing a researcher to draw biased conclusions (McFarland & McFarland, 2015) and feel confident on these conclusions because of the mere size of the sample. If we take the Literary Digest example to the Big Data era, we might conduct research for predicting voting behavior at the next election by using data from social media (e.g., Facebook), for example the posts or the likes for different types of contents. However, using this type of data would not guarantee that our sample would represent the voters' population, nor that all users would express their positions online with the same frequency: For example, individuals characterized by more extreme political attitudes might be more active online, as compared to those with more moderate attitudes. Thus, Big Data samples may not be representative of this layer of the population. To mitigate such bias, researchers should define the key characteristics of the target population and the procedures, considering the potential confounding variables that may affect the proper acquisition of data from representative samples.

2.2. Variety

Another important characteristic of Big Data lies in the wide variety of information, the heterogeneity of data modalities that are available for examinations.

In recent years there has been an explosion in the development of systems capable of tracking people's behavior, and the variety of information we can access, so much so that the interest of psychologists has skyrocketed accordingly.

Big data are varied in relation to the source that created them. Some of the data come from automated systems, while some others are produced by users (human-generated content).

An example of data created by automated systems is system logs registering data collected through smartphones sensors. Such smartphone data can be used to capture many behaviors such as social interactions (e.g., the size of in-person social groups; Chen, et al., 2014), daily activities (e.g., partying and studying habits; Wang, Harari, Hao, Zhou, & Campbell, 2015), and mobility patterns (e.g., routines in mobility patterns; Harari, Gosling, Wang, & Campbell, 2015). Smartphones represent a feasible and unobtrusive method for collecting information on the behavior of people as they go about their daily lives. Thus, Big Data can provide psychology with behavioral traces that occurs in authentic ecological contexts. However, the unobtrusive methods pose some critical issues about people's privacy, which I discuss in detail in the "Privacy Concerns" section.

Examples of data produced by users (human-generated content) are e-mails, materials posted on social media, but also data registered, for instance, by call center operators and hospital admittance records. There are many examples of psychological researches that leveraged user-generated content. One is provided by Youyou and colleagues (2015), who used Facebook data (e.g., users Likes) from 90000 users to evaluate the ability of computer models to predict individual personality characteristics. They found that data-driven models based on online behavior were more accurate than participants' Facebook friends in predicting

self-reported personality measures and had high external validity when predicting life outcomes such as substance abuse, political attitudes, and physical health.

Big Data are varied also according to the structure they can assume. In some cases, they have the typical format that psychologists are more used to, as they can be represented in quantifiable metrics in a matrix or are structured as (relatively) easily manageable text strings. In other cases, they consist in images or videos encoded in standardized file formats. These new formats are not only more available than in the past, but also more easily treated and analyzed, thanks to the recent technological developments. By crossing the variety of sources and formats, we can categorize data along a continuum of structured, semi-structured, and unstructured data. The phrase “Structured data” describes the type of data that can be represented in a spreadsheet, are highly organized, and can be easily understood by machine language. Semi-structured data do not have the same level of organization and predictability of structured data. These data do not reside in fixed fields or records but contain elements that can provide them with categorizing properties. An example of such type of data is a digital photograph. The image itself is unstructured, but if the photo was taken on a smartphone, for example, it would be date and time stamped, geo tagged, and would have a device ID. Unstructured data are data that do not conform to typical relational data models (e.g., two-way matrix) and for which it is not possible to identify hierarchies. Examples of this category are e-mails, videos, audio files, web pages, and social media data. Nowadays, most of the data is unstructured, with some estimates of it being 85% of all data generated (De Boe, 2014).

The increase in types of sources and types of data available has greatly expanded the possibilities for psychological research. This helps overcome some of the typical methodological issues of data acquired with the traditional techniques commonly used in psychological research. In existing procedures for collecting data on behaviors, researchers typically ask to estimate the frequency or duration of past behaviors. However, these self-reporting procedures are associated with biases, such as respondent's lack of attention to

critical behaviors, memory limitations, socially desirable responding (Gosling, John, Craik, & Robins, 1998; Paulhus, & Vazire, 2007). As a consequence, psychology has a great deal of data on what people believe they do, derived from the self-reports, but little data on what people actually do, derived from direct observations (Baumaister, et al., 2007). Instead, Big Data provide a precise, unprecedented view of people's behaviors. They can tell what we have done, where we have been, with whom we have been in contact. Importantly, they do so without asking us anything and, therefore, without the potential for conscious or unconscious distortions.

Finally, Big Data provide information that can complement the one retrieved with traditional approaches. For example, the integration of self-report data with sensor data permits the researcher to supplement objective behavioral estimates with the participants' reports of their experience. Consider a researcher who is interested in how socializing behaviors vary as a function of internal states (e.g., mood or stress level) or a person's situational context (e.g., talking with a friend or being at work). In such a case, we can combine data recorded through smartphone sensors (e.g., microphones) with ecological momentary assessment reports (e.g., "Who are with you?", "What are you doing here?", "What is your mood level?") (Shiffman, Stone, & Hufford, 2008). In this way, we can compute, for example, the talking duration (obtained through the microphone) when they report being with friends, at work, relaxed or in a bad mood). By combining multiple methods, researchers may overcome the weakness or intrinsic biases and the problems that come from single-method studies.

In sum, the new and enhanced data sources and technologies provide unprecedented opportunities for researchers and practitioners to improve the richness of their data and, hence, reach a more comprehensive view of the phenomena under investigation.

2.3. Velocity

The velocity through which new data become available is the third factor that characterizes Big Data. Velocity describes the speed at which data accumulate. It is a function of the rate by which a phenomenon is quantified or sampled into a digital object and then transmitted and retrieved (Wenzel & Van Quaquebeke, 2018). Events that occurred in the real world can be converted into data at different levels of granularity: From real-time streams of data to data streams that can be constant or variable with daily, weekly, monthly, or event-triggered peak loads (Troester, 2012). The latency through which events are recorded depends on the type of the object under investigation. For example, a wearable device will record body movements and states regularly (i.e., a constant flow of real-time data). A company will register information about firm-customer interactions whenever they happen if they happen (i.e., variable flow of data).

Social and individual phenomena have a temporal nature, and the corresponding research is often constrained by limited available observations (e.g., cross-sectional data), which offer a partial representation of reality (Mitchell, James & James, 2011). Having at our disposal observations that are recorded in a discrete but temporally ordered manner can extend our understanding of what happens, when it happens, and potentially why it happens (Roe, 2008).

The IT infrastructures through which Big Data are generated can record information in potentially infinite ways and with increasingly tight intervals. This characteristic has two consequences. On the one hand, it becomes possible to study and understand in a much more varied way the direction, frequency, speed, and change of the phenomena under investigation. On the other hand, the analysis of the time series would allow verifying the existence of causal links between phenomena, which require that a specific phenomenon X temporally precedes another phenomenon Y. For example, Kramer, Guillory, and Hancock (2014) uncovered causal evidence of emotional contagion by presenting nearly 700,000 Facebook users with contents characterized by varying degrees of emotional valence over one week.

They found that exposure to emotions led people to change their posting behaviors accordingly. This represents an alternative way through which researchers can use Big Data for testing the causality of investigated phenomena. Even though being temporally ordered does not intrinsically demonstrate causality, the fact that one phenomenon precedes the other in a time series provides a stronger empirical indication for testing whether certain values of X reliably precede the phenomenon Y and that the opposite does not occur.

2.4. Veracity

Another V that is commonly used as a defining feature of Big Data is Veracity, which refers to the accuracy of the data. This feature indicates the precision of the data and the possibility to use it for analysis. The correctness of the data will determine how important this data for the problem being studied and some researchers believe this is the biggest challenge of Big Data (Hamoudy, 2014). Indeed, when the data are characterized by substantial volume, variety, or velocity, it may be particularly challenging to ensure Veracity (Toninandel, et al., 2018). A well-known example of the risks of Big Data error is provided by the Google Flu Trends series that uses Google searches on flu symptoms, remedies, and other related keywords to provide near real-time estimates of flu spread in the USA. Compared to data collected by the US Centers for Disease Control and Prevention, the Google Flu Trends provided remarkably accurate indicators of flu incidence in the USA between 2009 and 2011. However, for the 2012-2013 flu seasons, Google Flu Trends predicted more than double the proportion of doctor visits for flu-like symptoms compared to those registered by the Centers for Disease Control and Prevention (Butler, 2013). A reason for the failure may be traced back to the fact that the data-generating engine was modified in such a way that the formerly highly predictive search terms eventually failed to work. For example, when a Google user searched on "fever" or "cough," Google's other programs started recommending searches for flu symptoms and treatments, the search terms the algorithm used to predict flu. Thus, flu-related searches artificially spiked as a result of these changes to the algorithm and the impact

these changes had on user behavior. In survey research, this would be similar to the measurement biases induced by interviewers who suggest to respondents who are coughing that they might have flu, then ask the same respondents if they think they might have flu. This example shows that having a large quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability (Lazer, Kennedy, King, & Vespignani, 2014). Indeed, within the context of Big Data, key constructs may be underrepresented (or may be omitted altogether) or poorly measured, data may remain limited in their generalizability across contexts, and inferences drawn from the use of inaccurate data may lead to wrong decisions. Thus, is of crucial importance to assess the quality of Big Data, because this determines the quality of insights and conclusions that can be derived from them. In data science, this statement is simplified with the motto "garbage in/garbage out". When it comes to Big Data, this motto holds. The data collected needs to be highly accurate, otherwise the data analytics process and its results will be unavoidably unreliable.

A major concern related to data veracity is data relevance, which represents the level of consistency between the data content and the area of interest of the user. The volume and variety of Big Data available to psychology researchers clearly offer a benefit for combating traditional concerns regarding, among other things, statistical power and replicability. However, just because data are numerous and are obtained from varied sources does not mean that they adequately operationalize the constructs of interest (Braun, & Kuljanin, 2015; Whelan, & DuVernet, 2015). Indeed, it is not guaranteed that the types and quantities of data available will adequately represent the underlying mechanisms of concern (i.e., construct representativeness) or that they will exhibit strong psychometric properties (e.g., reliability). For example, a job recruiter may believe that the answers provided by a job applicant to a Facebook quiz represent an indicator of their problem solving skills. However, should the recruiter put faith in such data for informing employee selection over scores on a traditional test? Keeping in mind that this "measure" is not designed or validated for such purposes,

researchers and practitioners should exercise particular caution in drawing inferences from their use. In mining such data for deriving insights, psychologists are equipped with the tools of classic test validation procedures to help data scientists identify key, representative indicators across the depth and breadth of theoretical constructs of interest. Even once such indicators are identified, translating this data (e.g., number of “like” on Facebook) into useable formats presents challenges for drawing meaningful conclusions (Braun & Kuljanin, 2015).

Another concern related to data veracity lies in the fact that Big data analytics are often conducted using existing or ongoing data collection (i.e., use of secondary data) that have not been registered with research purposes. As such, issues regarding experimental and statistical control are of particular concern for drawing conclusions from Big Data. Without statistical and experimental controls in place, Big Data users must be cautious of drawing causal inferences from the results of Big Data analytics. This is one area in which psychology researchers, who are classically trained in experimental and quasi-experimental design, can contribute to Big Data science. Involving psychology researchers in database design, data collection, data management, and subsequent data analyses, can contribute to 1) identify relevant comparison or control groups against which to test empirical differences in effects, 2) identify potential alternative explanations and third-variables for statistical control, and 3) advocate for cautious optimism in interpreting Big Data results.

Of course, data quality is one of the most central issues in psychological research, and one of the most important questions that researchers should ask themselves is: "Are my measures of good psychometric quality?". However, the meaning of data quality, when applied to Big Data, is different as compared to traditional measures. Within the survey research tradition, the concept of survey errors was developed in the early 1940s (Deming 1944) and has since evolved into the Total Survey Error framework (Biemer 2010), which refers to the accumulation of all errors that may arise in the design, collection, processing, and analysis of

survey data. If we apply the TSE framework to Big Data, we obtain the Big Data Total Error (Japec et al. 2015). Errors in Big Data arise mostly during three steps used to create a dataset from Big Data (Biemer 2014, 2016):

1. Data generation. It is specifically the data generation process that differentiates Big Data from surveys, and experiments (Kreuter and Peng 2014). Errors may arise because no technology is perfect. The instruments that generate the data may not be accurate in doing so due to technical problems. Moreover, similarly to traditional methods, also for measures based on new technologies we may have problems of reliability. For example, device components such as microphones may differ in their sensitivity, which can produce dampened or extreme signals (Chaffin, et al., 2015). Similarly, studies also suggest that the inter-device reliability for wearable activity trackers is commonly high for normal step count. However, problems occur at slow walking speeds or intense physical activity (Evenson, Goto, & Furberg, 2015; Mantua, Gravel, & Spencer, 2016). Such incongruities may not be problematic for usual device usage, but they may reflect a substantial bias in the inter-device variability that is a systematic function of some other quality. For instance, more expensive wearables may produce significantly better data than their cheaper counterparts. Under identical conditions, then, based on the data registered from their wearable devices a wealthy worker would be considered relatively more vocal, active, or rested than a poorer employee, but this inferences may not represent reality.
2. Extract, Transform, and Loading processes. In research on Big Data, it is often necessary to combine data from different sources. For example, to reveal a stable connection between political orientation and subjective well-being, Wojcik and colleagues (Wojcik, Hovasapian, Graham, Motyl, & Ditto, 2015) analyzed texts from tweets, public speaking records, photos from LinkedIn profiles and public pictorial directories. However, combining those sources in

the same computing environment through the Extract, Transform, and Loading processes[1] may introduce distortions and inconsistencies. Thus, for providing an accurate representation of the object under investigation, we must integrate information without making errors.

3. Analysis. The analysis of Big Data introduces risks for a number of errors. For example, Big Data can be subject to sampling errors when the data are filtered, sampled, or otherwise reduced to form more manageable or representative data sets. Big Data may also be subjected to nonsampling errors which can arise when the analysis involves further transforming the data (e.g., errors can occur while data are being coded, edited or imputed.), as well as can be errors due to modeling and estimation (e.g., using the wrong analytical tools).

In sum, verifying data quality is a process that should be applied both to the data generation level and to its management level once data is collected. Regardless of the source of flawed data, the consequences of analyzing data affected by poor quality might be drastic.

2.5. Big Data and Causality

Another debated issue related to the use of Big Data concerns the study of causality. Generally, Big Data provides an exploratory analytical framework that, devoid of specific guiding hypotheses, is mainly a data-driven search of correlations among sets of variables. This approach may not be compelling for psychologists who want to rely their conclusions on theory-driven researches. However, I see two ways of how Big Data can help theory-driven researches.

Even if causation and correlation are separated concepts, the latter may be suggestive of the former. Indeed, a data-driven approach can shed some light on novel (ir)regularities, which can generate new theoretical hypotheses that can then be evaluated in subsequent hypothesis-testing studies. For instance, Hernandez, Newman, and Jeon (2016) analyzed the

[1] This is the process when the data are brought together in the same computing environment with the process of extraction (data accessed, parsed and stored from multiple sources), transformation (e.g., coding, recoding, editing) and loading (integration and storage).

Twitter feeds from the 200 largest cities in the USA. By focusing on phrases related to "loving" or "hating" a job, Hernandez and colleagues found that the content of job-related tweets from people in cities with high SES, occupational prestige, and commute times was more negative than other cities. The consideration of "macro-level" attitudes and behavior that is enabled by Big Data opens up entirely new research questions and the potential for new streams of work. Using exploratory analyses to drive the development of new theories is not new in psychology. However, Big Data offers the opportunity to conduct such analyses more efficiently and effectively.

Big Data can be used in an experimental fashion to conduct controlled studies. Experimental conditions can be emulated in the field by using large samples and many features to construct controlled matched groups. For example, to investigate the effect of September 11, 2001, on the political behaviors of victims' families and neighbors, Hersh (2013) used the data from the New York State registered voters' database, which contained personal information of all 9,995,513 New Yorkers registered at that time. Then, he identified 9/11 victims and their families living in the State of New York and then traced how their families' political activities changed as compared to a control group chosen to be highly similar in the pre-9/11 period. Although social scientists have previously found psychological and behavioral effects of terrorism and acts of violence, never before have such effects been estimated with objective individual-level indicators rather than through self-reports or aggregated data. Not only the Big Data approach enabled the author to access large and ecological data, but also to get information about the people's political activity ten years before and after the terrorist attack. This data allowed Hersh to analyze the trend of the effect and its persistence over time. This study shows us that causal modeling is possible in a Big Data paradigm by conducting Big Data experiments. Google, for example, is famously known for conducting about a thousand experiments at any given instance (Varian, 2013). Telecom network operators themselves utilize such techniques when rolling out new services or for

figuring out pricing. Compared to traditional laboratory experiments, Big Data experiments do not only allow access to an enormous number of individuals in any given study, but they also allow access to an enormous number of studies conducted in different locations and under different conditions. Thus, Big Data experiments can boost external validity, which is the translation of the findings to new populations (Pearl & Mckenzie, 2018). Another advantage that comes from Big Data experiments is that they are easier to conduct. A researcher does not need to recruit and pay participants. Indeed, she can write a line of code to assign them to a group randomly. She does not need participants to fill out surveys. Indeed, she can measure mouse movement and clicks. She does not need to hand-code because he can build a program to do that automatically. In the era of Big Data, all the world is a lab (Stephens-Davidowitz, 2017).

In addition, Big Data samples contain abundant traces of natural experiments in which random or quasi-random subsets of people are exposed to a particular situational influence, such as a natural disaster, regional policy change, or geographically limited interruption in their access to digital services. Such natural experiments offer an opportunity to study the effects of factors that would be difficult to simulate in a controlled experimental setting. For example, in the United States Hurricane Ike was used to study the short- and long-term causal effects of natural disasters on social networks. A study based on a sample of 1.5 million Facebook profiles of US college students showed that people affected by this natural disaster formed stronger bonds than those who were unaffected (Phan & Airoldi, 2015). As the exact paths of hurricanes are difficult to predict and can change unexpectedly, they affect a quasi-random subset of communities, enabling the discovery of causal mechanisms.

2.6. Epistemology of Big Data

In the recent years there has been discussion about whether the rise of Big Data calls for a new kind of epistemological understanding of science (e.g. Floridi, 2012; Frické, 2015; Hey, Tansley, & Tolle, 2009; Kitchin, 2014). For instance, Kitchin stated that the development of

Big Data and its analytics offers the possibility of reframing the epistemology of science and such a reframing is already actively taking place across disciplines (Kitchin 2014).

This epistemological reframing is due to the idea that Big Data enable a novel form of inquiry, called *data-driven science*, which seeks to generate scientific hypotheses by discovering patterns in vast amounts of data (Kelling et al. 2009; Kitchin 2014). Data-driven science contrasts with the more traditional 'theory-driven science', where the hypotheses to be examined are derived from theory rather than data (Kelling et al. 2009). Thus, the data-driven design differs from the traditional, experimental deductive design because it seeks to generate insights "born from the data" rather than "born from the theory" (Kelling, et al., 2009).

Data-driven science, which implies working from observations to draw inferences about underlying patterns, is a declination of inductive research. Contemporary applications of Big Data analytics are characterized by evaluating information about individuals with an intention of identifying (and subsequently using) relations between measured variables. In short, the Big Data scientist often does not have any expectations, theories, or hypotheses about the underlying relationships, but rather uses the observed patterns in the data to guide future decisions.

Though not identical, such Big Data applications have similarities to applications of inductive reasoning in more traditional research. With some notable exceptions and despite recent efforts to boost more inductive research, inductive approaches in psychology have traditionally been avoided. Indeed, drawing strong conclusions from small sample is probably not the best application of inductive methods. In contrast, Big Data affords contemporary researchers with much richer and more abundant information than previously available.

However, also in the realm of Big Data, when one attempts to draw inferences, the quality of the data (e.g., their provenance, the technologies that generated them, the reasons why they were created) necessarily influences the conclusions that can be derived. Brooks (2013)

contends that Big Data analytics struggles with the social dimension (e.g., data excels at measuring the quantity of social interactions but not the quality); struggles with context (e.g., data are largely decontextualized from the social, political, economic and historical context); creates bigger haystacks consisting of many more spurious correlations, making it difficult to identify needles; and, identifies trends but not necessarily meaningful features that may become a trend. In other words, whilst Big Data might provide some insights, it needs to be recognized that they are limited in scope, produce particular kinds of knowledge, and still need contextualization with respect to other information, whether that be existing theory, small data studies, or historical records, that can help to make sense of the emerged patterns (Crampton et al., 2013). Beyond the epistemological and methodological approach, part of the issue is that much Big Data analysis is conducted with no specific questions in mind, or the focus is driven by the application of a method or the content of the data set rather than a particular question. For example, geotagged Twitter data has not been produced to provide answers with respect to the geographical concentration of language groups in a place and the processes driving such concentration. We should perhaps not be surprised, then, that it only provides a surface snapshot, albeit an interesting one, rather than deep insights into the geographies of language, agglomeration and segregation in particular places (Crampton et al., 2013).

How Big Data are generated or repurposed is guided by certain assumptions, underpinned by theoretical and practical knowledge. Big Data are both a representation and a sample, shaped by the technology and platform used, the data ontology employed and the regulatory environment, and it is subject to sampling bias (Crawford, 2013; Kitchin, 2013). As such, an inductive strategy of identifying patterns within data does not occur in a scientific vacuum and is discursively framed by previous findings and theories; by speculation that is grounded in experience and knowledge (Leonelli, 2012). Thus, the process of induction (i.e., insights emerging from the data) is contextually framed. The

derived insights provide the basis for the formulation of hypotheses and the deductive testing of their validity. In other words, data-driven science is a reconfigured version of the traditional scientific method, providing a new way in which to build theory. Rather than decree "the end of theory" (Anderson, 2008), researchers are arguing that data-driven science will become the new paradigm of scientific method in an age of Big Data because its epistemology is suited to extracting additional, valuable insights that traditional 'knowledge-driven science' may fail to generate (Kelling et al., 2009; Miller, 2010).

I consider that neither deductive nor inductive reasoning is 'better' in a broad, general sense, but rather that they both provide valuable ways of understanding and explaining people's psychological processes and behaviors. Additionally, Big Data can be used in either deductive or inductive research. Indeed, the richness of such data and the wealth of information included in big datasets may stimulate researchers to explore (in an inductive manner) new relations with the goal of generating theoretical models that could then be tested using deductive methods. In other words, the epistemological strategy adopted within data-driven science is to use guided knowledge discovery techniques to identify potential questions (hypotheses) worthy of further examination and testing. Thus, the argument is that Big Data can reorient the roles that data and theory play in research, and that therefore we should rethink our conception of how scientific knowledge production works.

2.7. Privacy Concerns

Big Data often is personal and sensitive. This raises many questions about how protecting and using this data to guarantee people's privacy, that is, people's right to control and maintain their autonomy, individuality, and free expression of it, the right to be let alone (Cooley, 1888). Breaches in data flow or inappropriate use, regardless of the presence of intentionality in doing that, can cause serious harm to the involved people (Richards & King, 2013). Problematic consequences may arise from identity disclosure, such as discrimination and stigma, or identity distortion, such as false profiling.

Big Data redefines the rules of ethical research practices and raises questions about what kind of controls need to be set to protect research participants. Psychological research already has a strong foundation in securing data and protecting the privacy of participants, as long as it frequently handles sensitive information about participants (e.g., personality traits, suicidal intentions). For conducting ethical research, psychologists must obtain the consent of the participants before including them as part of a sample, especially when they are affected by an experiment or an intervention that involves risk. However, in the Big Data realm, it could be difficult, for example, to provide meaningful privacy information to target people due to the complexity of the procedures involved in the analysis of this large amount of data. Furthermore, it is implausible identifying all the possible purposes that can be potentially be achieved from those data. This point is of fundamental importance since it could be challenging to obtain a "valid consent" from the subjects to whom the data refer, especially when participants do not directly provide data (e.g., answers to a questionnaire) because it is collected automatically by a multitude of technological devices. Thus, the characteristic of Big Data to use any form of data for new and unexpected purposes could conflict with the principle of finality, which considers the processing of personal data legitimate in relation to the declared purpose(s) of such a processing. It is true that privacy does not mean that all data must stay private, but just because a data is made publicly available does not directly mean that it can be ethically used for research.

In this context, the Institutional Review Boards (IRB) play an increasingly central role in extending and adapting ethical research principles to Big Data projects. Although IRBs cannot always foresee the damage that a particular study can cause, the value they bring is to induce researchers to think critically about the ethics of their projects. Moreover, it is also important to remark the role of governmental institutions in defining to what extent and in which terms data can be processed and used in Big Data applications. The latest critical step in regularizing privacy issues in Europe is the establishment of the General Data Protection

Regulation (2018), which regulates the processing of consumer data by organizations. The scope of this regulation is to provide consumers greater control of their data by setting limitations on what companies can do with that data and how long they can retain it.

2.8. The Skills Gap

When I addressed Big Data's Volume, I stated that psychology research is largely theory-driven. This characteristic is reflected in the type of knowledge that is taught in psychology courses. The knowledge of psychological theories can be extremely useful because theoretical knowledge can be applied to Big Data in order to provide a theoretical justification both for guiding the analysis and for interpreting the conclusions that can be drawn. However, psychologists' training on Big Data is still scarce.

In chapter II, I describe how and where it is possible to obtain data (data acquisition), how to approach these data once obtained (data preprocessing), the statistical skills useful for analyzing Big Data (data mining), technological and learning tools useful for psychologists who want to learn Big Data.

3. Big Data in Consumer Psychology

The main goal of consumer psychology is to explain the causal mechanisms that drive consumers' thoughts, choices and behaviors. This goal is epitomized in the ultimate tool of psychological science: A tightly controlled randomized experiment that focuses on a handful of carefully measured variables. To explain consumer behavior, research on consumer psychology generally gathers 'human-centric', primary consumer data (i.e., data collected by a researcher from first-hand sources) using surveys, focus groups, interviews, observation studies, and panel data (Wedel & Kannan, 2016). In the consumer research literature, the outlined methods have generated valuable customer knowledge that tapped into the underlying processes of consumers' behavior.

In contrast, the Big Data approach has been preoccupied with prediction. Datasets comprise many observations and variables combined with powerful analytical tools enable researchers to build models that encompass hundreds or thousands of predictors. Such models can be very good at predicting future outcomes and behavior, often with good levels of accuracy. Moreover, thanks to technological advances in the collection, storage and analysis of large amounts of data, consumer research is facing great opportunities (Lazer et al., 2009). Researchers can now gain valid insights on millions of consumers by looking at the digital records that are passively collected as consumers live their daily lives. For example, posts in social media, customer forums, and product reviews make it possible to observe large and natural focus groups at very little cost (Netzer, Feldman, Goldenberg, & Fresko, 2012). Similarly, observing consumers' behavior in a traditional retail store is very similar to analyzing the journey of a customer who is browsing a company's online store (e.g., one can examine the characteristics of products the customer has looked at and/or bought, measure the time they took to make a decision, or implement mouse-tracking technologies to study the decision process).

The contrast between traditional explanation-focused psychology and the prediction-focused Big Data approach does not mean that they are incompatible. Presumably, the ultimate motive of psychologists who strive to explain psychological mechanisms is to build theories that can predict real-life behavior. However, traditional tools employed by psychologists (such as controlled experiments) are typically focused on testing well-defined hypotheses in an artificial environment, like laboratories. As such, they are unlikely to discover mechanisms that were not hypothesized. Moreover, they are not well suited to examination of whether a given theory is predictive of real-life behavior. Consequently, while it is clear that both prediction and explanation are important, psychology has focused mainly on designing and testing models (theories) and somewhat less on testing whether such models predict behavior outside of the lab. Thus, elevating Big Data research to a stature equal to that

of traditional psychological research and incorporating prediction-focused statistical tools could greatly benefit the field of psychological science.

In Chapter IV, I will provide an empirical example of how Big Data and its analytics (i.e., data mining and machine learning) can be used to predict consumers' churn behavior and to extract psychological knowledge that might explain such behavior. By mining Big customer Data from an Italian energy provider database, I first predict future churn behavior through the construction of a predictive model. Then, using the results of the most predictive model, I provide post hoc explanations on the psychological underpinnings that may have driven that behavior. The derived explanations, in turn, shed light on new hypotheses that may be deductively tested both in the laboratory and "natural" settings.

The use of data mining on Big customer Data is widespread in the field of Customer Relationship Management (CRM). Customer data gathered by companies have the potential to become an asset for the analysis of the psychological underpinnings of customer behavior. The voluminous data produced by these CRM databases enables detailed examination of behavioral factors that contribute to customer acquisition, repurchase intentions (e.g., cross-selling), consumer retention, loyalty and other behavioral intentions such as the willingness to provide positive referrals or become brand advocates. Customers' data is the key to business success. This explains the fact that many who study consumer behavior come from disciplines (e.g., economics, marketing, advertising, retailing) that support business activities. But why do psychologists find it worthy for study? First, consumer behavior is a domain providing innumerable opportunities for important, socially meaningful research. Second, consumer theory and research have the potential to contribute to the development and extension of psychology itself. Examples of such contributions abound and include the conceptual as well as the empirical methods that enable psychologists to better test dynamic theories. Third, because of the universal and pervasive nature of consumer behavior, this context provides excellent opportunities for examining the validity and limits of psychological theory. Fourth,

consumer research provides numerous opportunities for working with dependent variables that, from the perspective of test subjects, are meaningful and consequential (Jacoby & Morrin, 2015).

Consequently, having a clear and comprehensive view of how data mining and machine learning have been used in the CRM literature can provide with useful information on how to deal with this type of data, the data mining tasks, and extract knowledge from it. In the next section, I introduce the relationship between the use of Big Data and data mining in CRM processes and what these processes are.

4. Big Data and CRM

Big Data is increasingly used to optimize business intelligence processes (Marr, 2019). The term "business intelligence" was used by Devens for describing how the banker Henry Furnese achieved an advantage over competitors by collecting and analyzing information relevant to his business activities in a structured manner. This is thought to be the first study of a business putting data analysis to use for commercial purposes. It sounds like something that happened only recently, does it? It was 1865. This anecdote shows us that the idea to leverage data and information for doing and improving business has a long steady history. However, only in the last 30 years, firms have acquired the capabilities (e.g., computational, IT infrastructures) to ground their business processes and decisions on data-driven insights effectively.

Companies leveraging Big Data can enhance their competitive advantage in a world where markets are global, and vast amounts of information about consumers is available on the internet (Verhoef et al., 2016). Big Data is a significant source for business intelligence activities aimed at creating, delivering, and capturing customer value (Verhoef et al.,2016). Thus, the emergence of Big Data has brought a new wave of Customer Relationship

Management (CRM) 's strategies in supporting personalization and customization of sales, services, and customer services (Anshari, Almunawar, Lim & Al-mudimigh, 2019).

CRM is a combination of business strategies through which companies manage relationships and interactions with potential and existing customers. In small business contexts, learning the customers' preferences, attitudes, and opinions is easily done. Think of running a small grocery shop in your hometown. Once a customer crosses the door, we already know who he is, what he likes, if he is vegetarian, and so on. Well-run small businesses naturally form relationships with their customers. Based on this knowledge, we can take the best actions that let him buy our products and put him in the condition to want to return. So, we adjust the way to welcome him, what products recommend, how we talk to him. The result is happy, loyal customers, and a profitable business. Now think of running a big distribution chain with thousands, or even millions of customers. In this case, it is highly implausible having a personal relationship with each customer. Therefore, such a firm must rely on other means to form a relationship with each customer. In this sense, the firm must take full advantage of something it has in abundance: customer data. There is a great variety of data that say something on customers: purchases and transactions, web and online behaviors, socio-demographic and socio-economic information, interactions with the company, and census data, among others (Artun & Levin, 2015).

Moreover, the availability of computers, mass and cloud storage, and advance statistical and analysis methods provide companies the key to accessing information. For extracting practically significant information from the data, companies rely on data mining, which provides a standardized procedure that assures quality and repeatable results. Data mining can help gain a better understanding of customers and their needs in various ways. For example, it can be useful to identify segments of customers with similar behaviors and needs or identify targets that are predicted to have a positive impact on revenues (Kumar & Reinartz, 2012).

The use of data mining in CRM processes allows companies to improve the way they manage their relationships with customers during the different phases of the customer lifecycle. In CRM terms, the customer lifecycle describes the various stages a consumer goes through before, during, and after they complete a transaction or cancel their subscription. The business relationship with customers changes over time. Depending on the business context, the type of relationship that can be established may be different because of the type of products or services offered or the way by which these products and services are sold (subscription relationship vs. event-based relationship[2]). Despite these differences, five general phases can describe the ways through which a relationship evolves (Dwyer, 1987; Berry & Linoff, 2011):

- The customer relationship process starts long before a person becomes a customer. A company should identify those people who are in the target market and may be turned into customers. These people are called prospects. This is the phase where a company begins the relationship by letting prospects become aware of the company brand, products, or services. People can learn about a company in many ways, such as through traditional marketing and word-of-mouth.
- After a person learns about the product, they might decide that they want to try or buy it. Those people who exhibited some interest are defined as responders. This begins the acquisition period. However, showing an interest towards a product or service is a necessary but not sufficient condition for becoming a customer. For example, a person might download an app, but to effectively become a customer, she must go through a registration process that may not be completed for different reasons (e.g., lack of time, uncertainty). Thus, the acquisition phase starts with the intent to buy a product and ends when the customer buys that product (or subscribes to a contract).

[2] Subscription-based relationships are based on subscriptions (e.g., subscribe a contract with a service provider), while event-based relationships are based on transactions (e.g., make a purchase from a grocery store).

- New customers are responders who have committed, such as subscribing to a contract or registering on a web site providing personal information.
- Just because someone has bought a product or signed a contract does not mean that he is a valuable customer. For example, someone who has a social network account but does not show any activity over time is a customer, but not an engaged one. Thus, new customers should be turned into established, and possibly, loyal customers.
- In the end, not all the relationships last forever. When a business relationship ends, customers (established or not) become former customers. The act of abandoning a company is defined in different ways across industries. Some call it attrition, others defection. Others again call it churning. Whatever the noun is, they share the ideas of what motivations lead a customer to move away. Customers may leave because of voluntary (e.g., they have churned to a competitor) or involuntary reasons (or forced; e.g., they have not paid the bills). In some cases, churn is expected (e.g., they shift out from the target market; for example, because customers have grown older and progress through the family lifecycle).
- The customer lifecycle is a central element because the CRM business processes supported by data mining are organized around the phases of this lifecycle. These CRM business processes are:
- Customer acquisition is the process of attracting prospects and turning them into customers (Berry & Linoff, 2011). The first step of a customer acquisition plan is the identification of prospects. Prospecting is a mining term, that means searching an area thought likely to yield a valuable mineral deposit. In CRM, it means searching for opportunities that might convert into strategically significant customers. Prospecting is a result of segmenting and targeting. Segmentation divides a heterogeneous market into homogeneous groups, even down to the level of the unique customer. Targeting is the process of choosing which market segments, clusters, or individuals, to approach with an offer. (Buttle, 2004)

- Customer development is the process of growing the value of customers. To increase customers' value, companies generally attempt to cross-sell and up-sell products or services to their customers. Cross-selling involves the sales of additional items related (or sometimes unrelated) to a previously purchased item. Up-selling involves the increase of order volume either by the sales of more units of the same purchased item, or the upgrading into a more expensive version of the purchased item.
- Customer retention includes the activities implemented by a company to retain its customers over time and minimize churn. In the broadest and most general sense, it means maintaining long-term relationships with customers. For many firms, the efforts to retain a customer are reactive; for example, an insurance company offers some incentive to a customer who indicates that they wish to cancel their contract. Increasingly, firms are becoming proactive, undertaking their retention marketing activities before the customer has the opportunity to churn.
- Customer win-back entails all the strategies to reactivate an ended relationship with former customers. Win-back tries to do that by providing former customers with incentives, products, or pricing promotions (Berry & Linoff, 2011).
- The notion of customer acquisition, development (i.e., cross-sell and up-sell), retention, and win-back being the four key drivers of the company's organic growth is widely accepted and has even made its way into core marketing teaching materials (e.g., Gupta 2014). The importance of such dimensions is even more emphasized by the fact that techniques such as data mining are used to improve them continually. Data mining is a valuable tool for analyzing the customer lifecycle because it can improve the profitability in each lifecycle stage.

Other than helping firms to improve their profitability and competitiveness in the market, using data mining on CRM data can be helpful also for understanding the psychological underpinnings of consumers behavior. Thanks to technological advances in the collection,

storage and analysis of large amounts of data, companies can now gain valid insights on millions of consumers by looking at the digital records that are passively collected as consumers go about their daily lives. The sources of information companies can tap into to learn more about their consumers are almost limitless. Among the most vital ones are historical purchasing data, credit card records, search queries, browsing histories, blog posts, social media profiles. Importantly, it is often possible to combine the information extracted from different sources to form a more holistic picture of a consumer's daily habits and preferences.

To understand how data mining has been used over the years on CRM data, in Chapter III, I will revise the literature published from 2000 and 2018 on the application of data mining techniques in the following CRM processes: customer acquisition, cross-sell, churn, and customer win-back.

5. Dissertation Outline

I present three chapters which aims are summarized in Table 1.

Table 1.1. Thesis Overview.

Chapter	Chapter Aim	Contributions
Chapter 2	Provides an overview of the KDD steps to those psychologists who want to learn how to conduct Big Data projects	Although there are papers in the literature that aim to provide technical guidance to psychologists who want to do Big Data research (e.g., Chen, & Wojcik, 2016), this is the first work that has extensively covered the data pre-processing and data mining phase.
Chapter 3	Provides a literature review on the use of data mining on CRM data about four CRM processes (i.e., acquisition, cross-selling, churn, and win-back)	This is the first literature review on the use of data mining and machine learning on CRM data about four CRM processes that considered a wide range of works during the last 18 years.
Chapter 4	Using data mining and machine learning techniques to predict customer churn for an energy retailer and shed light on the psychological underpinnings of churn behavior	This is the first work in which data mining and machine learning are used on a Big Data set to understand the psychological underpinnings of customers' churn behavior.

In Chapter II, I provide statistical and technical knowledge to those psychologists who want to learn how to conduct Big Data projects. By taking the Knowledge Discovery from Database

(KDD; Fayadd, et al., 1996) steps as the fil rouge, I discuss where to find data suitable for psychological investigations, the methods to preprocess them, the techniques through which analyze them, and the programming languages through which all these steps can be implemented. In Chapter III, I present a literature review on the use of data mining in Customer Relationship Management, particularly on the acquisition, cross-sell, churn, and win-back processes. By analyzing nearly 500 publications on the topic, I will discuss, among others, their characteristics, the techniques that have been used, the characteristics of the analyzed datasets, and the predictors that showed to be related to customer behaviors. In Chapter IV, I apply data mining and machine learning methods and techniques to predict customer churn for an energy retailer. Starting from those predictions, I address *a posteriori* psychological explanations of the discovered relationships between predictors and the behavioral outcome (i.e., whether a customer will churn). The study provides useful insights into the consumer psychology field, which can be used to drive future experimental research.

A stylistic note: the introduction and the conclusion of this dissertation are written using the first singular person ("I"), while the other chapters use the first plural person ("We"). This is motivated by the will to reflect different authoring process – mostly individual in Chapter I and V, and largely combined in Chapters II, III, and IV.

CHAPTER II

A GUIDE FOR CONDUCTING PSYCHOLOGICAL RESEARCH WITH BIG DATA

1. Introduction

The human beings' tendency to record and store information has a long history. It is thought that the first evidence of this behavior dates back to around 18000 BC when the Ishango Bone was created, which was classified as a tally stick[3]. Recording and storing information has gradually intensified to the present days, where the capabilities of computers and computational systems allow us to manage and use large amounts of data to generate insights. The term Big Data emerged in the late 1990s (e.g., Cox & Ellsworth, 1997), but it took several years for the concept to enter the scientific and public imagination. The current use of the term Big Data goes back to 2001 when Laney introduced the milestone definition of Big Data based on its essential dimensions: volume, variety, and velocity. Since then, it was clear that the Big Data would have changed the way of doing science entirely, so much so that Wired (2008) proclaimed the end of the "Theory Age" and welcomed the "Petabyte Age", where the data deluge would have made the scientific method obsolete. Big Data is a result of the information age and is changing how people do science. Although many scientific fields have leveraged Big Data since its appearance (e.g., computer science, business, physics), research in psychology has only recently started to apply Big Data to address psychological questions (e.g., Beaton, Dunlop, & Abdi, 2016). From a bibliographic research I conducted on Web of Science, it has emerged that 117 works dealing with the use

[3] A tally stick is an ancient memory aid device used to record and document numbers, quantities, or even messages. Thus, carving notches on bones or stones has been called by many the first information revolution (Marshack, 1972).

of Big Data in psychology have been published between 2005 and 2019 (see Appendix A for the full list). Three leading journals have dedicated a special issue on the topic (*Psychological Methods*, *Current Opinion in Behavioral Sciences*, and *Zeitschrift Fur Psychologie*). Most of the published materials discussed the use of Big Data in general as a new and valuable methodological framework in psychological research without referring to any specific psychological field. A part of the published material, instead, has dealt with the use of Big Data in health psychology (e.g., Yetton, Revord, Margolis, Lyubomirsky, & Seitz, 2019), industrial & organizational psychology (e.g., Kobayashi, Mol, Berkers, Kismihók, & Den Hartog, 2018), clinical psychology (e.g., Hollon, Cohen, Singla, & Andrews, 2019), and cognitive psychology (e.g., Stevens & Soh, 2018). The relevance of the topic is further emphasized by the fact that research institutes have organized conferences (Big Data in Psychology 2018 and 2019, by Leibniz Institute for Psychology Information - ZPID[4]) and forums (Big Data in personality and social psychology, by the Society of Personality and Social Psychology[5]).

It may be expected that larger and larger datasets will be available in psychology in the future (Cheung & Jak, 2016). Although psychologists often show a sense of excitement when talking about these new data opportunities, this enthusiasm has not yet let to extensive use of Big Data in the psychological community. There are several reasons why the use of Big Data is still limited. According to Paxton & Griffiths (2017), these reasons can be summarized in three categories: what they call an "imagination gap", cultural reasons, and lack of skills.

By "imagination gap" they refer to the observation that often psychologists do not consider their research questions as a Big Data problem. Nowadays researchers know that companies, governments, and other organizations are capturing massive amounts of data. However, very few can envision a dataset that would address an important theoretical question in their field.

[4]https://conferences.leibniz-psychology.org/index.php/index/index/index/index

[5]http://www.spsp.org/events/SPF/2019

Bridging the imagination gap will take some work to adjust psychology field's idea of the possible scope of data beyond experimentally generated datasets. This requires the curiosity to continue hunting down new possible datasets, the theory-guided creativity to see their potential, and the ethical constitution to critically question their use. A wealth of data can be used to address a variety of research areas.

For example, researchers interested in understanding categorization might investigate tagging behavior in the Yahoo Flickr Creative Commons 100M dataset (Thomee et al., 2015). Play-by-play sports records might be useful for studying team dynamics (Horowitz, 2015), and decades of online chess game records could shed light on expertise and decision making (e.g., Free Internet Chess Server Database).

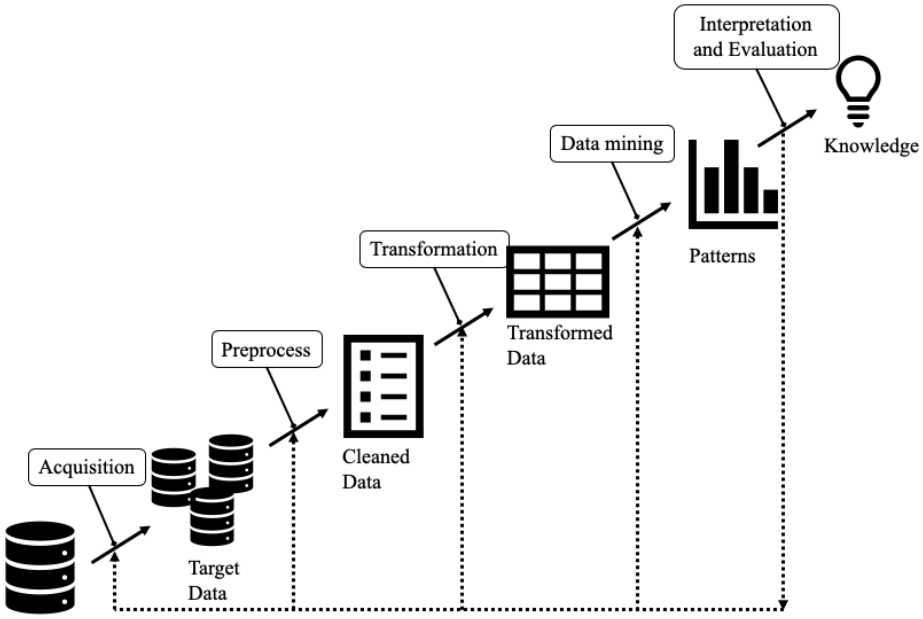
The cultural reasons Paxton & Griffiths refer to the difficulty in getting Big Data perspective adopted by individuals and institutions within psychology. The difference between interest in Big Data research and utilizing Big Data in research can be partially attributable to a lack of role models and acceptance of these new data resources. To date, only few publications encouraged the use or actually used data mining (Li, 2019), machine learning techniques (Yarkoni & Westfall, 2017; J. Zhao, Zhang, He, & Zuo, 2019). The internet provides everyone the ability to access data at any time, for any need. Unfortunately, it does not help guarantee that the data is valid, or clean. In general, successful Big Data research requires having to build a whole range of skills and competencies necessary to conduct that kind of projects. In this chapter, I wish to provide some useful information to whoever intends to approach Big Data analytics. Firstly, I will describe the Knowledge Discovery from Database (KDD; Fayyad, Piatetsky-Shapiro, & Smyth, 1996), that is the process of discovering useful knowledge from a collection of data. Next, I will describe in separate sections how and where it is possible to acquire data and the fundamental techniques for preprocessing and transforming it. Later on, I will introduce the data mining and machine learning techniques (i.e., what they are, what they do, and why to use them), and finally the

available tools to implement them. Expanding the capabilities will be key for psychologists to making an impact in the Big Data era. If nothing else, psychologists have the chance to learn how to speak the language of Big Data methods and data analytics.

2. The Knowledge Discovery in Databases Process for Analyzing Big Data

The first concept a psychologist should learn when approaching the Big Data world is the KDD, which is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. This process can satisfy the requirements that Big Data analytics entail and provide a structured step-by-step path for analyzing large and complex sets of data. The KDD process consists of five essential and iterative steps (Figure 2.1).

Figure 2.1. The KDD process.



The iterative nature of KDD indicates that knowledge discovery often involves experimentation, iteration, user interaction, and Thus, one can move back to adjust previous steps whenever it is required. As a consequence, KDD has many “degrees of freedom”, meaning that one cannot present one formula or make a complete taxonomy for the right choices for each step and application type. Thus, it becomes essential to understand the

process as a whole and to know what are the aims to be achieved at each step. The KDD five steps are:

1. *Defining application domain and goals.* This step requires the definition of the goals. It includes the identification of relevant knowledge. It also converts these goals into a data mining problem definition, which is essential to design a preliminary project plan that will regulate the next phases.
2. *Data acquisition and selection.* Based on the defined goals, psychologists should determine and identify the data useful for achieving them. This includes finding out what data are available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one data set, including the attributes that will be considered for the process (Maimon & Rokach, 2010). This process is critical because the data mining techniques learn and discover from the available data. If some important variables are missing, then the entire knowledge discovery process may fail. On the one hand, it may be successful to consider as many data as possible at this stage. On the other hand, to collect, organize, and operate complex datasets may be too expensive. Thus, the selection of data and variables should be the result of a compromise between those two aspects. In this sense, the theoretical training of psychologists can be of great value because it can drive the selection of useful data on which conduct Big Data projects.
3. *Data Preprocessing*. In this stage, data reliability is enhanced. It includes operations for data integration (where multiple data sources may be combined into one), handling missing and inconsistent values, identifying outliers. Among all the steps of KDD process, data preprocessing plays a vital role. This is because it removes duplicate records, unnecessary data fields, and standardize the data format.
4. *Data Transformation*. In this stage, the generation of better data for the data mining is outfitted and developed. Methods here include attribute transformation (e.g.,

normalization) and data reduction (e.g., dimensionality reduction). This step is to transforms the data more efficiently based on the desired goals.

5. *Data mining.* Data mining is the set of automatic mechanisms designed to allow the exploration of the preprocessed data in search of consistent trends and systematic relationships between variables. Data mining classical statistics, machine learning, and artificial intelligence (Han, Kamber, & Pei, 2012). The data mining step entails making a set of actions: Firstly, choosing the data mining task, then choosing the algorithm, and finally building and validating the model. I will describe these substeps thoroughly in the dedicated section.
6. *Evaluation, interpretation, and use of the discovered knowledge.* In this stage, we evaluate and interpret the extracted patterns, with respect to the goals defined in the first step. This step focuses on the comprehensibility and usefulness of the data mining model. The results of the model are evaluated to identify those patterns which contain useful information. The knowledge becomes active, in the sense that we can use it to address new psychological explanations of the detected effect and to drive new psychological research. The success of this phase determines the effectiveness of the entire KDD process.

Usually, the data mining stage is considered the key step of the entire KDD process, and this emphasis makes it increasingly difficult, especially in practical terms, to distinguish the KDD process from data mining. However, data mining is only one step in the overall process. Even if some use the terms interchangeably, I will use them as separate concepts and consider the data mining as a step of the KDD process, which indeed denotes the overall process of extracting high-level knowledge from low-level data. Thus, the quality of the mined knowledge is a function of both the effectiveness of the data mining technique used and the quality of the data acquired and preprocessed. If researchers select the wrong data or data mining algorithms, use irrelevant variables, or transform the selected data inappropriately, the process will likely be a failure (Sumathi & Sivanandam, 2006).

2.1. Data Sources and Acquisition

A relevant category of data sources is social media. Social media have been defined as web-services that allow individuals to create a public or private profile within the system and connect to other people using that system (Boyd & Ellison, 2008). Social media differ in how individuals interact. For example, Twitter allows communicating with the broadest audience possible, while Facebook interactions are more selective and personal, and LinkedIn promotes selective but professional connections. Compared to archival repositories, social media often lack the user-friendly interfaces that make it easy to extract data from their systems. Social networks, indeed, require the use of Application Programming Interfaces (APIs), a set of protocols and technologies that enable users to access the rich data that these platforms collect about users. Generally, the process for requesting data require: 1) the registration of the user on the platform (e.g., in Twitter users wanting to scrape data must sign up as developers and declare why and how they are going to use the requested data), 2) the data request by calling an endpoint (i.e., an address that corresponds to a specific information; for example a Twitter hashtag like *#ilovebigdatainpsychology*), and 3) the automatic provision of requested data to the user (e.g., accessing all the tweets that contain the called hashtag and other information related to the user that tweeted it). Some APIs are open access and do not impose too many constraints on the quantity of data that can be retrieved. For having a broad idea of how to use the API, I suggest reading the paper by Chen and Wojcik (2016), showed how to extract data from Twitter.

Streaming data, recorded through sensors, is another valuable source of data. In light of their ubiquity and the fact that they come already equipped with numerous embedded sensors, such as accelerometer, GPS, Wi-Fi, light sensors, microphones, log data (Lane et al., 2010), smartphones offer a unique opportunity to collect streaming data. Smartphone sensors can obtain a great deal of information about their owners' behavioral lifestyles. They routinely record sociability (who we interact with via calls, texts, and social media apps) and mobility

behaviors (where we are via accelerometer and global positioning system) as part of their daily functioning. Smartphone-sensing research is flourishing in the field of computer science but recently smartphones have begun to enter the methodological toolkit of psychological researchers (Gosling & Mason, 2015; Wrzus & Mehl, 2015). Sensing methods have been already used to study topics such as sleeping patterns and postures (Wrzus et al., 2012), interpersonal behaviors in group settings (Schmid Mast, Gatica-Perez, Frauendorfer, Nguyen, & Choudhury, 2015), and emotional variation in daily life (Rachuri et al., 2010; Sandstrom, Lathia, Mascolo, & Rentfrow, 2017). Studies have explored the links between mobility patterns and depression (Chow et al., 2017), schizophrenia (Wang et al., 2016), bipolar disorder (Abdullah et al., 2016), physical activity and well-being, mental health, educational outcomes, happiness, and face-to-face encounters (Lathia, Sandstrom, Mascolo, & Rentfrow, 2017; Wang et al., 2014; Wang, Harari, Hao, Zhou, & Campbell, 2015). Nowadays, researchers are able to detect more complex behavior when they combine data from different sensors and sources. For example, combining data from microphones and accelerometers enables the detection of common behaviors such as clapping, vacuuming, or taking out the trash (Lu, Pan, Lane, Choudhury, & Campbell, 2009). Cartographic data merged with GPS data can be used to identify a user's location (e.g., workplace, home, or shopping center) and particular routes that he or she has traveled to get to those places (Eagle & Pentland, 2009). Merging GPS data and signals from WiFi transmitters or accelerometers can be used to track individuals within a building (Chon & Cha, 2011), detect modes of transportation (Hemminki, Nurmi, & Tarkoma, 2013), and monitor pedestrian behavior (Wang, Guo, Peng, Zhou, & Yu, 2016). Information about phone usage, combined with data from other sensors, such as ambient light sensors, microphones, or accelerometers, can reveal users' sleep patterns (Hao, Xing, & Zhou, 2013) or alertness (Murnane et al., 2016).

Although Big Data research in psychology is often closely associated with social media and streaming data (Chen & Wojcik, 2016), there are many systems that store data that

psychologists can use to conduct Big Data projects. The digitalization of modern life enables the acquisition of data from a broad range of areas. Commercial data produced by companies is the most substantial source of Big Data. Many company platforms and systems involve large segments of the population while also capturing outcomes of psychological interest. However, accessing these data could be complicated for people external to the company. Firstly, the procedures for accessing and acquiring data are often expensive for companies. Companies should, in fact, acquire new or reallocate existing resources (e.g., workforce, technology) to allow researchers to use their data. Secondly, collaborations with research institutes may expose companies to the risk of spreading strategic information that would ruin the competitive advantages on the market, receiving negative press, invading the customers' privacy, as well as violating the terms of service of the company itself. These issues arise especially when researchers are willing to perform experiments on a randomized sample of customers (e.g., to test the effectiveness of marketing tools) or if they are willing to use the communication network of the customers (e.g., if they want to know how negative word-of-mouth spreads in the customers' network within the company).

An alternative option concerns the purchase of data from third-party companies. Several companies sell their data such as Acxiom (demographic data for direct marketing), Nielsen (audience measurement for TV, radio, music, and newspapers) and Equifax (demographic data on people around the world). However, if limits prevent the purchase of data, an alternative is to use archival repositories that allow acquiring data for free. Archival repositories contain data uploaded and made public online. Among these are the archives of political institutions (e.g., the European Union), and non-profit international organizations (e.g., World Health Organization, OpenPsychometrics, Word Association Lexicons). In these repositories, data owners often provide interfaces that facilitate data retrieval. In Appendix B, I have enlisted some data repositories along with the access requirements and the description of the downloadable data.

Researchers can use other research tools to obtain data. An option is provided by digital crowd-sourcing platforms, which are website through which hire remotely located "crowdworkers" to perform discrete on-demand tasks, such as experiments. Digital crowd-sourcing platforms have been increasingly used by researchers to recruit study participants for online surveys and experiments. In this context, a widely used platform is Amazon Mechanical Turk (AMTurk), which has been described as "the internet's hidden science factory" (Marder, 2015). Some have demonstrated that AMTurk samples have similar or better quality compared to undergraduate samples, which are widely used in psychology (Paolacci, Chandler, & Ipeirotis, 2010; Gabriele Paolacci & Chandler, 2014; Peer, Vosgerau, & Acquisti, 2014) and they map rather closely to American nationally representative sample (Berinsky, Huber, & Lenz, 2012; Coppock, 2019). However, there are about 100,000 Mechanical Turk workers who participate in academic studies each year, the most active of them complete hundreds of studies each month. Participating in multiple studies may entail that the same participants are continuously recycled across researches generating the problem of "non-naivete" (Chandler, Mueller, & Paolacci, 2014), which may be a concern if one uses a fairly common experimental paradigm.

Technology developments have also changed the way in which the traditional research strategies can be performed, allowing the collection of richer information (Adjerid & Kelley, 2018). For example, Qualtrics, which is a tool widely used to administer surveys online, also includes features that collect data on respondents' behavior. It is now possible to record how long it takes to answer questions, capture the number of clicks and time spent on a page, and produce heat maps, which are graphical representations of data that uses a system of color-coding to represent different values, such as where people have clicked on a page, or how far they have scrolled down a page.

All these methods through which acquire data are different from traditional psychological methods such as surveys and laboratory experiments. Indeed, traditional methods allow

researchers to carefully design their studies, determine how to measure variables of interest, and hence to have more control on how data are acquired. With Big Data researchers often need to work with second-hand data collected by others such as social media services, energy retail companies, or archival repositories. Similarly to small data, Big Data are subjected to the same kinds of errors and noises (e.g., missing values, outliers) that, if not identified and corrected, can negatively affect the quality of the analyses results. However, with Big Data, other types of problem may arise. For example, working with Big Data usually entails acquire data from different sources. These datasets must be integrated for being analyzed. However, integrate different datasets can produce redundancies and inconsistencies which can also negatively affect the quality of the results. Another common issue with Big Data is the presence of a high number of variables (i.e., high-dimensionality), which can introduce bias in the analysis. In such a situation, we should understand what variables retain for the analysis or how to transform them to make data less complex. In the next section, I discuss the problems that may arise when dealing with Big Data and illustrate pre-processing methods through which is possible to make them of quality and, hence, ready for the analyses.

2.2. You Cannot Make Something Out of Nothing: Data Preprocessing

Big Data does not entail only great opportunities. It also contains critical elements that could nullify its advantages. These critical issues are widely linked to the quality of the data acquired. After receiving the raw data, the first thing researchers must do is validate them.

Imagine that you want to analyze customers' buying behavior of healthy foods. A supermarket chain provides you with two datasets: One dataset contains information on customers (e.g., age, from how long is a customer), while the other contains all the transactions made by customers in the last year (e.g., item purchased, price, when the item was purchased). For doing the analysis, you must integrate the two datasets. Still, they have different unit levels (e.g., in the first one, a row represents a customer, while in the other, a row represents a transaction). You start to carefully inspect each dataset separately, looking

for interesting variables to include in your analysis. However, you notice that several of the variables for various observations have no recorded value. For your analysis, you would like to include information as to whether each item purchased was advertised as on sale, but you discover that this information has not been recorded. Furthermore, the datasets contain errors, unusual values, and inconsistencies in the data recorded for some observations.

This brief introduction aims to make the reader aware of the fact that "real-world data is a real mess" because data entry and acquisition is inherently prone to errors, which can be either simple or complex. The quality of the data is determined by a series of characteristics, such as completeness, consistency, and accuracy of information (Han et al., 2012). Similarly to what happens with small datasets, we cannot expect to obtain quality results if the data are not qualitatively suitable for analysis. To perform analyses that produce reliable results and from which valid conclusions can be drawn, having data that satisfy the characteristics just outlined is indispensable. Indeed, it is a myth to be able to apply algorithms on raw data and expect them to provide useful insights as output (Lohr, 2014). Then, it comes as no surprise that data preprocessing and transformation is the process that requires the most time and effort.

Data preparation is a mandatory step. It converts raw data into new data that can fits the data mining process. First of all, if data is not prepared, the data mining algorithm may not work. For example, there are techniques (e.g., regression techniques) that require the absence of missing values. Thus, if I would use data with missing value, the model will report errors during its runtime. With other types of data errors, the algorithm will work, but its results will not make sense or will not be considered as accurate knowledge.

To sum up, real-world data is usually dirty, incomplete, and inconsistent. Therefore, data preprocessing techniques are needed to improve the accuracy and efficiency of the subsequent data mining technique used. The rest of the section further describes the basic techniques used to perform the preprocessing of the data set.

2.2.1. Data Integration

As stated above, Big Data may come from different sources. To analyze them, the researcher must integrate the sources into the same dataset. Gathering all the data elements together is not an easy task. If the integration process is not performed properly, redundancies and inconsistencies will soon appear, resulting in a decrease in the efficiency of the subsequent data mining phases and in an increase of errors and distortions of the results.

A variable may be redundant if it can be "derived" from another attribute or set of attributes. Inconsistencies in variable naming (e.g., having the same information in two datasets but with different names) can also lead to the redundancies in the dataset. Some redundancies can be identified through feature selection techniques (for further details, see the dedicated section "Feature selection").

A useful tool for the integration process is building an Entity Relationship Diagram (P. P.-S. Chen & Pin-Shan, 1976), a type of flowchart that illustrates how "entities" such as people, objects, or concepts relate to each other between datasets.

2.2.2. Outlier Identification

Outliers are defined as extreme and atypical values on a single variable (univariate outliers), or atypical combinations of variable values (multivariate outliers), which arouse suspicions that they are generated by a different mechanism (Hawkins, 1980).

Outliers may be induced in the data for a variety of reasons. They may be the results of human errors or instrument errors. They may have been created unintentionally through data manipulation or data sampling. In some other cases, they may not represent an error rather a novelty. Identifying the reason for outlier existence is relevant for determining the actions to take. If outliers are the results of errors, then outlier removal is the way to go. Even though some data mining, as for example decision trees, can tolerate outliers, with many other techniques modeling noisy data may deliver unreliable results. In those cases, outliers can be removed before performing any data analysis through either noise removal (Teng, Chen, &

Lu, 1990) or noise accommodation (Rousseeuw & Leroy, 1987). If outliers are recognized as novelties, identifying those values becomes necessary for unveiling new phenomena (novelty detection; Markou & Singh, 2003). Novelty detection plays a central role in many applications, including fraud detection (van Capelleveen, Poel, Mueller, Thornton, & van Hillegersberg, 2016), illegal parking detection (Xie, Wang, Chen, Shi, & Zhao, 2017), and patients monitoring and alerting (Hauskrecht et al., 2013). Novelty detection represents great opportunities in many fields, also in psychology. Indeed, if we have enough data to be able to detect novelty, we may face the opportunity to study new phenomena. For example, Pan and colleagues (Pan, Zhou, Liu, & Wang, 2019) have recently developed a model based on novelty detection to capture anomaly behaviors in the crowds recorded in video (e.g., prisons fight).

The most straightforward methods for identifying outliers are visualization tools (e.g., histograms, box plots and scatter plots), capping methods (e.g., points that fall out of the range of 5th and 95th percentile), and statistical indicators (e.g., points above three standard deviations from the mean value). However, there are also more advanced methods, such as Extreme Value Analysis, Statistical Modeling, Regression models, Proximity-based models, and Information-theoretic models (Aggarwal, 2015).

2.2.3. Missing Values

Missing data are records that were intended to be obtained but for some reasons were not. The problem of missing data is relatively common in almost all research (traditional and Big Data) and can have a considerable negative effect on the conclusions that can be drawn from the data.

There are various methods through which handling missing data. The choose of a method over the other should be directed by the understanding of what mechanism has caused a value to be missing (Little & Rubin, 2002). Despite the name, the mechanism is not a causal explanation for missing data. Indeed, a missing data mechanism represents the statistical

relationship between observations and the probability of being missing. There are three mechanisms:

- Missing Completely at Random (MCAR) occurs when the probability of missing values depends neither on the variable values nor on the values of other variables.
- Missing at Random (MAR) occurs when the probability of missingness depends on the observed values of other variables. For example, the presence of missing values in an income variable may not be linked to the value of the income itself, but, for example, to education. It is indeed possible that people with a higher level of education are more reluctant to disclose their income than those who have a lower level of education.
- Missing Not at Random (MNAR) occurs when the probability of missingness depends on the value of the variable with missing. This mechanism is common in situations where people do not want to reveal personal or embarrassing information about themselves. For example, it is less likely that a person who earns a lot or very little will reveal his income compared to those who earn an average sum. As a result, the income will be missing due to its value.

In general, there are three ways through which handling missing values: elimination, imputation, and the use of robust analytical techniques. Elimination is probably the most common approach for handling missing data. The techniques through which eliminate missing are listwise and pairwise elimination. Listwise elimination (or complete case analysis) concerns the exclusion from analysis of an entire record if any single value is missing. Despite its popularity, the indiscriminate use of this method could introduce errors in the results and should, therefore, be discouraged. If we believe our missing are MCARs, then we can use this method without concerns (Donner, 1982). Since MCAR represents a random subset of the data set, their elimination does not affect the results we can obtain and the conclusions we can draw from them. The pairwise elimination (i.e., available case analysis) removes the specific missing values from the analysis, not the entire record. Compared to the

listwise elimination, pairwise elimination has the advantage of maximizing the number of the sample to be used. The major drawback of pairwise elimination is the distorted estimation of the model parameters because they will be based on different sample sizes and different standard errors.

Imputation methods concern the replacement of missing values with plausible alternatives obtained from the data, from external sources, or the combination of both, following established rules and methods. Imputation methods aim to reduce the distortions introduced by the presence of missing data. There are various types of imputation methods, and generally, their use is recommended for MAR and MNAR missing. A first imputation method provides that the missing values are replaced with the central tendency value of the variable (e.g., average, median, fashion, according to the variable measurement scale). This method is widely used due to its simplicity. However, this method could lead to the reduction of variability within the variable and, thus, model performance. Another imputation method involves replacing the missing value with the value of the variables in observations similar to the one with the missing value (hot-deck imputation). The main problem of this method lies in the definition of the concept of "similar". A further imputation method concerns the replacement of missing with a value predicted by a statistical model (e.g., regression, decision tree, random forest) whose predictors are the other variables present in the dataset. The advantage of this method lies in the leveraging of the existing relationships between the variables. However, the imputation through predicted values tends to distort variances and covariances, and hence, to amplify multi-collinearity between variables. A possible strategy to avoid this shortcoming, and at the same time, allow uncertainty, consists of using multiple imputations methods (Rubin, 1987) to create several different plausible imputed data sets and combining the results obtained from each of them (Little & Rubin, 1987). Thus, imputed values are not random guesses of what a particular missing value may be. Instead, multiple imputation intends to create educated guesses of data values that maintain the overall

variability in the population while preserving relationships with other variables. A disadvantage of multiple imputation is that it assumes the data to be MAR (i.e., the missing values must be predictable from the values of other variables). Nevertheless, multiple imputation has been studied extensively, and most scholars agree that it is a compelling technique (Donders, van der Heijden, Stijnen, & Moons, 2006; Graham, 2009; Rubin, 1976; Tsikriktsis, 2005; van Ginkel, van der Ark, & Sijtsma, 2007).

Both imputation and elimination assume, to some extent, that missing values are errors to get rid of. However, the fact that values are missing may in itself carry valuable information, besides the specific value, the record would have taken had it been recorded. In data mining, it may be useful to have indicators (i.e., dummy variables) for the missing values because they may be predictive. When a variable contains missing values, each observation will assume either the value 0 (i.e., the value is observed) or 1 (i.e., the value is missing). Missing indicators should be created for each variable with missing values, and then all the indicators will be added to the data mining model.

The third method through which handling missing values is the use of analytical techniques that are robust to the presence of missing values. In the realm of data mining, there are techniques, such as decision trees, that can handle the presence of missing values in the datasets. Thus, they require neither elimination nor imputation of missing values.

Dealing with missing is very common in all kinds of research. Even when Big Data is considered, the problem of the missing data handling is still of primary importance for the successful implementation of data mining techniques. Usually, the probability of having missing data increases with the number of features in the dataset and with the number of samples, making the imputation task in Big Data context extremely important (Petrozziello, Jordanov, & Sommeregger, 2018). If the number of missing values systematically grows with the size of the dataset, the imputation is necessary to preserve, or even increase, the statistical power of the data or, in general, to not lose too many samples during the pre-processing stage.

2.3. Data Transformation

Ending up with a consistent and almost error-free data set does not mean that the data is in the best form for a data mining algorithm. Data transformation techniques have been devised to adapt a dataset to the needs of the data mining algorithm that will be applied afterward. For example, some data mining algorithms work much better with normalized attribute values, such as Artificial Neural Networks or clustering algorithms. Therefore, it is common to transform the original to generate new attributes with better properties (e.g., normalization) or to reduce the complexity of the information enclosed into the dataset (e.g., data reduction). All these transformations make the data suitable for data mining when using data in their original format may be computationally or inferentially intractable.

2.3.1. Normalization

Often, numerical variables are measured on scales that are entirely different from each other. For some data mining algorithms, such differences in the ranges will lead to a tendency for the variable with greater range to have undue influence on the results. This happens, for instance, to neural networks, and to algorithms that make use of distance measures, such as clustering techniques. To avoid these potential detrimental effects, the original data can be normalized, or in other words, transformed without distorting values distribution. Generally, variables are scaled to fall within a smaller range, such as from -1 to 1, or from 0 to 1. Commonly used methods for normalizing data are the Min-Max, the Z-score standardization, and the Decimal Scaling normalizations (Han et al., 2012).

2.3.2. Feature Selection

Collecting large datasets nearly always entails collecting a wide range of variables. However, only a subset of them will be useful for the data mining analysis we want to perform. Keeping irrelevant attributes or leaving out relevant attributes may be detrimental. This is why selecting a subset of variables is important. By definition, feature selection aims

to find a minimum set of important variables in the data set and discard others that are redundant or irrelevant for the task. Mining on a reduced set of attributes has many benefits, such as increasing the reliability of the data mining models, improving computational efficiency (e.g., reducing storage requirements and computational cost), and improving the understanding of the data and the results of the model.

But, how do we know which features are relevant and which are not? Feature relevance can be determined in two ways: domain knowledge-driven feature selection and data-driven feature selection. Since KDD processes are almost data-driven, it sounds weird to select variables based on the knowledge we possess on the topic we aim to study. However, using just a data-driven process can result in models that may not correspond to true relationships present in the data set due to overfitting (Groves, 2013). Thus, theoretically driven feature selection may be useful because keeping information that is known a priori can facilitate the development of a more reliable model and a better understanding of its results. For example, Pan and colleagues (2019) build a novelty detection model for identifying anomalies in crowd's behavior. They showed that model was robust because they were able to integrate social psychology knowledge in the selection of the features (e.g., conformity behaviors and collective behaviors)

Feature selection methods based on data-driven techniques can be divided into filter, wrapper, and embedded methods (Guyon, Bitter, Ahmed, Brown, & Heller, 2005).

Filter methods select variables on the basis of their scores in various statistical tests for their correlation with the outcome variable. For example, a businessman aims to predict which customers will accept a cross-sell proposal. He collects a dataset that contains hundreds of variables and decides that he wants to use only those variables that may have a relationship with the behavioral outcome. Thus, he performs some correlation analyses to select only the variables that show to be correlated with the outcome. Various statistical means can be used to determine predictive power, such as correlation, chi-square statistic, latent discriminant

analysis, and ANOVA. The variables that demonstrate to be linked with the outcome are kept in the subsequent analysis.

Wrapper methods (Ron Kohavi & John, 1997) build models with a certain subset of features and evaluate the importance of each feature. Then they iterate and try a different subset of variables until the optimal subset is reached. Compared to filter methods, which measure the relevance of features by their correlation with the dependent variable, wrapper methods measure the relevance of a subset of features by actually build a data mining model on it. Wrapper methods are usually computationally very expensive, especially when the number of variables is substantial. Common wrapper methods include Forward Selection, Backward Elimination, and Recursive Feature Elimination. The forward Selection method is an iterative method that starts with having no feature in the model. In each iteration, we keep adding the feature which best improves our model until the addition of a new variable does not improve the performance of the model. Backward Elimination methods start with all the features and remove the least important feature at each iteration, which improves the performance of the model. We repeat this until no improvement is observed on the removal of features. Recursive Feature Elimination aims to find the best performing feature subset by repeatedly creating models and keeping aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

Embedded methods combine the qualities of filter and wrapper methods. Embedded methods perform feature selection as part of the model construction in the data mining phase. A well-known method is the Least Absolute Shrinkage and Selection Operator (LASSO) method for constructing a regression model, which penalizes the regression coefficients, shrinking many of them to zero. In this way, any variables which have non-zero regression coefficients after the penalization are “selected” by the LASSO algorithm.

2.3.3. Data Reduction

Data reduction techniques can be applied to achieve a reduced representation of the data set, which is much smaller in volume and tries to keep most of the integrity of the original data (Han et al., 2012). The aim is to provide the data mining process with a mechanism to produce the same (or almost the same) model outcome when it is applied over reduced data instead of the original data, and, at the same time, make the process more efficient. Usually, data reduction methods are categorized into three families: dimensionality reduction, sample numerosity reduction, and cardinality reduction.

2.3.3.1. Dimensionality reduction

Dimensionality reduction methods aim to overcome a major problem in mining large data sets, which is the "curse of dimensionality" (Bellman, 1961). Curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of variables) that do not occur in low-dimensional settings.

High dimensionality (i.e., a high number of variables) can become a severe obstacle for the efficiency and effectiveness of most of the data mining algorithms because of their computational complexity. High dimensionality increases the size of the search space exponentially and also increases the chance to obtain invalid models (Hwang, Lay, & Lippman, 1994). Moreover, large datasets require many computational skills that would require a long time to be mined or, at worst, make the analysis infeasible. Dimensionality reduction ensures the reduction of the number of attributes or random variables in the data set through two different methods. On the one hand, there is the feature extraction and selection (see the "Feature Selection" section), and the transformation or projection of the original data onto a smaller space. To alleviate this problem, many dimension reducers have been developed over the years, some of which are linear methods, such as Principal Component Analysis (Dunteman, 1989), Factor Analysis (Jae-on. Kim & Mueller, 1978) and Multidimensional Scaling (Kruskal, 1964); some others are non-linear, such as Locally Linear

Embedding (Roweis & Saul, 2000), and Isometric Feature Mapping (Tenenbaum, de Silva, & Langford, 2000).

2.3.3.2. Sample numerosity reduction.

Sample numerosity reduction methods replace the original data by an alternative smaller data representation. They include data sampling methods (e.g., simple random sample without replacement and stratified sampling) and different forms of grouping (e.g., data condensation and clustering). Data sampling methods aim to reduce the number of observation submitted to the data mining algorithm, support the selection of only those target cases to use in the data mining phase, assist regarding the balance of data and occurrence of rare events (i.e., class imbalance problem), and divide a data set into two or three data sets to build and validate data mining model (see Validation Process and Methods for further details).

2.3.3.3. Cardinality reduction

Data cardinality refers to the number of distinct values in a variable. Generally, categories are supposed to be mutually exclusive and unrelated, with a rather small, fixed, known set of possible values. However, in many datasets, variables do not encase a small number of categories: The set of all possible categories may be huge and not known a priori, and often carry some morphological or semantic links (Cerda & Varoquaux, 2019). The variables with many distinct values are called high-cardinality variables (e.g., ZIP codes). In general, it is helpful to reduce the number of categories in a variable as it decreases the complexity of that information, especially in predictive modeling, where many algorithms hardly cope with such a high number of distinct values. For example, imagine that the ZIP code variable has a thousand distinct values. To reduce the cardinality, we may consider group distinct values according to the region where the ZIP codes are located. This can reduce the number of distinct values from a thousand to a few categories. There are several methods for dealing with high cardinality. The most straightforward way to handle high cardinality is binning classes into a smaller number of classes through semantic grouping, which aims to aggregate

all the distinct values into the same higher-ordered category (e.g., like ZIP codes and regions). Another way to reduce the number of categories implies the transformation of the categorical variable into a continuous variable through transformation, such as the Weight of Evidence (Hand & Henley, 1997), Supervised Ratio (Moeyersoms & Martens, 2015), and the Perlich Ratio (Perlich & Provost, 2006).

2.4. May Data Mining Be with You

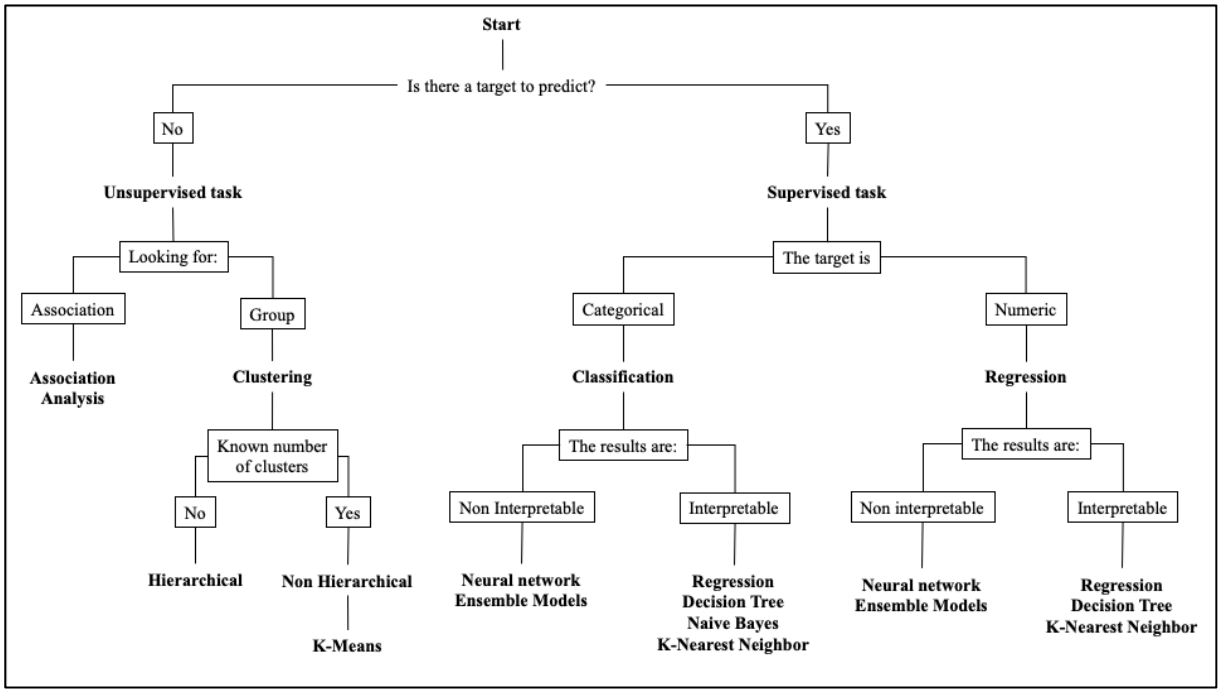
The developments of data mining and the various fields in which it can be applied, make data mining a rather vague concept characterized by various definitions. Some, indeed, define data mining with more generic words (e.g., extraction of information through different technical approaches; Frawley, Piatetsky-Shapiro, & Matheus, 1992), others refer to it according to the context where it finds the application (e.g., a business process; Linoff & Berry, 2011). What remains unchanged in those definitions is the reference to the approaches, tasks, and techniques that constitute data mining.

Even though data mining can be used for top-down analysis (i.e., verification-driven data mining), data mining methods emphasize the importance of bottom-up analysis where researchers start digging into the data in search of previously unknown information (Han et al., 2012), as mining gold from the earth. In order to extract this new knowledge from data, data mining uses machine learning techniques (Bishop, 2006). The bottom-up approach represents a considerable opportunity for psychologists to open their practices to different analytical and computational approaches through which derive structure and meaning from the massive amounts of data.

2.4.1. Data Mining Tasks and Algorithms

When a decision has to be made about which type of algorithm to use to model the data, certain information must be considered. One concerns the nature of the data to be analyzed: What variables to consider, what are the characteristics of these variables, and whether or not a target variable is present. Other information concerns the objectives that must be pursued and the resources (e.g., computational resources and technical skills) that are available. From this introduction, it is clear that the choice of the task to be performed and, therefore, of the algorithm to be used, must be guided by a decision-making process that leads to the best option or set of options (Figure 2.2).

Figure 2.2. A possible decision-making process for choosing the algorithm.



In this section, I describe the options available, outline their characteristics, and discuss some methodological issue that psychologists approaching data mining and machine learning should think about.

2.4.1.1. Supervised Learning

Supervised learning methods attempt to discover the relationship between independent variables and a dependent variable (Maimon & Rokach, 2010). To be performed, they require

a dataset containing one or more predictors and a dependent variable to predict. The purpose of supervised learning techniques is to create a model that is able to predict a specific outcome in the future. Starting from a set of data in which the outcome is known, the supervised learning model learns rules and patterns that determine the occurrence of a given outcome.

Numerous supervised learning techniques can be used in the context of psychological research, for example, predicting the effectiveness of a treatment or identifying profiles at risk of mental disorders. Supervised learning techniques can perform two types of tasks: regression and classification. The main difference between the two methods lies in the type of outcome variable that should be predicted. If the outcome variable is categorical, the supervised task is called classification. More specifically, it is a binary classification when the outcome variable takes two distinct values and multinomial classification when it has more than two distinct values. If the outcome variable is numerical, the task is called regression. Although classification and regression are different methods, most machine learning algorithms manage both of them. In table 2.1, I have enlisted some of the most used techniques and compared them according to some characteristics.

Table 2.1. List of some of the most used supervised techniques and their characteristics

	Algorithm						
	KNN	**Naive Bayes**	**Regression**	**Decision Tree**	**Ensemble**	**Neural Networks**	**SVM**
Parametric or Non Parametric	Non-Parametric	Parametric	Parametric	Non-Parametric	Non-Parametric	Non-Parametric	Non-Parametric
Assumptions on data	None	Attribute values independence	Many assumptions	None	None	None	None
Task Type	Classification	Classification	Classification and Regression	Classification and Regression	Classification and Regression	Classification and Regression	Classification and Regression
Ease of implementation	Very easy	Very easy	Very easy	Easy	Mildly easy	Mildly easy	Mildly easy
Interpretability	Inherently Interpretable	Inherently Interpretable	Inherently Interpretable	Inherently Interpretable	Post-hoc Interpretable	Post-hoc Interpretable	Post-hoc Interpretable
Dataset size	It works well even on small datasets	It works well even on small datasets	It works well even on small datasets	Works best on large datasets	Works best on large datasets	Works best on large datasets	It works well even on small datasets

Feature type	It performs better on numerical than categorical variable.	It performs better on categorical than numerical variable.	Work well on both categorical and numerical variables	Work well on both categorical and numerical variables	Work well on both categorical and numerical variables	Work well on both categorical and numerical variables	Work well on both categorical and numerical variables
Normalization	Depends on the distance metric	Not needed	Yes	Not needed	Not needed	Helps the convergence of training	Helps in solving linear equations
Sensitivity to outliers	High	High	High	Low	High	Low	Low
Missing data handling capability	Need complete data	Need complete data	Need complete data	Yes (e.g., surrogate splits)	Need complete data	Need complete data	Need complete data

Following I provide a brief description of those techniques.

- Regression consists of a set of methods that allow predicting the value of an outcome variable based on one or several predictors. Regression have been used in psychological research for predicting. An example of the use of such technique in psychological research, is provided by Swartz and colleagues (Schwartz et al., 2014). They built a regression model to predict the degree of depression of Facebook user. They combined 28,749 users' Facebook status with survey data. They found that user mood worsens in the transition from summer to winter. Generally, the goal of the regression model is to build a mathematical equation that defines the outcome variable as a function (linear for regression task and sigmoid for classification task) of the independent variables. Next, the derived equation can be used to predict the outcome based on new values of the predictor variables. Regression techniques are widely used because they are very easy to implement, efficient to train, do not require too many computational resources. Moreover, regression techniques can offer a good baseline for judging the quality and for debugging more sophisticated approaches (Kosinski, Wang, Lakkaraju, & Leskovec, 2016). Moreover, results are highly interpretable. A major shortcoming of simple regression is its reliance on the many assumptions on independent variables (i.e., absence of multicollinearity, independent variables and residuals are uncorrelated, and independent variables are expected to be

normally distributed), the dependent variable (in linear regression the dependent variable should be normally distributed, while in logistic regression it must be binary), their relationship (i.e., linear regression expects the relationship between dependent and independent variables to be linear, whereas logistic regression expects the independent variables to be linearly related to the log odds), and residuals (i.e., normality, the residuals mean should be zero, and homoscedasticity). The violation of the assumptions can be, in some cases, circumvented by using a different form of regression. For instance, when the linearity assumption is violated, it may be suitable to use non-linear regression techniques, such as polynomial or spline regression (e.g., MARS regression). A further shortcoming of simple regression concerns the difficulty of dealing with many variables. In such a context, the simple regression technique is often outclassed by its regularized counterparts (LASSO, Ridge, Lars, and Elastic-Net regressions). Regularization is a technique that discourages learning a more complex or flexible model by setting a penalty term that shrinks the coefficient estimates towards zero. The variables that show zero estimated coefficients are excluded from the model. The disadvantage of penalized regressions is that they require the tuning of the penalty strength.

- K-Nearest Neighbor (KNN) is a non-parametric algorithm, easy to understand, and equally easy to implement. KNN assumes that similar things exist in close proximity. In other words, similar things are near to each other. Thus, to predict the value of an unknown observation, KNN uses the values of the K observations that are near to that point. What defines “near” depends on the distance metric one chooses to adopt. There are different types of methods for calculating distance (e.g., Manhattan distance, Euclidean distance, Hamming distance, Minkowski distance) and the choice of a specific method depends on the complexity of data on which the classification task takes place and on the type and the nature of attributes (e.g., numerical vs. discrete and numeric vs. string). Typically, the application of this classifier is effortless and intuitive for spaces with two or three

dimensions, while it is complicated for complex multivariate spaces. The most critical aspect of KNN is the selection of the parameter K. This choice is crucial because it indicates the number of elements that must be taken into account when making predictions. If K is too small, the classification is likely to be sensitive to noise; conversely, if K is too large, the classification can be computationally expensive. Despite its simplicity, the KNN algorithm has never been used in psychological research with Big Data.

- Bayes is one of the most efficient inductive machine learning algorithms based on the Bayes Theorem. It is used in a wide variety of classification (binary and multinomial) tasks, especially text classification because it is fast and easy to implement (Xu, 2018 Bayesian Naïve Bayes classifiers to text classification). In psychological research, Naïve Bayes have been used for classifying infants into higher risk group at Autism Spectrum Disorder (ASD) or control (Bosl, Tierney, Tager-Flusberg, & Nelson, 2011) or to forecast the engagement levels of social media posts (Hwong, Oliver, Van Kranendonk, Sammut, & Seroussi, 2017). Naïve Bayes classifier describes how the probability of an event is affected by the prior knowledge of conditions. Naïve Bayes algorithm assumes that the effect of a variable on a given event is independent of the values of the other variables. This assumption, called conditional independence of classes, aims to simplify the calculations, which explains why the algorithm is called "naïve". The violation of this assumption can extremely limit the ability of the model to classify new instances correctly and, in reality, this assumption is nearly always violated. This limit can be overcome through the use of other algorithms, such as the Bayesian networks (if one wishes to stick to Bayes Theorem) which is a type of probabilistic graphical model that uses Bayesian inference for probability computations. However, when the assumption of independence holds, a Naive Bayes classifier can perform better than other models, like regression, and can achieve better performance with less training data. A further disadvantage is represented by the chance of occurring into the "zero conditional probability problem", that

emerges when there is no frequency in a class of a categorical variable, and thus, the estimation of the posterior probability will be equal to zero. As a consequence, the model is unable to make predictions. There are several sample smoothing techniques to overcome the zero conditional probability issue, such as "Laplacian Correction" (Manning, Raghavan, & Schütze, 2008). Beside its limits, the Naïve Bayes model is easy and fast to build, with no complicated iterative parameter estimation (Rezzani, 2013), and it is an interpretable model because the contribution of each variable can be inferred from the estimated posterior probability.

- Decision Tree methods are a set of nonparametric methods that have significantly increased in popularity because they closely resemble human reasoning, are easy to understand and interpret (Kotsiantis, 2013). Decision trees have been used in psychological research for identifying patients with higher risk of committing suicide (Kessler et al., 2017) or predicting behavioral and emotional problems in children (Ahmed, Afzal, Siddiqi, Amjad, & Khurshid, 2018). A decision tree is a diagram of decision-making rules used either for classification or prediction tasks. A tree consists of branches and leaves in which the value of the instances is predicted downstream. A decision tree predicts the outcome value by posing a series of questions about independent variables. Each node contains an "if-then" question and every node point to one child node out of each possible answer to that question. Thus, each instance follows a path from the root node to the leaf according to the answers that apply to that instance. Decision tree algorithms are attractive because they do not require any assumption of linear relationships, normal distribution, or homoscedasticity (Jinhwa Kim, Won, & Bae, 2010), they are not influenced by data transformation, and they are resistant to outliers (Breiman, Friedman, Olshen, & Stone, 1984; Steinberg, 2009). Furthermore, tree-based methods can handle missing values without requiring imputation (i.e., they use surrogate splits) and handle heavily skewed data without requiring data transformation (Song & Lu, 2015). A decision tree is a diagram

of decision-making rules used either for classification or prediction tasks. Despite the decision trees are highly straightforward to understand, they have some disadvantages. Firstly, decision tree models tend to overfit, and therefore they may not generalize to new data. Secondly, they may show high variance (i.e., they can get unstable due to small variations in data). Thirdly, preparing decision trees, especially large ones with many branches, is a complex and time-consuming affair. Finally, too complex trees may have detrimental effects on model interpretability, since large trees are not intelligible and pose visualization difficulties. There are different types of decision tree and, probably, the most known and used algorithms for building a decision tree are CART, C5.0, and its precursor C4.5 (Rokach & Maimon, 2015).

- Ensemble algorithms liken human behavior under decision making. In which a situation, we hardly base a decision on the judgment of a single expert, and typically prefer to subject the problem to different experts who provide various solutions. Drawing a parallel with data mining, it is possible to compare the model built through a single model with the advice of one single expert, while an ensemble can be compared to an assembly of experts who decide together the solution to be adopted. In psychological research, ensemble learning has been adopted, for example, for predicting psychological traits (i.e., Big Five personality traits, materialism, and self-control) from online spending transactions (Gladstone, Matz, & Lemaire, 2019). The authors used an ensemble model (i.e., a Random Forest) rather than a single model (e.g., decision tree) because ensemble model usually provides better predictions and higher generalization than single models (Dietterich, 2000). Although ensemble model can outperform single classifiers in many situations, ensemble classifiers require the presence of enough diversity in the base classifiers to ensure good performance (Kuncheva, 2004). The aim is to combine single models that make as few errors as possible, but they should make different errors to be able to learn from each other. When nearly identical single models vote for a prediction, indeed, they would all agree and

behave like a single model. Bagging (Breiman, 1996) and boosting (Freund & Schapire, 1997) are two well-known techniques that adjust the original training data to include diversity in the training process. The former method uses different training sets for every component classifier in parallel, while the latter method trains multiple classifiers in sequence on training sets using different weighted observations. The most well-known algorithm that uses the Bagging method is Random Forest (Breiman, 2001). Random Forest is an ensemble of decision trees. Each tree in the forest makes its predictions that are further aggregated to make one weighted prediction. The random forest takes advantage of the sensitivity of the decision tree to changes in data by allowing each tree to randomly choose variables from the dataset with replacement. The most well-known algorithms that apply the Boosting method are AdaBoost (Freund & Schapire, 1997), Gradient Boosting Machine (Friedman, 2001), and Extreme Boosting Machine (Chen & Guestrin, 2016).

- Artificial Neural Network is a set of algorithms inspired by the structure and functioning of brain neurons. Neural networks can be used for both classification and regression tasks and can model nonlinear data. They can provide reliable results also in complex multidimensional spaces. In a neural network, nodes are units capable of efficiently processing the signals they receive from other nodes. The nodes are located on different layers that are connected with each other (i.e., interlayer connection). Each connection is estimated by an iterative weighting procedure. The simplest structure of a neural network consists of three layers of nodes that can be classified into three categories: The input nodes are the independent variables, the output node is the dependent variable, and the hidden nodes are latent variables. Neural networks with three layers (input, one hidden, and output layers) are called shallow neural networks. The most known neural network algorithms are the Single-Layer Feedforward Neural Network, and the Radial Basis Function Neural Networks. Neural networks can have multiple hidden layers, and more

hidden layers result in higher flexibility of the model. Neural networks with several hidden layers are called deep learning networks (Deng & Yu, 2014). In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. By further advancing into the neural network, the more complex the features our nodes can recognize, since they aggregate and recombine features from the previous layer (i.e., feature hierarchy). Besides providing greater flexibility to the model, using too many hidden layers makes the model more prone to overfitting. Thus, the optimal number of hidden nodes needs to be determined by cross-validation (see Cross-Validation Process and Methods section). Deep neural networks have produced major breakthroughs in computer vision, natural language processing (LeCun, Bengio, & Hinton, 2015) and time series forecasting (Gamboa, 2017), because of their ability to model unstructured data such as images and texts. However, whereas neural network have been already used in psychological research, such as forecasting individuals' mood (Mikelsons, Smith, Mehrotra, & Musolesi, 2017; Taylor, Jaques, Nosakhare, Sano, & Picard, 2017), classifying social interactions (Aghaei, Dimiccoli, Canton Ferrer, & Radeva, 2018; for a review, see Vieira, Pinaya, & Mechelli, 2017). The most known deep neural network algorithms are Convolutional Neural Networks, Deep Boltzmann Machine, and Recurrent Neural Networks. An example of the application of Deep neural networks in psychological research is the prediction of sexual orientation from facial images (Wang & Kosinski 2018).

- Support Vector Machines (SVM) became extremely popular around the time they were developed in the early 1990s, and they remain the go-to method for a high performing algorithm with little tuning efforts, good generalization performance and ability to handle high dimensional data. In psychological research, SVMs have been successfully applied in the classification of relevant emotional states from the real-life spoken human-human interactions (Devillers, Vidrascu, & Lamel, 2005), the classification of eyes' iris positions

to predict personality traits (Ramli & Nordin, 2018). Although they can perform well both on classification (support vector classification) and regression tasks (support vector regression), SVMs are more used for addressing classification problems. In such a task, SVMs try to find a separating hyperplane that separates data with the largest margin. The margin of the hyperplane is defined as the shortest distance between instances of separate classes that are closest to the hyperplane (i.e., the support vectors points). The intuition behind searching for the hyperplane with a large margin is that a hyperplane with the largest margin should be more resistant to noise than a hyperplane with a smaller margin. If the linear hyperplane exists, all the instances at one side of the hyperplane will belong to one class, and all the instances on the other side will belong to the other class. If the linear hyperplane does not exist (e.g., nonlinear, and high dimensional space data), SVM address non-linearly separable cases by introducing two concepts: Soft margin and Kernel trick. The former tries to find a line to separate but tolerate few misclassification errors, while the latter tries to find a nonlinear decision boundary through nonlinear functions such as polynomial, radial basis or sigmoid functions.

2.4.1.2. Unsupervised learning

Unsupervised learning methods aim to group variables or observations based on their degree of similarity or covariation. Thus, in unsupervised learning, there is no outcome variable that we wish to predict. Unlike supervised learning methods, unsupervised learning is commonly used in psychological research. For example, data reduction methods, such as Principal Components Analysis and Exploratory Factor Analysis (see Data Reduction section), are quite common in psychology as are methods for grouping participants, such as cluster analysis and finite mixture modeling. For certain aspects, unsupervised learning is more complicated than supervised learning. For example, compared to supervised learning, which results can be estimated through well-accepted evaluation measures, the performance

of the unsupervised model cannot be mathematically estimated (Tan, Steinbach, & Kumar, 2005).

Cluster analysis and Association Rules are two of the most widely used techniques in unsupervised learning:

- *Cluster analysis***.** Cluster analysis is the most used descriptive data mining method. Clustering algorithms examine data to find groups of items that are similar (Bramer, 2013). A cluster is, therefore, a set of objects that are similar to one another within the same cluster, but which are dissimilar to the objects present in other clusters. Thus, to achieve a good representation of the data, cluster analysis should maximize the differences between clusters and maximize the internal homogeneity within the cluster. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. Clustering is either done hierarchically or non-hierarchically. One type of hierarchical clustering (agglomerative clustering) consists in initially considering each observation as a cluster, and then repeatedly linking pairs of clusters based on similarity until every data object is included in one unitary cluster. Another type of hierarchical clustering can also be performed as a divisive process (divisive clustering): In this case, all the elements are initially in a single cluster that is gradually subdivided into sub-clusters. Nonhierarchical methods require that the user pre-specify the number of clusters desired. Often, when nonhierarchical methods are employed, a hierarchical cluster analysis is carried out beforehand to specify the starting number of clusters for the nonhierarchical algorithm. However, hierarchical clustering may become unfeasible with very large datasets. The most commonly used cluster algorithm in data mining is k-mean (James, Witten, Hastie, & Tibshirani, 2014). Clustering has been widely used in psychological research. In particular, speaking about Big Data, it has been applied to detect structures in digital data (Eisenberg et al., 2019; Stachl et al., 2019). However, clustering methods may help psychologists in finding relevant relationships among study participants that often are

assumed a priori (Moustafa et al., 2018). For example, Crouse and colleagues (Crouse, Moustafa, Bogaty, Hickie, & Hermens, 2018) used an agglomerative clustering to subtype psychosis-prone individuals. As stated above, this approach assumes each element has its own cluster, and then clusters are merged based on similarity in a hierarchical manner. Results show that there are three clusters, which differ in IQ and social functioning. Future research should use similar methods to subtype participants instead of using a priori (assumed) taxonomy.

- *Association rules.* Association rule algorithms are used to search for recurring configurations within data. For example, consider a university psychology department that maintains a database of courses taken by its students. Suppose the department wants to discover popular combinations of courses taken by its students and finds, for example, that social psychology and social cognition courses tend to be taken together. Such information could be used for curriculum planning to drop low-demand courses or introduce courses related to popular ones. Despite that an associative rule tells whether two or more elements of the dataset are associated or not, it tells us nothing about the nature of this association. There are many possible association rules derivable from any given dataset, most of them of little or no value. Thus, it is usual for association rules to be stated with some additional information indicating how reliable they are, for example, the support (i.e., $P(A \cap B)$) of such rule and its confidence (i.e., $P(A \cap B)/P(A)$) (Nabareseh, 2014). The higher the support and confidence, the more trust one can have on the results. Algorithms such as Bayesian classifiers, FP-Growth, and especially Apriori algorithms are the most used in association rule mining.

2.4.2. Validation Process and Methods

Researchers build supervised models when they want to predict the outcomes of future data. To check if the model will be able to make accurate predictions once placed in a new context, we need to submit it to a validation process. The validation process prevents the risk

of using a model that may have overfitted the data. Overfitting happens when the trained model has learned idiosyncratic relationships that might not hold when the model is applied to data that is similar, but not identical, to that used to construct it (Hastie, Tibshirani, & Friedman, 2009).

However, it must be said that overfitting is only a flip of the coin. Underfitting a model to the data is also an important issue. Underfitting occurs when a model is not able to account for the true complexity (e.g., non-linear effects and interactions) in the data and, therefore, the model cannot grasp the systematic variance. Underfitting will cause lower predictive performance on new data than an adequately flexible model could achieve.

For preventing the risk of overfitting and underfitting, we need validation. Validation consists of a set of techniques for subdividing data into complementary sets to obtain an independent, objective, and unbiased estimate of learned model performance (Browne, 2000; Han et al., 2012; Kohavi, 1995). Usually, validation requires the splitting of the dataset into sub-datasets, and to test different built models and to select the one with the best predictive performance on unseen data. Dataset splitting can be performed in multiple different ways. Some of the common ways to perform validation are:

- Holdout validation is the most known, most used, and most straightforward method for validating a model. Holdout validation concerns the subsetting of the dataset in two parts: the training set, which is used to build the model, and the test set, which is used to prove its predictive ability. To assess how well the supervised model would perform on new unseen data, the test set must not be used in any way to create a classifier. There is no definitive rule on which is the best proportion with which to divide the dataset. In general, it is preferable to build the model on 70% / 80% of the cases and test it on the remaining portion, since the more data we give to the algorithm to build the model, the more are the possible patterns that could be available to learn. There are many ways to split the data into training and testing sets. The most recommended approach is to use some version of

random sampling (e.g., completely random sampling, stratified random sampling). Some algorithms involve two building-stages, one to come up with the basic structure of the model and the second to optimize the parameters involved in that structure. In such cases, model construction and test must involve three datasets instead of two: The training set to build the basic structure, the validation set to optimize the parameters, and the test set to evaluate the performance of the final, optimized model.

- K-fold cross-validation retakes the underlying logic of the holdout validation, but instead of building the model on one fixed train-test split, it creates the model multiple times based on different train-test combinations. Thus, with k-fold validation, the original dataset is divided randomly into k parts (i.e., folds) of a similar size. In general, K-fold cross-validation is performed by taking one-fold as the test data set, and the other k-1 folds as the training data, fitting and evaluating a model, and recording its prediction performance. This process is repeated with each fold as the test data, and all the scores averaged to obtain a more comprehensive model performance. K-fold cross-validation, as the name let intends, involves the setting of the parameter k. The value for k is chosen such that each train/test group of data samples is large enough to be statistically representative of the broader dataset. The choice of the k value is not clear cut. However, the most commonly used value for k is the ten-fold cross-validation (10-fold cross-validation).
- Leave-one-out cross-validation is K-fold cross-validation taken to its logical extreme where K is equal to N, which is the number of cases in the dataset. Thus, the model is trained on all the instances, and one instance is used as a test set. The results of all N predictions, one for each instance of the dataset, are averaged, and that average represents the predictive performance. This procedure is attractive because the highest amount of data is used for training the model, and no random sampling is involved in creating data partitions. However, this method is computationally expensive and, perhaps worst, the test

set cannot represent the correct distribution of the target variable because the test set is one instance.

Other than being an essential technique in machine learning, the use of cross-validation in psychological research may have an important role in positively contributing to solving the replication crisis that is affecting the field. A systematic study designed to assess the reproducibility of psychological science (Open Science Collaboration, 2015) found that the mean effect size of the replicated studies was half of that of the originally conducted studies, and, even more strikingly, only 36% of replication studies had statistically significant results. Although replication studies are very important, it is not always feasible to conduct them. Think, for example, of research conducted to investigate the impact of exceptional events. In those cases, cross-validation provides a system for conducting a third type of replication, which is simulated replication (Koul, Becchio, & Cavallo, 2018). In fact, other than avoiding the risk of overfitting, cross-validation increases the confidence that the effects obtained in a specific study will be replicated, instantiating a simulated replication of the original study. In this way, cross-validation mimics the advantages of independent replication (Yarkoni & Westfall, 2017).

2.4.3. Performance Measures

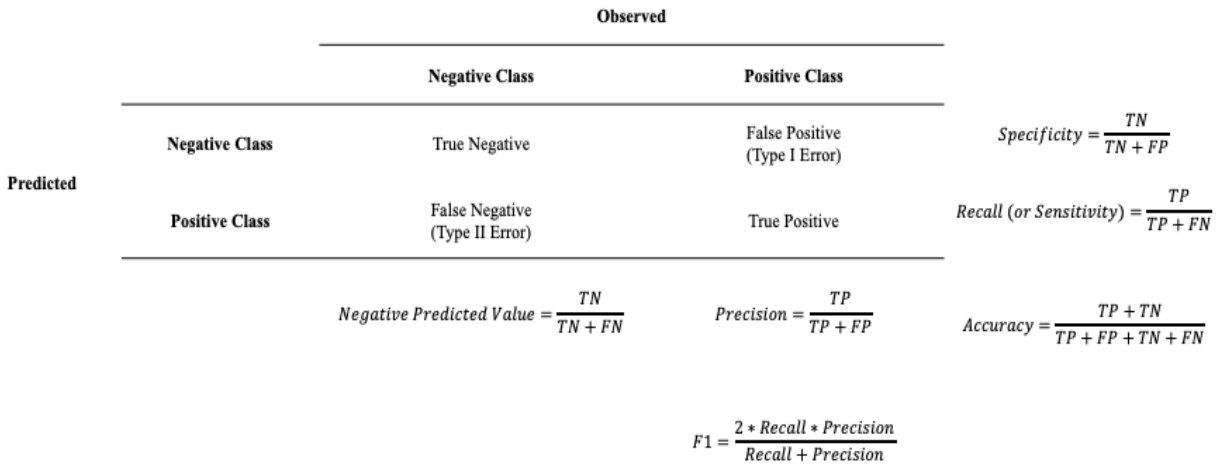
As discussed in the previous section, a model affected by overfitting or underfitting makes "poor" predictions. To evaluate the quality of a prediction, it is necessary to introduce some formal metrics that measure the algorithm’s performance. Evaluating the performance of a model is a fundamental aspect of machine learning. When an algorithm receives a training set as input, it constructs a model that can predict the value or the class of an unseen instance. The model's ability to correctly predict that value is measured through one or more criteria (or performance measures). The evaluation is essential for judging the quality of the model, for refining choices in the KDD iterative process (e.g., feature selection, tuning the algorithm) and for selecting the most performative model from a given set of models built with different

algorithms (Maimon & Rokach, 2010). While often overlooked, the metric used to assess the effectiveness of a model to predict the outcome is very important and can affect the conclusions (Kuhn & Johnson, 2019). The metric we select to evaluate model performance depends on the type of supervised learning task (classification vs. regression).

Classification algorithms can produce two types of outcomes: classes (e.g., binary classification output will be either "0" or "1") and probabilities (i.e., the output is the probability that a specific case belongs to each of the possible categories). Different algorithms can produce either the former or the latter. For example, SVM and KNN create a class output. Algorithms like Logistic Regression, Random Forest, and Naïve Bayes provide probability outputs, which can be converted into class output by setting a threshold probability (e.g., if an instance has a probability below 0,5 will belong to the "0" class, and if it is above 0,5 it will belong to the "1" class). The majority of the classification performance measures consider the class output.

The most straightforward way to assess the performance of a classification model is using the confusion matrix (R. Kohavi & Provost, 1998), a table with many different combinations equal to the product of the number of output classes that are predicted and observed (e.g., binary classification confusion matrix will be composed of four combinations, whereas a four-

Figure 2.3. The structure of a confusion matrix and some performance measures.



classes classification will have 16 combinations). The cells of the matrix summarize with count values the number of correct and incorrect predictions (see Figure 2.3).

In other words, the confusion matrix shows how much the classification model is "confused" when it makes predictions. Besides the number of errors, it provides insights into the types of errors that a classifier has made. In practice, a binary classification model can make two types of errors: It can incorrectly assign, for example, a depressed individual to the no depression category (i.e., False Negative), or it can incorrectly assign a non-depressed individual to the depression category (i.e., False Positive). The confusion matrix is useful for measuring the most commonly used performance metrics, such as accuracy, specificity, precision, recall, and F1 measure (in Figure 2.3, I have reported the metrics that can be computed and how). A visual way to measure the performance of a binary classifier is the Receiver Operating Characteristic (ROC) curve. ROC plot is a popular performance measure, and it is intuitive to interpret. A ROC plot displays on the x-axis the 1 – specificity metric and on the y-axis the recall metric at all threshold levels. The combination of such metrics produces a line that represents the performance. A classification model with a random performance level always shows a straight line from the origin to the top right corner of the ROC space. A classification model with the perfect performance level shows a combination of two straight lines: From the origin to the top left corner and further to the top right corner. By showing the performance level in the plot, it becomes straightforward to compare the performance of different classification models: Curves close to the perfect ROC curve have a better performance than the ones close to the random straight line. From the ROC plot, it can be derived a measure called the Area Under ROC Curve (AUC). As the name indicates, it is the portion of the area under the ROC line. Although the theoretical range of the AUC score is between 0 and 1, the actual scores of meaningful classifiers are greater than 0.5, which is the AUC score of a random classifier.

Unlike classification outcomes, the output of a regression algorithm is always continuous. Thus, compared to classification measure, it requires no further treatment (e.g., setting a probability threshold). Even though there is a comparison between the predicted and actual values (i.e., residuals), errors in regression tasks are not just present or absent because they come in different sizes. There are several alternative residual-based performance metrics, but the most widely known are the Root Mean Squared Error (RMSE), Root Mean Squared Logarithmic Error (RMSLE), and Mean Absolute Error (MAE). The value of these metrics is usually interpreted as either how far (on average) the residuals are from zero or as the average distance between the observed and predicted values (Kuhn & Johnson, 2013). Another commonly used metric is R Squared (R^2 or coefficient of determination) which is the squared correlation between the predicted and actual values. R^2 indicates the proportion of information that is explained by the model. Even though this is an easily interpretable statistic, R^2 is a measure of correlation, not accuracy. Moreover, R^2 has the disadvantage of being dependent on the variation in the outcome variable. This means that if we validate a model on a test set where the outcome variable shows less (or more) variance than the outcome used in the training phase, the R^2 performance would be affected mainly by this variance change.

In practice, all these measures are straightforward to compute and understand. However, applying the right measure is far from trivial as different measures tell different parts of the story.

2.4.4. Models Interpretability

When I described the data mining and machine learning techniques, I introduced the concept of interpretability by stating that some models are "black boxes" (e.g., neural networks), whereas some others are interpretable (e.g., decision trees).

However, what is interpretability, and how to achieve it? Miller (2017) has non-mathematically defined interpretability as the degree to which a human can understand the cause of a decision. In machine learning terms, Kim and colleagues (Kim, Khanna, & Koyejo,

2016) defined interpretability as the degree to which a human can consistently predict the model's result. Generally, it becomes easier to comprehend why certain decisions or predictions have been made when the interpretability of a machine learning model is higher.

Several reasons drive the demand for interpretability (Doshi-Velez & Kim, 2017). The first straightforward reason is inherently linked to our human nature: Human beings are curious and want to know the reasons behind behaviors. Humans possess a mental model of their environment that is updated whenever something unexpected happens by finding an explanation for the unexpected event. When "black box" models are used in research, new research questions may remain hidden if the model only gives predictions without allowing for interpretation. Thus, interpretability is crucial because it facilitates learning and satisfies curiosity as to why machines create certain predictions.

Interpretability becomes important also regarding the goals of the disciplines where machine learning is applied. In scientific fields, like psychology, which aim to gain knowledge on why something happened, the interpretable model itself becomes the source of knowledge. Indeed, model interpretability permits to derive new research questions and opportunities for new research. From psychologists' point of view, it becomes crucial using a model that not only achieves satisfactory predictive performance but also that enables the interpretation of the extracted patterns. Thus, the selection of a model should represent the best compromise between achievable accuracy and the degree of explanation.

Interpretability assumes a role according to the specific research questions. If someone is interested in the prediction per se (e.g., the probability that an experimental manipulation will be effective, or the probability that a client will accept a commercial offer), the model's interpretability is not necessary. If, instead, someone would like to know why predictions have been made (e.g., why there is a relationship between a predictor and manipulation effectiveness, why certain clients are more prone to accept the commercial offer), we need an interpretable model.

Finally, interpretability may become important in relation to the consequences of its predictions. Some models may not require explanations because they are used in a low-risk environment, where mistakes will not have serious consequences, (e.g., a movie recommender system). In high-risk environments (e.g., disease diagnosis), instead, explanations may be critical. The more a model's decision affects people's lives, the more critical it is for the machine to allow for explanations. Thus, even though a single metric, such as classification accuracy, is essential to understand the quality of the predictions, it provides a partial and incomplete description of most real-world tasks (Doshi-Velez & Kim, 2017).

In general, when we aim to interpret a model, we require information on what variables in the model are most important, the effect of each variable in the data on that particular prediction, and the effect of each variable over a large number of possible predictions. Machine learning algorithms are characterized by different degrees of interpretability. We can distinguish two classes of machine learning models: intrinsically interpretable (e.g., regression and decision tree) and post-hoc interpretable models (e.g., ensemble models and neural networks). Intrinsic interpretability means selecting and training a data mining model which results are directly interpretable (e.g., the coefficients of a regression model). Post hoc interpretability means selecting and training a black-box model (e.g., a neural network) and applying interpretability methods after the training.

Interpretability methods can tell what independent variables have the biggest impact on predictions. Commonly used interpretability methods are:

- Permutation importance (Breiman, 2001) is a measure of how much independent variables contribute to a prediction model’s accuracy. The intuition behind permutation importance is that if a feature is not useful for predicting an outcome, then altering or permuting its values will not result in a significant reduction in a model’s performance. Given a trained model, a test dataset, and an evaluation metric, the permutation importance module creates a random permutation of a feature column (i.e., shuffles the values of that column) and

evaluates the performance of the original trained model on the modified dataset. This is done iteratively for each of the variables, one at a time. The module then returns a list of the feature variables and their corresponding importance scores. The importance score is defined as the reduction in performance after shuffling the feature values. If prediction performance gets worse, the variable is important. On the contrary, if the prediction performance does not change, the variable is unimportant because the model ignored that information for the prediction (Molnar, 2019). In the end, variables are ranked according to the impact of their shuffling on the prediction, which is used as an index of their importance.

- Partial Dependence Plot (PDP) was first introduced by Friedman (2001) based on the intuition that visualization is one of the most powerful interpretation tools. PDP shows how each value of the independent variable affects the prediction of the outcome variable. Moreover, PDP can show whether the relationship between the outcome and an independent variable is linear, monotonic or more complex. In general, the PDP has many advantages. PDP is easy to implement and provide a useful mean through which interpret relationships intuitively. Also, PDP intervenes on a feature and measures the changes in the predictions (Zhao & Hastie, 2019) and, therefore, it can give insights on possible causal relationships between the predictors and the criterion of interest. The major disadvantage of PDP is its reliance on the independence assumption which requires the predictors in the PDP to be not correlated with each other, and this assumption is often violated. A solution to this problem is using the Accumulated Local Effect plots (ALE plot; Apley & Zhu, 2016) which shows the conditional instead of the marginal distribution of predictions.
- Shapley Additive Explanation (SHAP; Lundberg, Scott & Lee, 2017) is a unified approach for explaining the impact of each feature on prediction. SHAP is based on Shapley values (Shapley, 1988), a technique used in game theory to determine how much each player has contributed to the success in a collaborative game. In machine learning, Shapley values are

used to determine, for a predictive model (which is treated as if it were a collaborative game among the predictors), the impact of each predictor (considered as a player) on outcome prediction (which is treated as success in the game). Shapley values calculate the importance of a variable by comparing what a model predicts with and without that variable. Since the order in which variables are entered in a model can affect predictions, Shapley values are calculated for every possible combination, to guarantee that the variables are fairly compared. The SHAP approach offers two benefits. Firstly, Shapley values can be calculated for any model, regardless of its complexity. Second, each observation in the dataset will have its own set of SHAP values. Compared to traditional feature importance algorithms, which are a one-size-fits-all approach for determining features importance, Shapley values can pinpoint which independent variables are most impactful for each instance, allowing us to customize our next actions accordingly.

3. Big Data in Hypothetic - Deductive Approach

Traditionally, researchers in psychology make sense of empirical data using theory-driven approaches to explain phenomena (i.e., how things happen) rather than merely describe them (i.e., what has happened). The theory-based focus increases our understanding of causal relationships in psychological processes and underlying mechanisms. However, with the emergence of Big Data research, where computer scientists often use data-driven methods such as machine learning, psychologists have started to adopt bottom-up data-driven approaches.

When integrating Big Data approaches in the psychological sciences, one is, thus, well-advised to carefully consider decades of psychology research and foundational issues of theorizing, measurement, and analysis. There is a general trade-off faced by researchers of all disciplines that is also crucial for successfully integrating Big Data approaches into psychological science, which is the epistemological trade-off.

Traditional deductive (i.e., knowledge-driven or theory-driven) approaches heavily rely on hypothesis testing (see also Coveney, Dougherty, & Highfeld, 2016; Kitchin, 2014). The deductivists' view relies on a priori hypotheses based on what (we think) we know and their critical, empirical test.

The accumulation of large quantities of data has brought forward a scientific practice that generates insights purely from data and stands in contrast with the more traditional deductive approach in psychology. This inductive, data-driven approach permits to learn from actual observed behaviors in a bottom-up fashion, enabling researchers to derive theories from data. The inductive work is much more promising in the context of Big Data than in more traditional research approaches due to Big Data's large power to detect effects reliably, even when they are small, and their a priori likelihood is low. Further advantages have been pointed out by several Big Data researchers (e.g., Anderson, 2008; Prensky, 2009; Siegel, 2013; Steadman, 2013), some going so far to claim that Big Data 'makes the scientific method obsolete' (Anderson, 2008). I argue that Big Data might help to converge inductive and deductive methods and reasoning better. The epistemological strategy should be to produce rich empirical findings and theoretical concepts as a result of both inductive and deductive practices.

Using Big Data does not imply that cognitive and methodological procedures, which have been refined during centuries of philosophical and scientific thought, will be superseded. There is no "end of theory" but only new opportunities. This empirical and conceptual knowledge basis then guides the identification of valid questions and multiple alternative explanations in data-driven science (see also Kitchin, 2013, 2014) that can then be tested against each other (Burnham & Anderson, 2002). The procedure can be repeated infinitely to refine increasingly specific research questions and produce strong inferences (Platt, 1964). Big Data, in combination with well-formalized theories and the testing of increasingly refined research questions, seems to be a promising hybrid approach to professionalize psychological

theory development and knowledge acquisition. Framing the issue of Big Data in terms of oppositions, that is, deduction versus induction, theory-driven versus data-driven, misses the point that both strategies are necessary and can complement each other. Data-driven and theory-driven approaches should be integrated and together provide a rich knowledge base that allows scientists to investigate increasingly fine-grained questions and explanations.

Other than being used for integrating data-driven and theory-driven approaches, thereby providing a rich knowledge base that allows to investigate increasingly fine-grained questions and explanations, Big Data can be used for addressing theory-based research questions. A way through which using Big Data for addressing theory-based research questions is to conduct controlled studies. Experimental conditions can be recreated in the field by using large samples and many features to construct controlled groups. For example, to study visual search, Mitroff and colleagues (Mitroff et al., 2015) developed a mobile game in which respondents had to detect illegal items in X-rays of bags, acting as if they were an airport security officer. One of the research goals was to investigate errors in the visual search of (ultra) rare items. A large number of trials available allowed the investigation of visual search of very rare events, with targets being presented in 1 out of 1000 trials.

4. R or Python? This is the Dilemma

The skills that psychologists prone to conduct Big Data projects should develop are not limited to statistical skills. For implementing the procedures and the techniques described in the last 40 pages, technical skills are required. Since the KDD process relies heavily on programming, learning a language with which performing its steps is fundamental. For Big Data Analytics, programming skills are undoubtedly necessary because the size and the type of the available data may not be processed and analyzed with standard statistical tools, such as SPSS and Excel, which are the most learned software by psychologists.

Nowadays, there exist 269 programming languages (a list is provided in (Ronin, 2019). The considerable number of programming languages raises the question of which language someone needs to know. According to the Kaggle Survey (2018) conducted on more than 23000 professionals, the most used and recommended programming languages for data analytics are Python and R. Both of them are quite robust languages and learn to use one of them is sufficient to carry out the Big Data projects. However, there are some advantages and disadvantages for both of them that learning both of them would be better than learning just one.

4.1. R

R is a programming language and a statistical environment developed by Ihaka and Gentleman in 1993. In its essence, R is rooted in statistics, data analysis, data exploration, and data visualization. Nowadays, R is widely used in the psychological community.

R runs on almost any standard computing platform and operating system. Its open-source nature means that anyone is free to use and adapt the software to whatever platform they choose. R contains a wide variety of packages for acquiring data (e.g., download tweets through TwitteR package), preprocessing, and analyze data. To date, there are more than 15000 packages on the Comprehensive *R* Archive Network (CRAN), which is a network of web servers around the world where you can find the source code, manuals and documentation, and contributed packages. R can communicate with Python, other languages (e.g., Python and C++) can be called in the R environment, and it can be connected to different databases like Spark or Hadoop. Since it was criticized for using only one CPU at a time, R has been evolved for allowing parallelized operations to speed up computations. Developers have created packages like *parallel* (R Development Core Team, 2017) and *foreach* (Microsoft Corporation & Weston, 2015) which can perform tasks on different computer clusters. Another advantage over many other statistical packages is R's sophisticated graphics capabilities. R's base graphics system allows for good control over every aspect of a

plot or graph. Other graphics packages, like *ggplot2* (more than 6 million downloads only in 2018; data retrieved from CRAN), allow for complex and sophisticated visualizations of high-dimensional data. Moreover, R has excellent utilities for reporting and communication, such as *Shiny* (i.e., a tool for building prototype web applications) and *RMarkdown* (i.e., a method for integrating code, graphical output, and text into a journal-quality report).

4.2 Python

Python is a high level, interpreted, and general-purpose dynamic programming language that focuses on code readability, developed by von Rossum in 1991. Differently from R, Python is rooted in computer science and mathematics. The language was developed to be easy to read and cover multiple programming paradigms and applications. The language supports both procedural programming and object-oriented programming. Python versatility includes database connectivity, web frameworks, web scraping, networking, scientific computing, text, and image processing, many of which features lend themselves to various tasks in machine learning, including deep learning, image recognition, natural language processing, and machine learning. Since its release, Python has been extremely popular and is widely used, especially in data preprocessing activities. Python has become popular for a variety of reasons, including the fact that standard platforms and operating systems can support it, it is easy to understand, and it has a considerable number of available packages. The Python library is called PyPI and has more than 113000 packages, which makes it the largest ecosystem of any programming language. Moreover, Python is a better choice, compared to R, if one uses Big Data platforms like Spark or Hadoop.

4.3 Similarities and Dissimilarities

Both R and Python are open-source languages used in a wide range of data analysis fields. Their main difference is R has traditionally been geared towards statistical analysis, while Python has not such a specific nature. To a large extent, both R and Python perform the same

activities in slightly different ways. Thus, R and Python have more points in common than divergences.

As open-source projects, both R and Python are released frequently. Major releases are published annually, where the main new features are incorporated and released to the public. Throughout the year, smaller-scale bugfix releases will be made as needed. The frequent releases and regular release cycle indicate the active development of the software and ensure that bugs will be addressed promptly. Of course, while the core developers control the primary source tree, many people around the world make contributions in the form of new features, bug fixes, or both. The major drawback of relying on a community is that the capabilities of the systems generally reflect the interests of the communities built around them. If no one feels like implementing what a specific user needs, then it is the user's job to implement it.

Both R and Python can be implemented on the Integrated Development Environment (IDE), which is an application that facilitates application development and offers a central interface featuring all the tools a developer needs, such as code editor, compiler, and debugger. There are many IDE that can support R and Python (e.g., Jupyter Notebook, Spyder for Python, and RStudio or StatET for R). Moreover, each IDE is featured for executing and converting the code of a foreign language into its ecosystem. For example, in RStudio is possible to convert Python code in R code using the *reticulate* package (Ushey, Allaire, Tang, Lewis, & Geelnard, 2019), or to insert Python code chunks into the RStudio notebook. Similarly, Jupyter Notebook can implement R code by the creation of notebooks with R as execution context.

Even though there is no official support from the developers, there are a large number of users' communities out there (e.g., Stackoverflow and R-Bloggers) created specifically to provide all the necessary support. They are platforms where one can ask questions about the encountered problem or where one can find existing discussions that allow addressing the

problem. In my experience, 9 out of 10 times, there is someone in the world that had experienced the same issue. Thus, although there is no official support, it is generally easy to find answers regarding issues related to R or Python.

All things considered, the characteristics and the similarities between R and Python decide to study one of them becomes a highly personal choice and adherence to the mainstream choice of one's own scientific community. Thus, choosing a language over the other is not a dilemma. In general, I would not recommend choosing one of them exclusively, but instead, I would suggest working towards having both in the psychologists' research methods toolbox. Since Python and R users intersect a lot, it becomes likely to get involved in projects done in both languages. Thus, at least a basic understanding of both R and Python would be highly recommended.

4.4. Learning Tools

Today, knowing how to program code can be one of the most important skills you can learn that will directly affect career advancement, both in academia as well as the business world (Sijbrandij, 2017). This change is happening because programs like Microsoft Excel and SPSS are powerful tools, but they show limitations in terms of the amount of data one can work with, the kind of analyses one can do, and the types of visualizations one can make. Using a programming language can seem confusing at first, but hard work will be rewarded by making one's work much easier. As stated above, use a programming language can make replication attempts easier, and processes that used to take too much time can take a few minutes of coding.

The best kind of learning is learning by doing. The maximal level of performance in a given domain is not attained automatically as a function of extended experience, but the level of performance can be increased even by highly experienced individuals as a result of deliberate efforts to improve (Lave, 1988). Thus, the first thing a beginner should do is installing the language (from the R Project and Python websites) and its IDE on the computer

(from R Studio and Jupyter websites). A way to proceed concerns starting with basic projects with the appropriate difficulty level and develop the competencies by continually adding higher levels of complexity. Clearly, learn to program on ourselves is compelling but hard. Thus, it is essential to put ourselves into mentors' and peers' hands. There are many online and offline resources for learning how to program with programming languages. In table 3, I have enlisted some resources that helped me to learn R and Python.

The first type of resource is online courses. Different platforms provide their ways to teach programming, including premade learning paths (e.g., Codecademy) and individual courses that can be taken individually (e.g., Coursera). Most of the platforms provide basic content for free, but they require a fee payment for accessing advanced content (e.g., Data Quest). Some others, let access to all content free but only for audition purposes (e.g., Cousera). Thus, if someone wants to get a certificate, fee payment is required. Generally, online platforms provide courses both on R and Python. In general, all the learning platforms provide courses that touch different levels of complexity and difficulty. Thus, one can choose a platform and stick with it throughout the entire learning process, especially if one is willing to pay the fee.

Books are the second essential way to learn a programming language. There are so many great books out there and, unfortunately, not enough time to read them all. In Table 2.2, I have enlisted some good reads that are also available online for free. Some of them provide a great introduction on how to use languages from the beginning (Grolemund, 2014; Shaw, 2013), whereas some others provide an introduction to programming languages in the data science and mining fields (McKinney, 2017; Wickham & Grolemund, 2017).

Table 2.2. List of recommended materials.

Type	Name	Python or R	Pay or Free	Useful websites
Learning Platform	Codecademy	Both	Free	https://www.codecademy.com/catalog/language/python https://www.codecademy.com/learn/learn-r
Learning Platform	DataCamp	Both	Pay (basic contents for free)	https://www.datacamp.com/courses/intro-to-python-for-data-science https://www.datacamp.com/courses/free-introduction-to-r

Learning Platform	Coursera	Both	Free auditory only	https://www.coursera.org/specializations/python https://www.coursera.org/learn/r-programming
Learning Platform	DataQuest	Both	Pay (basic contents for free)	https://www.dataquest.io/course/python-for-data-science-fundamentals/ https://www.dataquest.io/path/data-analyst-r/
Learning Platform	EdX	Both	Free auditory only	https://www.edx.org/course/introduction-to-python-fundamentals-5 https://www.edx.org/course/introduction-to-r-for-data-science-5
Learning Platform	Udemy	Both	Pay (basic contents for free)	https://www.udemy.com/course/complete-python-bootcamp/ https://www.udemy.com/course/r-programming-for-statistics-and-data-science/
Learning Platform	Real Python	Python	Pay (basic contents for free)	https://realpython.com/courses/
Learning Platform	CodeAvengers	Python	Pay	https://www.codeavengers.com/
Learning Platform	Udacity	Both	Pay (basic contents for free)	https://www.udacity.com/course/introduction-to-python--ud1110 https://www.udacity.com/course/programming-for-data-science-nanodegree-with-R--nd118
Learning Platform	Lynda from LinkedIn	Both	Pay	https://www.lynda.com/search?q=python https://www.lynda.com/search?q=r
Book	Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (Mckinney, 2017)	Python	Free online version	https://github.com/Jffrank/Books/blob/master/Python%20for%20Data%20Analysis.%20Data%20Wrangling%20with%20Pandas%2C%20NumPy%2C%20and%20IPython%20(2017%2C%20O%E2%80%99Reilly).pdf
Book and Tutorial	Learn Pyhon. The hard way (Shaw, 2013)	Python	Free online version	https://learnpythonthehardway.org/book/intro.html
Book	A Bite of Python	Python	Free online version	https://python.swaroopch.com/
Book	An Introduction to Statistical Learning: with Applications in R (Witten, James, Hastie & Tibshirani,)	R	Free online version	http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf
Book	R for Data Science (Wickham, 2016)	R	Free online version	https://r4ds.had.co.nz/
Book	Hands-On Programming with R (Grolermund, 2014)	R	Free online version	https://data-flair.training/blogs/python-tutorials-home/ https://data-flair.training/blogs/r-tutorials-home/
Tutorial	Data Flair	Both	Free	https://www.tutorialspoint.com/python/index.htm https://www.tutorialspoint.com/r/index.htm
Tutorial	Tutorials Point	Both	Free	https://www.listendata.com/search/label/Python https://www.listendata.com/p/r-programming-tutorials.html
Tutorial	Listen Data	Both	Free	https://www.listendata.com/
Tutorial	Quick-R	R	Free	https://www.statmethods.net/
Tutorial	PythonTutorials	Python	Free	https://pythonspot.com/
Tutorial	LearnPython	Python	Free	https://www.learnpython.org/
Tutorial	Storybench	Both	Free	http://www.storybench.org/

Online Community	Stack Overflow	Both	Free	https://stackoverflow.com
Online Community	RBloggers	R	Free	https://www.r-bloggers.com/how-to-learn-r-2/
Online Community	KDNuggets	Both	Free	https://www.r-bloggers.com/
Online Community	Python.org	Python	Free	https://www.python.org/

If a particular concept does not make sense or a code continuously end with an error, it would be useful and, hopefully, decisive to look for alternative online resources to disentangle that content. The online resources to learn computer programming are endless, and they are provided by many online tutorials (e.g., Data Flair or Learn Python) as well as online communities (e.g., Stack Overflow or KDNuggets).

5. Conclusion

In this chapter, I described the KDD process and outlined some concepts and tools that are relevant to those psychologists who want to conduct Big Data projects. Throughout the chapter, I also have discussed some methodological issues and highlighted some related pitfalls that need to be considered when applying machine learning models. Nevertheless, I am convinced that machine learning concepts, such as data reduction techniques, out of sample validation (e.g., via cross-validation), data mining techniques and methods of interpretable machine learning (e.g., PDP plots) will contribute to the generalizability and robustness of psychological studies.

I see two ways in which machine learning techniques will play a critical role in psychological research. Firstly, data mining and machine learning comprise a set of techniques that will be useful in addition to the researchers' toolbox of research methods. Along with the advent of large datasets, data mining and machine learning techniques will help to handle their high dimensionality and complexity. While conventional statistical models can quickly reach their limits due to factors such as multi-collinearity, data mining

and machine learning techniques can be that flexible to model highly complex data. If evaluated correctly, data mining models will also provide the chance to more efficiently test which variables provide the most information for the development and validation of theories in psychology. In this sense, interpretable techniques can play an essential role in this process. In the course of unceasing enhancement, large numbers of behavioral indicators could be used to predict behaviors, traits, and other psychological constructs. The resulting models will then provide more information about which indicators were most predictive for the investigated constructs. Finally, the most predictive indicators could then be used after new data is collected to build an updated model, moving towards the creation of cumulative knowledge in the discipline (Eisenberg et al., 2019).

Second, data mining and machine learning techniques will allow insights from psychology to be transferred to applications in a more reliable way. We have seen how cross-validated models can provide a more realistic estimate of how well a model can generalize to unseen data. The fact that out of sample validation is directly connected to the success of machine learning in other contexts, such as business contexts, should make psychologists think carefully. This conception of model performance is applied because companies' health depends on the predictive success of the algorithms. Thus, more generalizable models will increase the relevance of data mining models in applied contexts. In the same vein, psychologists will be challenged with the fact that predictions can often be made without the availability of an explanation and outside of the context of a verified theory (Yarkoni & Westfall, 2017). In other scientific areas such as genetics or natural language processing, the "prediction over explanation" perspective has led to the successful development of models and indirectly to scientific insights and progress (Shmueli, 2010).

Since the use of data mining methods in psychological research is expected to increase and as collaborations with data scientists may become necessary, psychologists must familiarize themselves with both the concepts and the methodology of data mining and machine learning.

Knowledge about these methods will give them the bricks to pave the way for a fruitful implementation in the field of psychological research. As illustrated throughout this chapter, there are many things to know and on which develop competencies and skills. However, the correct usage of data mining and machine learning methods may, in the end, lead to a better understanding of psychological constructs and human behavior.

As this overview shows, many concepts and technique that we can use for the treatment of Big Data are not completely new for psychologists and psychometrists. Sometimes it is simply a matter of learning a new jargon. To mention just a few examples, think about supervised and unsupervised learning, which might be described as techniques for forecasting and techniques to summarize the data, or about the term “features” which we usually call variables, or predictors. Therefore, one initial step to bring psychologists closer to the study of big data is to make them aware that they already know at least some of the analytical tools that they could use. Of course, there are statistical techniques that are known, but not much used in the psychological field, as for example prediction trees, which may be more well suited for big datasets than for the ones typically used by psychologists. And, then, there are techniques that are totally new to the psychological field, probably because they require large samples, as for example the ensemble techniques and the association rules discussed before.

There are also many new techniques required to do research with Big Data. First, the shift from a predominantly hypothesis testing approach to a much more exploratory one, with the adoption of the KDD cycle instead of the traditional hypothesis - operationalization - measurement - statistical test - conclusion cycle comes with specific requirements in terms of technical knowledge.

As in Big Data research, differently from more traditional psychological research, results are not evaluated anymore in terms of consistency with the hypotheses, but in terms of quality of the model, specific instruments are needed. Only part of them (e.g., AUC, explained variance) is already in the psychologist's toolbox. Others must be learned and mastered to

conduct this type of research. Moreover, this shift in strategy requires specific competences, as for instance in the fine-tuning of parameters, which are completely outside traditional psychological research. And of course, in the case of really big datasets, there is the necessity of competencies in cloud and parallel computing.

Appendix

APPENDIX A. List of Publications that dealt with Big Data Analytics and Psychology.

Search Keywords: Big Data, Data Mining, Machine Learning and Psychology

Authors	Year	Title	Journal	Document Type	Content Type	Psychological Field
Li	2019	Innovation Research on Psychological Health Education of Contemporary College Students under the Background of Big Data	2018 international workshop on advances in social sciences	Proceedings Paper	Theoretical	Health Psychology
Hollon, Cohen, Singla & Andrews	2019	Recent Developments in the Treatment of Depression	Behavior therapy	Article	Theoretical	Clinical Psychology
Dechesne & Bandt-Law	2019	Terror in time: extending culturomics to address basic terror management mechanisms	Cognition & emotion	Article	Empirical	Social Psychology
Feng & Sun	2019	On simulating one-trial learning using morphological neural networks	Cognitive systems research	Article	Empirical	Cognitive Psychology
Li	2019	Government accounting optimization based on computational linguistics	Cognitive systems research	Article	Empirical	Cognitive Psychology
Troian, Arciszewski & Apostolidis	2019	The Dynamics of Public Opinion Following Terror Attacks: Evidence for a Decrease in Equalitarian Values From Internet Search Volume Indices	Cyberpsychology-journal of psychosocial research on cyberspace	Article	Empirical	Social Psychology
Meuleman, Moors, Fontaine, Renaud & Scherer	2019	Interaction and Threshold Effects of Appraisal on Componential Patterns of Emotion: A Study Using Cross-Cultural Semantic Data	Emotion	Article	Empirical	Cognitive Psychology
Johns	2019	Mining a Crowdsourced Dictionary to Understand Consistency and Preference in Word Meanings	Frontiers in psychology	Article	Empirical	Psycholinguistic
Kang, Wu & Wang	2019	Principles, Approaches and Challenges of Applying Big Data in Safety Psychology Research	Frontiers in psychology	Article	Theoretical	Safety Psychology
Provoost, Ruwaard, van Breda, Riper & Bosse	2019	Validating Automated Sentiment Analysis of Online Cognitive Behavioral Therapy Patient Texts: An Exploratory Study	Frontiers in psychology	Article	Empirical	Clinical Psychology
Yetton, Revord, Margolis, Lyubomirsky & Seitz	2019	Cognitive and Physiological Measures in Well-Being Science: Limitations and Lessons	Frontiers in psychology	Article	Empirical	Health Psychology
Qin, Liao, Zheng & Liu	2019	Stock Market Exposure and Anxiety in a Turbulent Market: Evidence from China	Frontiers in psychology	Article	Empirical	Cognitive Psychology
Zhao, Zhang, He & Zuo	2019	Data-Driven Research on the Matching Degree of Eyes, Eyebrows and Face Shapes	Frontiers in psychology	Article	Empirical	Neuro Psychology
Montag & Elhai	2019	A new agenda for personality psychology in the digital age?	Personality and Individual Differences	Article	Theoretical	Personality Psychology
Bleidorn & Hopwood	2019	Using Machine Learning to Advance Personality Assessment and Theory	Personality and Social Psychology Review	Article	Theoretical	Personality Psychology

Shatte, Hutchinson & Teague	2019	Machine learning in mental health: a scoping review of methods and applications	Psychological medicine	Review	Review	Health Psychology
Gladstone, Matz & Lemaire	2019	Can psychological traits be inferred from spending? Evidence from transaction data	Psychological science	Article	Empirical	Personality Psychology
Young	2018	The Long History of Big Data in Psychology	American journal of psychology	Article	Theoretical	History of Psychology
Hesse	2018	Can Psychology Walk the Walk of Open Science?	American psychologist	Article	Theoretical	General
Adjerid & Kelley	2018	Big Data in Psychology: A Framework for Research Advancement	American psychologist	Article	Theoretical	Research Methods
Dwyer, Falkai & Koutsouleris	2018	Machine Learning Approaches for Clinical Psychology and Psychiatry	Annual review of clinical psychology	Article	Review	Clinical Psychology
Koul, Becchio & Cavallo	2018	Predpsych: A toolbox for predictive machine learning-based approach in experimental psychology research	Behavior research methods	Article	Empirical	Research Method
Cangelosi & Schlesinger	2018	From Babies to Robots: The Contribution of Developmental Robotics to Developmental Psychology	Child development perspectives	Article	Empirical	Developmental Psychology
Seeboth & Mottus	2018	Successful Explanations Start with Accurate Descriptions: Questionnaire Items as Personality Markers for More Accurate Predictions	European journal of personality	Article	Empirical	Personality Psychology
Moustafa, Diallo, Amoroso, Zaki, Hassan & Alashwal	2018	Applying Big Data Methods to Understanding Human Behavior and Health	Frontiers in computational neuroscience	Article	Theoretical	Health Psychology
Yaden, Eichstaedt & Medaglia	2018	The Future of Technology in Positive Psychology: Methodological Advances in the Science of Well-Being	Frontiers in psychology	Article	Theoretical	Positive Psychology
Hampton, Asadi & Olson	2018	Good Things for Those Who Wait: Predictive Modeling Highlights Importance of Delay Discounting for Income Attainment	Frontiers in psychology	Article	Empirical	Industrial & Organizational Psychology
Guadagno, Nelson & Lee	2018	Peace Data Standard: A Practical and Theoretical Framework for Using Technology to Examine Intergroup Interactions	Frontiers in psychology	Article	Theoretical	Social Psychology
Lowery, Nadler & Putka	2018	Allies From Within: I-O Practitioners in Organizations	Industrial and organizational psychology-perspectives on science and practice	Editorial Material	Editorial	Industrial & Organizational Psychology
Nai, Narayanan, Hernandez & Savani	2018	People in More Racially Diverse Neighborhoods Are More Prosocial	Journal of personality and social psychology	Article	Empirical	Social Psychology
Kobayashi, Mol, Berkers, Kismihok & Den Hartog	2018	Text Mining in Organizational Research	Organizational research methods	Article	Theoretical	Industrial & Organizational Psychology
Luciano, Mathieu, Park & Tannenbaum	2018	A Fitting Approach to Construct and Measurement Alignment: The Role of Big Data in Advancing Dynamic Theories	Organizational research methods	Article	Theoretical	Industrial & Organizational Psychology
Putka, Beatty & Reeder	2018	Modern Prediction Methods: New Perspectives on a Common Problem	Organizational research methods	Article	Theoretical	Industrial & Organizational Psychology

Diener & Seligman	2018	Beyond Money: Progress on an Economy of Well-Being	Perspectives on psychological science	Article	Theoretical	Health Psychology
Stahl & Pickles	2018	Fact or fiction: reducing the proportion and impact of false positives	Psychological medicine	Review	Review	Clinical Psychology
Stevens & Soh	2018	Predicting similarity judgments in intertemporal choice with machine learning	Psychonomic bulletin & review	Article	Empirical	Cognitive Psychology
Soldatova	2018	Digital socialization in the cultural-historical paradigm: a changing child in a changing world	Social psychology and society	Article	Review	Developmental Psychology
Pargent & Albert-von der Gönna	2018	Predictive modeling with psychological panel data	Zeitschrift fur psychologie-journal of psychology	Article in Special Issue	Theoretical	Research Methods
Bittermann & Fischer	2018	How to Identify Hot Topics in Psychology Using Topic Modeling	Zeitschrift fur psychologie-journal of psychology	Article	Empirical	Research Methods
Vijayakumar & Cheung	2018	Replicability of machine learning models in the social sciences: A case study in variable selection	Zeitschrift fur psychologie-journal of psychology	Article in Special Issue	Theoretical	Research Methods
de Schipper & Van Deun	2018	Revealing the joint mechanisms in traditional data linked with big data	Zeitschrift fur psychologie-journal of psychology	Article in Special Issue	Theoretical	Research Methods
Schoedel, Au, Völkel, Lehmann, Becker, Bühner & Stachl	2018	Digital footprints of sensation seeking: Atraditional concept in the big data era	Zeitschrift fur psychologie-journal of psychology	Article in Special Issue	Theoretical	Cognitive Psychology
Zhang, Liu, Xu, Yang & Zhang	2018	Integrating the Split/Analyze/Meta-Analyze (SAM) Approach and a Multilevel Framework to Advance Big Data Research in Psychology Guidelines and an Empirical Illustration via the Human Resource Management Investment-Firm Performance Relationship	Zeitschrift fur psychologie-journal of psychology	Article in Special Issue	Theoretical	Research Methods
Cheung & Jak	2018	Challenges of Big Data Analyses and Applications in Psychology	Zeitschrift fur psychologie-journal of psychology	Editorial Material	Editorial	Research Methods
Kausel	2018	Big data at work: the data science revolution and organizational psychology		Book		Industrial & Organizational Psychology
Hwong, Oliver, Van Kranendonk, Sammut & Seroussi	2017	What makes you tick? The psychology of social media engagement in space science communication	Computers in human behavior	Article	Empirical	Cyberpsychology
Bleidorn, Hopwood & Wright	2017	Using big data to advance personality theory	Current opinion in behavioral sciences	Article in Special Issue	Review	Personality Psychology
Ruggeri, Yoon, Kacha, van der Linden & Muennig	2017	Policy and population behavior in the age of Big Data	Current opinion in behavioral sciences	Article in Special Issue	Review	Research Methods
Matz & Netzer	2017	Using Big Data as a window into consumers' psychology	Current opinion in behavioral sciences	Article in Special Issue	Review	Consumer Psychology
Greenberg & Rentfrow	2017	Music and big data: a new frontier	Current opinion in behavioral sciences	Article in Special Issue	Review	Music Psychology

Chamorro-Premuzic, Akhtar, Winsborough & Sherman	2017	The datafication of talent: how technology is advancing the science of human potential at work	Current opinion in behavioral sciences	Article in Special Issue	Review	Industrial & Organizational Psychology
Lai, Lee, Chen & Yu	2017	Research on Web Search Behavior: How Online Query Data Inform Social Psychology	Cyberpsychology Behavior and Social Networking	Article	Review	Social Psychology
Yarkoni & Westfall	2017	Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning	Perspectives on psychological science	Article	Theoretical	Research Methods
Canal, Roux, Bruchez & Santiago-Delefosse	2017	Digital health: Promises, challenges, and fears. A literature review	Pratiques psychologiques	Review	Theoretical	Health Psychology
Jia, Jia, Hsee & Shiv	2017	The Role of Hedonic Behavior in Reducing Perceived Risk: Evidence From Postearthquake Mobile-App Data	Psychological science	Article	Empirical	Cognitive Psychology
Hauser, Linos & Rogers	2017	Innovation with field experiments: Studying organizational behaviors in actual organizations	Research in organizational behavior: an annual series of analytical essays and critical reviews, vol 37 - research in organizational behavior	Book Chapter	Theoretical	Industrial & Organizational Psychology
Marinelarena-Dondena, Errecalde & Solano	2017	Knowledge discovery applying text mining techniques in Psychology	Revista argentina de ciencias del comportamiento	Article	Review	Research Method
Cheung, Hebl, King, Markell, Moreno & Nittrouer	2017	Back to the Future: Methodologies That Capture Real People in the Real World	Social psychological and personality science	Article	Theoretical	Research Methods
Parigi, Santana & Cook	2017	Online Field Experiments: Studying Social Interactions in Context	Social psychology quarterly	Article	Theoretical	Social Psychology
Bluemke, Resch, Lechner, Westerholt & Kolb	2017	Integrating Geographic Information into Survey Research: Current Applications, Challenges, and Future Avenues	Survey research methods	Article	Theoretical	Research Methods
Gray	2017	Game-XP: Action Games as Experimental Paradigms for Cognitive Science	Topics in cognitive science	Article	Theoretical	Cognitive Psychology
Falk & Bassett	2017	Brain and Social Networks: Fundamental Building Blocks of Human Experience	Trends in cognitive sciences	Article	Review	Neuro Psychology
Ye, Xu, Zhu, Liang, Lan & Yu	2016	The characteristics of moral emotions of chinese netizens towards an anthropogenic hazard: a sentiment analysis on weibo	Acta psychologica sinica	Article	Empirical	Cognitive Psychology
O'Brien	2016	Lamp Lighters and Sidewalk Smoothers: How Individual Residents Contribute to the Maintenance of the Urban Commons	American journal of community psychology	Article	Empirical	Community Psychology
Vinson, Davis, Sindi & Dale	2016	Efficient n-gram analysis in R with cmscu	Behavior research methods	Article	Empirical	Research Methods
Lipizzi, Dessavre, Iandoli & Marquez	2016	Towards computational discourse analysis: A methodology for mining Twitter backchanneling conversations	Computers in human behavior	Article	Theoretical	Research Method

Cheung & Jak	2016	Analyzing Big Data in Psychology: A Split/Analyze/Meta-Analyze Approach	Frontiers in psychology	Article	Theoretical	Research Methods
Sahdra, Ciarrochi, Parker & Scrucca	2016	Using Genetic Algorithms in a Large Nationally Representative American Sample to Abbreviate the Multidimensional Experiential Avoidance Questionnaire	Frontiers in psychology	Article	Empirical	Research Methods
Reindl	2016	People Analytics: an organizational psychology perspective on data-oriented leadership	Gio-gruppe-interaktion-organisation-zeitschrift fuer angewandte organisationspsychologie	Article	Theoretical	Industrial & Organizational Psychology
Pettit	2016	Historical time in the age of big data. Cultural Psychology, Historical Change, and the Google Books Ngram Viewer	History of psychology	Article	Theoretical	Cultural Psychology
Chamorro-Premuzic, Winsborough, Sherman & Hogan	2016	New Talent Signals: Shiny New Objects or a Brave New World?	Industrial and organizational psychology-perspectives on science and practice	Article	Theoretical	Industrial & Organizational Psychology
Dehghani, Johnson, Hoover, Sagi, Garten, Parmar, Vaisey, Iliev & Graham	2016	Purity Homophily in Social Networks	Journal of experimental psychology-general	Article	Empirical	Social Psychology
Sakaluk	2016	Exploring Small, Confirming Big: An alternative system to The New Statistics for advancing cumulative and replicable psychological research	Journal of experimental social psychology	Article	Theoretical	Research Methods
Shahbazi, Raizada & Edelman	2016	Similarity, kernels, and the fundamental constraints cognition	Journal of mathematical psychology	Article	Theoretical	Research Methods
Wang, Gan, Zhao, Liu & Zhu	2016	Chinese mood variation analysis based on sina weibo	Journal of university of chinese academy of sciences	Article	Empirical	Clinical Psychology
Harari, Lane, Wang, Crosier, Campbell & Gosling	2016	Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges	Perspectives on psychological science	Article	Theoretical	Research Methods
Beaton, Dunlop & Abdi	2016	Partial Least Squares Correspondence Analysis: A Framework to Simultaneously Analyze Behavioral and Genetic Data	Psychological methods	Article in Special Issue	Empirical	Research Methods
Kern, Park, Eichstaedt, Schwartz, Sap, Smith & Ungar	2016	Gaining Insights From Social Media Language: Methodologies and Challenges	Psychological methods	Article in Special Issue	Theoretical	Research Methods
Chapman, Weiss & Duberstein	2016	Statistical Learning Theory for High Dimensional Prediction: Application to Criterion-Keyed Scale Development	Psychological methods	Article in Special Issue	Theoretical	Research Methods
Brandmaier, Prindle, McArdle & Lindenberger	2016	Theory-guided exploration with structural equation model forests	Psychological methods	Article in Special Issue	Empirical	Research Methods
Jones, Wojcik, Sweeting & Silver	2016	Tweeting negative emotion: an investigation of Twitter data in the aftermath of violence on college campuses	Psychological methods	Article in Special Issue	Empirical	Cognitive Psychology

Miller, Lubke, McArtor & Bergeman	2016	Finding Structure in Data Using Multivariate Tree Boosting	Psychological methods	Article in Special Issue	Empirical	Research Methods
Stanley & Byrne	2016	Comparing Vector-Based and Bayesian Memory Models Using Large-Scale Datasets: User-Generated Hashtag and Tag Prediction on Twitter and Stack Overflow	Psychological methods	Article in Special Issue	Empirical	Research Methods
Harlow & Oswald	2016	Big Data in Psychology: Introduction to the Special Issue	Psychological methods	Editorial Material	Editorial	General
Chen & Wojcik	2016	A Practical Guide to Big Data Research in Psychology	Psychological methods	Article in Special Issue	Theoretical	Research Methods
Landers, Brusso, Cavanaugh & Collmus	2016	A Primer on Theory-Driven Web Scraping: Automatic Extraction of Big Data From the Internet for Use in Psychological Research	Psychological methods	Article in Special Issue	Theoretical	Research Methods
Kosinski, Wang, Lakkaraju & Leskovec	2016	Mining Big Data to Extract Patterns and Predict Real-Life Outcomes	Psychological methods	Article in Special Issue	Theoretical	Research Methods
Griffiths, Abbott & Hsu	2016	Exploring human cognition using large image databases	Topics in cognitive science	Article	Theoretical	Cognitive Psychology
Pope	2016	Exploring Psychology in the Field: Steps and Examples from the Used-Car Market	Topics in cognitive science	Article	Theoretical	Cognitive Psychology
Armayones, Gomez-Zuniga, Hernandez & Pousada	2015	Big Data and Psychology: an opportunity for the Internet of people?	Aloma-revista de psicologia ciencies de l educacio i de l esport	Article	Theoretical	Research Methods
Gamble, Boyle & Morris	2015	Ethical Practice in Telepsychology	Australian psychologist	Article	Theoretical	Clinical Psychology
Arfer & Luhmann	2015	The predictive accuracy of intertemporal-choice models	British journal of mathematical & statistical psychology	Article	Empirical	Mathemathical Psychology
Guzzo, Fink, King, Tonidandel & Landis	2015	Big Data Recommendations for Industrial-Organizational Psychology	Industrial and organizational psychology-perspectives on science and practice	Article	Theoretical	Industrial & Organizational Psychology
Iliev, Dehghani & Sagi	2015	Automated text analysis in psychology: methods, applications, and future developments	Language and cognition	Article	Theoretical	Research Method
Liikkanen, Jakubowski & Toivanen	2015	Catching earworms on twitter: using big data to study involuntary musical imagery	Music perception	Article	Empirical	Music Psychology
Armayones, Boixados, Gomez, Guillamon, Hernandez, Nieto, Pousada & Sara	2015	Psychology 2.0: opportunities and challenges for the psychology professional in the field of ehealth	Papeles del psicologo	Article	Theoretical	Health Psychology
Ma	2015	Investigation and Interrogation of Criminal Behavior Under the Background of Big Data	Proceedings of international symposium on psychology and behavior in china's social transformation under the	Proceedings Paper	Theoretical	Forensic Psychology

			background of informatization			
Chang, Hou & Pei	2015	An Interrogation Assisting System Based on Psychological Behavior Analyses of the Suspects	Proceedings of international symposium on psychology and behavior in china's social transformation under the background of informatization	Proceedings Paper	Theoretical	Forensic Psychology
Clough & Casey	2015	The Smart Therapist: A Look to the Future of Smartphones and MHealth Technologies in Psychotherapy	Professional psychology-research and practice	Article	Theoretical	Clinical Psychology
Curini, Iacus S & Canova	2015	Measuring idiosyncratic happiness through the analysis of twitter: an application to the Italian case.	Social indicators research	Article	Empirical	Cognitive Psychology
Koedinger, D'Mello, McLaughlin, Pardos & Rose	2015	Data mining and education	Wiley interdisciplinary reviews-cognitive science	Article	Theoretical	School Psychology
MacLachlan	2014	Macropsychology, Policy, and Global Health	American psychologist	Article	Theoretical	Macropsychology
O'Donnell, Falk & Konrath	2014	Big data in the new media environment	Behavioral and brain sciences	Editorial Material	Theoretical	Research Methods
Kern, Eichstaedt, Schwartz, Park, Ungar, Stillwell, Kosinski, Dziurzynski & Seligman	2014	From "Sooo Excited!!!" to "So Proud": Using Language to Study Development	Developmental psychology	Article	Empirical	Developmetal Psychology
Block, Stern, Raman, Lee, Carey, Humphreys, Mulhern, Calder, Schultz, Rudick, Blood & Breiter	2014	The relationship between self-report of depression and media usage	Frontiers in human neuroscience	Article	Empirical	Clinical Psychology
King & Resick	2014	Data Mining in Psychological Treatment Research: A Primer on Classification and Regression Trees	Journal of consulting and clinical psychology	Article	Theoretical	Clinical Psychology
Montag, Blaszkiewicz, Lachmann, Andone, Sariyska, Trendafilov, Reuter & Markowetz	2014	Correlating Personality and Actual Phone Usage Evidence From Psychoinformatics	Journal of individual differences	Article	Empirical	Personality Psychology
Poldrack & Gorgolewski	2014	Making big data open: data sharing in neuroimaging	Nature neuroscience	Article	Review	Neuro Psychology
Winham & Biernacka	2013	Gene-environment interactions in genome-wide association studies: current approaches and new directions	Journal of child psychology and psychiatry	Article	Theoretical	Developmental Psychology
van der Maaten & Hendriks	2012	Action unit classification using active appearance models and conditional random fields	Cognitive processing	Article	Empirical	Cognitive Psychology

Kelly, Gooding, Pratt, Ainsworth, Welford & Tarrier	2012	Intelligent real-time therapy: Harnessing the power of machine learning to optimise the delivery of momentary cognitive-behavioural interventions	Journal of mental health	Article	Theoretical	Clinical Psychology
Golder & Macy	2011	Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures	Science	Article	Empirical	Clinical Psychology
Jakel, Scholkopf & Wichmann	2007	A tutorial on kernel methods for categorization	Journal of mathematical psychology	Review	Theoretical	Research Method
Helie, Chartier & Proulx	2006	Are unsupervised neural networks ignorant? Sizing the effect of environmental distributions on unsupervised learning	Cognitive systems research	Article	Theoretical	Cognitive Psychology
Devillers, Vidrascu & Lamel	2005	Challenges in real-life emotion annotation and machine learning based detection	Neural networks	Article	Empirical	Cognitive Psychology
Baayen	2005	Data mining at the intersection of psychology and linguistics	Twenty-first century psycholinguistics: four cornerstones	Proceedings Paper	Theoretical	Psycholoing uistic

APPENDIX B. List of data sources

Name	Link	Access information	Description
Awesome Public Datasets	https://github.com/caesar0301/awesome-public-datasets	Varies by database/repository	List of public datasets and repositories
Data on the mind	http://www.dataonthemind.org/	Varies by database/repository	List of public datasets and repositories
APA Responsible Conduct of Research	http://www.apa.org/research/responsible/data-links.aspx	Varies by database/repository	List of public datasets and repositories
brainspell	http://brainspell.org/	Free; account optional	Neuroimaging studies, accompanied by human-generated classification system
Cambridge Centre for Ageing and Neuroscience	http://www.cam-can.com/datasharing/	Information not yet available	Not-yet-available repository of data from a group studying aging and cognition through epidemiological, behavioral, and neuroimaging methods
Dryad	http://datadryad.org/	Free; account required	Repository for sharing and storing nearly any data type
European Union Open Data Portal	https://open-data.europa.eu/	Free	Datasets from institutions and other entities within the European Union
figshare	https://figshare.com/	Free; account required	Repository for data and research sharing (including datasets, papers, and associated files)
Football-Data.co.uk	http://www.football-data.co.uk/data.php	Free	Historical data from football (soccer) matches. Includes results data (from 1993), in-depth match statistics (from 2000), and betting odds (from 2000).
iDASH (Integrating Data for Analysis, Anonymization, and Sharing)	https://idash.ucsd.edu/	Availability varies by user affiliation/role and specific datase	Repository of biomedical and informatics data with some data on health psychology
Internet Archive	https://archive.org/	Varies by dataset	Multimodal archived data from websites, books, videos, audio, television, software, images, concerts, and collections.
inventory.data.gov (US)	https://inventory.data.gov/	Free; account required	Databases on a government information on a wide range of topics
Kaggle	https://www.kaggle.com/datasets	Free	Repository of data from a variety of academic fields
Linguistic Data Consortium	https://www.ldc.upenn.edu/	Paid access only (through LDC membership dues or with nonmember license fee); access also available through some institutions or through scholarship program; account required	Repository of spoken and text corpora in multiple languages (including Arabic, English, German, Japanese, Mandarin, Spanish, and more)
Mendeley Data	https://data.mendeley.com/	Account required; may be subject to fees (depending on size of data)	Research data across a variety of fields for sharing and archiving data
National Archive of Computerized Data on Aging	http://www.icpsr.umich.edu/icpsrweb/NACDA/	Account required; availability may be limited unless affiliated with ICPSR member institution	Datasets related to aging
National Data Archive on Child Abuse and Neglect	http://www.ndacan.cornell.edu/	Free	Data related to child abuse and neglect from a wide range of related areas (e.g., law, health, psychology, public policy)
NBAstuffer.com	http://www.nbastuffer.com/download_nba_box_score_stats_in_excel.html	Free	Datasets on NBA statistics for players (including lineups and shot statistics) and games (including play-by-play datasets and odds)
NeuroSynth	http://neurosynth.org/	Free; account optional	Repository and meta-analysis tool for processed neuroimaging data

NeuroVault	http://neurovault.org/	Free; account required for uploading data	Repository for sharing and storing statistical maps from MRI and PET data
Newcastle Cognition Lab	http://www.newcl.org/	Free	Repository of experimental and survey data related to human memory, attention, and learning
NYC Open Data	https://data.cityofnewyork.us/data	Free	Repository of more than 1600 individual datasets on a range of metrics about New York City (NY), including datasets on health, housing, public safety, and more
Oakland Data Catalog	http://data.openoakland.org/dataset	Free	Datasets on a range of metrics about Oakland (CA), including datasets on health, housing, public safety, and more
Office for National Statistics (UK)	http://www.ons.gov.uk/ons/datasets-and-tables/index.html	Free	Government datasets from the United Kingdom
Online Speech/Corpora Archive and Analysis Resource	https://oscaar.ci.northwestern.edu/	Free; account required	Repository of speech data
Open Access Series of Imaging Studies (OASIS)	http://www.oasis-brains.org/	Free	Data of structural MRI
OpenfMRI	http://openfmri.org/	Free	Data on raw structural, functional, and diffusion MRI
Pecan Street Data Port	https://dataport.pecanstreet.org/	Account required; free for educational use and for university-affiliated researchers (upon verification)	Datasets about consumer energy behavior
Pew Research	http://www.pewresearch.org/data/download-datasets/	Free	Data from numerous Pew surveys
PhyloPic	http://phylopic.org/image/browse/	Free	Data of silhouettes of living organisms and corresponding phylogenic taxonomy
Pittsburgh Science of Learning Center DataShop	https://pslcdatashop.web.cmu.edu/	Free; account required	Multimodal repository and web-based analysis application for research on human learning
Repository of Neural and Cognitive Models	http://models.nengo.ca/	Free	Computational neural and cognitive models (largely created with Nengo software)
ReShare: UK Data Service	http://reshare.ukdataservice.ac.uk/	Free; account required	Lightly curated repository of self-stored UK-related digital data
Santa Barbara Corpus of Spoken American English	http://www.linguistics.ucsb.edu/research/santa-barbara-corpus	Free	Data from audio recordings of human interaction across various regions of the United States and including a variety of speakers and contexts
SF OpenData	https://data.sfgov.org/	Free	Data released by San Francisco city and county
Social Security Administration (US): Hearings and Appeals Public Data Files	https://www.ssa.gov/appeals/publicusefiles.html	Free	Information on hearings and appeals for disability
Social Security Administration (US): Socioeconomic Characteristics	https://www.ssa.gov/policy/data_sub100.html	Free	Data on socioeconomic issues within the USA, including poverty, wealth, and employment
Stanford Network Analysis Project	https://snap.stanford.edu/data/	Free	Data for network analyses

TalkBank	http://talkbank.org/	Free; membership optional but strongly encouraged	Multimodal data on human and animal communication
Tasmanian Cognition Laboratory	http://www.tascl.org/data-repository.html	Information not yet available	Not-yet-finished repository of data from a group studying decision making, language, attention, memory, and learning using experimental and modeling methods
Tennis-Data.co.uk	http://www.tennis-data.co.uk/data.php	Free	Historical data from tennis matches. Includes match and tournament results (from 2000) and head-to-head betting odds (from 2001).
TV News Archives	https://archive.org/details/tv	Varies by dataset	Data on televised news, including (for many) captions and rough statistics for content
Cognitive Modeling Repository	http://shatterboxwebdesign.com/cmr/2-uncategorised	Free; account required	Datasets and computational models of cognition; only includes data from published, peer-reviewed articles
Chess.com Database of Games	https://www.chess.com/downloads/database+of+games	Free; account required	Data on chess matches; specific type of content varies by dataset
CHILDES	http://childes.psy.cmu.edu/	membership optional but strongly encouraged	Multimodal data on child language and communication (subset of TalkBank)
Databrary	https://nyu.databrary.org/	Account required	Data for developmental research on humans and animals, specializing in video, audio, and other unstructured data
Datahub	https://datahub.io/	Account required for some datasets	Self-stored repository for data
Datastage	http://datastage.stanford.edu/	Permission required	Data from Stanford's online courses on multiple Massive Online Open Courses platforms
Dataverse	http://dataverse.org/	Free; account required for some datasets	Repository of data from a variety of academic fields
data.world	https://data.world/datasets/psychology	Free; account required	Open platform for sharing and finding data sets.
re3data.org	https://www.re3data.org/	Varies by dataset	Global registry of research data repositories that covers repositories from different academic disciplines
UCI Machine Learning Repository	http://archive.ics.uci.edu/ml/index.php	Free	Various datasets to assist in machine learning
Wordbank	http://wordbank.stanford.edu/	Free	Data on child language acquisition and vocabulary growth from multiple languages, as measured by MacArthur-Bates Communicative Development Inventory (MB-CDI)
Yahoo Labs Webscope Program	http://webscope.sandbox.yahoo.com/#datasets	Free; may require affiliation with research university	Datasets from Yahoo (e.g., search, image, news)
1000 Functional Connectomes Project (INDI)	http://fcon_1000.projects.nitrc.org/	Free; account required (may require association with research or educational institution) for some data	Data associated with the 1000 Functional Connectomes project (based on resting-state fMRI)
4TU.Datacentrum	https://data.4tu.nl/	Free; account required	Data from a variety of academic fields (e.g., health sciences, business and management)
Alzheimer's Disease Neuroimaging Initiative (ADNI)	http://adni.loni.usc.edu/	Free; account required	Multimodal datasets related to Alzheimer's disease
Board of Governors of the Federal Reserve System (US)	http://www.federalreserve.gov/datadownload/	Free	Data related to the Federal Reserve
Wikipedia	https://meta.wikimedia.org/wiki/Research:Data	Free	Data and reports released by Wikipedia

Wikipedia: List of online music databases	https://en.wikipedia.org/wiki/List_of_online_music_databases	Varies by database/repository	List of music-related databases
Word Association Lexicons	http://www.saifmohammad.com/WebPages/lexicons.html	Free	List of lexicons for word-emotion, word-sentiment, and word-color associations derived from a variety of sources (including Amazon, Yelp, Amazon Mechanical Turk, and Twitter)
American Rhetoric Speechbank	http://www.americanrhetoric.com/speechbank.htm	Free; copyright use varies by specific speech	Data from public speeches, including transcript, audio, and/or video (varies by speech)
Bookworm	http://bookworm.culturomics.org/	Requires collaboration	List of culturomics-based analyses and visualizations
Crowdflower: Data for Everyone	https://www.figure-eight.com/	Free	List of datasets that were created as open data jobs on Crowdflower on Artificial Intelligence
Cross News	http://demeter.inf.ed.ac.uk/cross/publications.html	Access varies by resource	Various Twitter dataset collected for academic studies (largely focusing on news)
/r/datasets	https://www.reddit.com/r/datasets	Varies by database/repository	A subreddit (i.e., message board for a single topic on reddit.com) about datasets, including requests for datasets and notifications about dataset
Bureau of Justice Statistics (U.S.): All Data Collections	https://www.bjs.gov/index.cfm?ty=dca	Free	Data from the U.S. Department of Justice's Bureau of Justice Statistics.
corpus.byu.edu	http://corpus.byu.edu/	Varies by database/repository	Language corpora (primarily English) from a variety of sources, including historical and contemporary uses from speeches, printed texts, and the internet
data.gov (US)	http://catalog.data.gov/dataset	Varies by database/repository	List of government datasets and resources
Deep and interesting datasets for computational journalists: a quick list	http://cjlab.stanford.edu/2015/09/30/lab-launch-and-data-sets/	Free	Blog post from Stanford's Computational Journalism Lab
Deep Learning Datasets	http://nces.ed.gov/edfin/datatools.asp	Varies by database/repository	Deep learning and categorization, including datasets of faces, speech, text, and other images
Education Finance Statistics Center (US)	http://nces.ed.gov/edfin/datatools.asp	Varies by database/repository	Financial information on public primary and secondary education institutions
EEG / ERP data available for free public download	http://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html	Varies by database/repository	EEG datasets and ERP datasets from human subjects. Includes one dataset hosted on the list page.
Elsevier	https://www.elsevier.com/books-and-journals/content-innovation/data-base-linking/supported-data-repositories	Varies by database/repository	Various academic fields
Gateway to Global Aging	https://g2aging.org/	Varies by database/repository	Aging-related social and behavioral research
Learning, Recognition, & Surveillance Lab Downloads	http://lrs.icg.tugraz.at/download.php	Free	Datasets for automatic classifications, including team behavior, consumer behavior, and face detection
LearnSphere	http://learnsphere.org/index.html	Varies by database/repository	Learning-related research
List of databases for machine learning research (Wikipedia)	https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research	Varies by database	Image data, text data, sound data, signal data, physical data, biological data, anomaly data, and multivariate data. Intended to help machine learning data.

National Center for Education Statistics (US)	http://nces.ed.gov/datatools/	Varies by database/repository	National and international education-related statistics
National Institute on Aging Division of Behavioral and Social Research (US)	https://www.nia.nih.gov/research/dbsr/publicly-available-databases-aging-related-secondary-analyses-behavioral-and-social	Varies by database/repository	Aging-related research funded in part or in whole by the NIA/SBR
Neuroscience Information Framework (NIF) Data Federation	https://neuinfo.org/mynif/databaseList.php	Varies by database/repository	Neuroscience
Open Data Stack Exchange	http://opendata.stackexchange.com/	Varies by database	Stack Exchange for the discussion and publicizing of open data.
openMorph	https://github.com/cMadan/openMorph	Varies by database	Brain morphology research
Opinion Mining, Sentiment Analysis, and Opinion Spam Detection	https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets	Free; Permission required	Opinion mining, sentiment analysis, and opinion spam detection
Police Open Data Census	https://codeforamerica.github.io/PoliceOpenDataCensus/	Free	Police-citizen interactions within U.S. cities
Roper Center	http://ropercenter.cornell.edu/polls/dataset-collections/	Varies by database/repository	Political science
Social Security Data Page (US)	https://www.ssa.gov/data/	Varies by database/repository	Datasets from government agencies and initiatives
Sports Betting Data Downloads	http://corywaters.com/sports-betting-data-downloads.html	Free	Sports data, including sections on game/match results, player data, and betting odds.
The Language Goldmine	http://linguisticdata.github.io/	Varies by database/repository	Language-related corpora across multiple languages
U.S. City Open Data Census	http://us-city.census.okfn.org/	Free	18 public metrics of cities within the U.S. (e.g., crime, zoning, health inspections, transit)
U.S. Department of Labor: Consumer Expenditure Survey	http://www.bls.gov/cex/	Free	Consumer spending and income, including data broken down by various demographic measures and family size
OpenPsychometrics	https://openpsychometrics.org/_rawdata/	Free	Datasets on psychological tests

CHAPTER III

THE USE OF DATA MINING IN CUSTOMER RELATIONSHIP MANAGEMENT: A LITERATURE REVIEW ON 18 YEARS OF PUBLICATIONS (2000 - 2018)

1. Introduction

So far, I have dealt with the topic of Big Data and data mining in very general terms with the intention to introduce psychologists to the use of these tools in psychological research. Since this dissertation aims to understand the psychological underpinnings of consumer behavior by using data mining techniques on customers' data, understand how the literature have addressed the use of data mining on customer data could help us to know what types of research have been carried out, what results have been obtained from them, and whether there are studies in which psychological knowledge has been extracted. The review presented in this chapter focuses on the use of data mining on customer data in Customer Relationship Management (CRM), which is the field where customer data has been extensively used to create and maintain profitable relationships throughout the customer lifecycle.

The application of data mining tools in CRM is an established trend in the global economy. With the development and the ongoing improvement of the IT resources and capabilities of digital technologies, organizations register and store a wealth of data about their customers. Given its capacity to turn information into valuable and usable knowledge, data mining has become the privileged method to analyze customers' data. Data mining is generally defined as an exploring process. It uses statistical, mathematical, artificial intelligence, and machine learning techniques to identify regularities and extract meaningful patterns from large amounts of data (Turban, Aronson, Liang, & Sharda, 2007). Thus, through data mining, data

is explored to identify new information, not immediately evident at first glance. An organization can enhance its value by leveraging the power of data, that would be otherwise lost if important conclusions are not extracted quickly enough to direct actions while the business opportunity is still present.

The concept of CRM has been thoroughly examined in marketing and organizational literature. However, the term has not gained a common definition as CRM can reflect several different themes or perspectives (Nevin, 1995). Despite the absence of a univocal definition among researchers, diverse CRM definitions share the idea that it is a comprehensive process of acquiring and retaining customers. A process that, with the help of business intelligence, can maximize the customer value to the company (Ngai, Xiu, & Chau, 2009). Therefore, CRM can be conceived as a set of processes and enabling systems useful for supporting business strategies to build long-term and profitable relationships with customers (Ling & Yen, 2001). CRM has established a new approach to the market that places customers at the center of the business. There exist three different perspectives on CRM: strategic, operational, and analytical (Buttle, 2004).

Strategic CRM is a top-down perspective that focuses on the development of a business culture, centered on customers, that aims at winning and keeping customers by generating and distributing value better than competitors. Operational CRM focuses on the automation of the parts of the business the customers interact with (e.g., marketing automation, sales-force automation, and service automation). Analytical CRM takes advantage of customers' data to enhance both customer and company value. It utilizes technology as an enabler to accumulate, analyze, and disseminate current and prospective customers' data to identify customers' needs more precisely. This review focuses on this last perspective on CRM.

Compared to strategic and operational CRM, analytical CRM represents a bottom-up perspective that focuses on the intelligent mining of customer data for strategic reasons (Buttle, 2004). Thus, analytical CRM, through the usage of data mining, can assist businesses

in finding and selecting the relevant information that may then be used to get a holistic view of the customer lifecycle. A notable characteristic of data mining concerns the ability to answer a wide variety of CRM questions, such as: Which prospects have the highest propensity to be (re) acquired? Which customers have the highest likelihood to switch to competitors? Which customers would be most likely to respond to a cross-selling offer? Who are the most valuable customers?". Given the type of questions that data mining can answer, it becomes clear why data mining has become one of the best supporting tools of business strategies. As a great deal of research has highlighted, companies that use data mining increase customers' loyalty, sales, and company value (Berry & Linoff, 2011). By adopting data mining, a firm's managers and employees can act sooner rather than later, be proactive rather than reactive, and know rather than guess. From the organization's perspective, data mining outputs can improve the quality of decisions and actions that could also be taken for handling other business processes. For example, by knowing who are the most valuable customers, operation teams can allocate sales efforts more appropriately. From the customers' perspective, data mining can help in determining how to deliver better, more timely, more customized services. It can also help identify what products fit customers' needs more closely and thereby enhance their satisfaction. Thus, companies that want to improve their business and excel in competitive markets must recognize the crucial role of data mining and include it in their CRM strategic plans.

The assumption behind analytical CRM is that customer behavior is an event that occurs under certain circumstances. Through prediction models, future customers' behavior may be predicted by discovering hidden relationships and patterns (Rygielski, Wang, & Yen, 2002). The rationale that underlies predictive models is to look at what happened in the past to forecast what is likely to occur in the future (Blattberg, Kim, & Neslin, 2008).

Several years have passed since it was learned that using data mining in CRM could have led to new opportunities for the development and maintenance of companies' competitiveness

and improvement of customer experience and engagement. Data mining can achieve these goals by making the most of the large amount of data that companies are accumulating on their customers. Moreover, organizations can use data mining because of the increasingly advanced IT resources that can be leveraged. Considering the meaningful contribution that data mining brings to the development of effective CRM strategies, it does seem right to ask how this relationship has been addressed in the past literature. We asked ourselves how and in which contexts data mining was used for improving CRM processes? How this relationship has developed over the years and what are the possible developments? What techniques have been used and what methods have been developed to improve the quality of the results and in turn, the quality of the strategical business decisions?

The purpose of this chapter is to answer these questions through a systematic review of the literature (i.e., articles in scientific journals and conference proceedings) published between 2000 and 2018. The focus of the chapter will be directly on the use of data mining techniques on four CRM processes, namely customer acquisition, cross-selling, customer churn, and customer win-back. We choose to analyze these business processes because they are crucial activities for moving customers from one phase of the customer lifecycle to the next. These business processes are critical because they can make customers more valuable over time (Berry & Linoff, 2011). Data mining is a valuable tool throughout the customer lifecycle because it can improve the profitability in each lifecycle stage.

2. CRM Dimensions

2.1. Customer Acquisition

The first task in managing the customer lifecycle is to acquire customers. Customer acquisition is hugely important to companies in many contexts, such as the creation of new business through start-ups, the expansion towards new geographic or customer market segments, the launch of new products or the exploitation of new applications for an existing product, among others. Moreover, even with well-developed customer retention plans,

customers may need replacing because the company lost them for some reason (Buttle & Ang, 2004).

However, companies aim is not just to acquire the largest number of customers, but to acquire the "right" customers, who generate more profits than the costs needed for acquiring them (F. Buttle & Maklan, 2019). Therefore, how to quickly and accurately identify the right prospects to carry out active marketing has become one of the keys for a distributor to gain competitive advantage and reduce marketing costs (Olson & Chae, 2012; Uncles, East, & Lomax, 2013). The identification of prospects to target implies, on the one hand, the estimation of the likelihood that the prospect will convert into a customer, and on the other hand, the profitability of that prospect once he/she is a customer. Data mining can underpin the success of these applications. By analyzing customer data, companies are better informed about which prospects are most promising, and what offers should be addressed.

2.2. Cross-sell

Cross-selling is a marketing strategy that organizations use to develop the relationships with customers (Neslin, et al., 2006). Cross-selling concerns the selling of other products to a customer who has already purchased a product from the vendor. As a consequence, cross-selling helps sellers to increase customer-wallet share, and generate additional sales revenue (Bolton, Lemon, & Verhoef, 2004; Kotler, Bowen, & Makens, 2014). Besides the economic benefits, cross-selling is an effective tool for retaining customers and developing a relationship with them (Kamakura, Ramaswami, & Srivastava, 1991). It can increase the customer's reliance on the company and decrease the likelihood of switching to another provider. Cross-selling activities can increase switching costs and enhance customer loyalty. As a consequence, these activities directly contribute to customer profitability and lifetime value: The more services a customer uses with the firm, the higher the costs of switching to other firms, which leads to loyalty and tenure. Moreover, cross-selling exerts a generally

positive influence on the relationship with the customer, strengthening the link between provider and user (Kamakura, Wedel, de Rosa, & Mazzon, 2003).

Cross-sell campaigns are often based on intelligent data mining. The rationale behind the use of data mining for supporting cross-sell is that customers who purchased one or more products from a specific set are usually prospects for other products in the same set (McFall, 1969). Thus, the types of products a consumer has already bought can be used to predict which product is likely to be purchased next (Paas, 1998). Data mining models can tell marketers the probability of a customer buying any other products based on their transactional history or profile (Buttle & Maklan, 2019). Practically, cross-sell models are constructed by looking at what customers have bought at time (t-1) and what they have bought at time *t*. Then, the profiles of new customers at time *t* are compared with profiles of past customers at time *t*. Finally, the model predicts which product those new customers are most likely to buy at *t+1*.

2.3. Customer Churn

Customer churn (i.e., customers who stop using a company's services) is one of the most important problems for every firm and one of the principal challenges many firms face worldwide (Neslin, Gupta, Kamakura, Lu, & Mason, 2006). Two different types of churn can be identified based on the reasons that cause the failure of the consumer-company relationship (Morita, Lee, & Mowday, 1993; Van den Poel & Larivière, 2004). Voluntary churn occurs when consumers purposely close their contract and move to another provider. Involuntary churn happens for two reasons. Firstly, it might happen when the customers are discharged by the company, for example, because they are consistently late with payments. Secondly, it might happen when customers close the contract without the aim of switching to a competitor, for example, because of financial problems, death, or relocation. Generally, companies aim to predict voluntary churn because it typically occurs due to factors over which companies might have control.

Data mining can help identifying those customers who are more prone to leave the company by detecting the likelihood of the event or forecasting when it is going to happen. Knowing in advance who will churn allows companies to allocate the right resources for addressing the right retention strategies. This attention to retention strategies may have important effects on companies' financial health. In fact, as it has been extensively demonstrated, traditional marketing strategies that focus primarily on customer acquisition may be too expensive. On average, attracting new customers costs five to six times more than retaining existing ones (Athanassopoulos, 2000). Moreover, long-term customers generate more profits, become less costly over time, tend to be less sensitive to competitor's marketing campaigns, and may provide new referrals through positive word-of-mouth (Colgate, Stewart, & Kinsella, 1996; Ganesh, Arnold, & Reynolds, 2000; Verbeke, Martens, Mues, & Baesens, 2011; Zeithaml, Berry, & Parasuraman, 1996).

2.4. Customer Win-Back

Companies with high churn typically spend vast sums on marketing to try to replace all those defectors. Using win-back strategies is one possible and beneficial way to repopulate a company's customer base. Customer win-back, or reacquisition strategies, aims to reacquire lost customers in order to revitalize relationships with them. Thus, win-back strategies can be framed as reactive revival measures (Thomas, Blattberg, & Fox, 2004). According to Kumar and colleagues (Kumar, Bhagwat, & Zhang, 2015), there are three main reasons why a company should rely on win-back strategies. First, lost customers have demonstrated a need for the service. This need makes them more eligible prospects than people that have never been with the firm. Second, lost customers are familiar with the company. This familiarity eliminates the need to create brand awareness and thus reduces the cost of marketing strategies. Finally, knowing how customers used the service throughout their first lifetime is helpful for tailoring more-successful win-back offers and for identifying the most profitable

defectors. Therefore, the best win-back strategy is the one that starts with the segmentation and profiling of lost customers (Griffin & Lowenstein, 2002).

In contractual settings, there are two types of reacquisition efforts: early win-back and late win-back (Lopes, Brito, & Alves, 2013; Stauss & Friege, 1999). Early win back means that a company would retain customers who have given notice of termination but are still contractually tied to the firm (e.g., when the customer decides not to renew the contract). This group of customers typically uses the firm's services until the end of their contract period is reached and the cancellation becomes effective. In this early win-back phase, marketers have to do with the challenge of motivating customers to revoke the termination announcement and restore their commitment up to previous levels, for example, by asking to remain in a postpaid contract or accept a prepaid offer. Late win-back means that the company would reacquire customers who terminated the relationship by canceling the contract or by stopping to use the firm's services.

3. Data Mining Techniques in CRM

Within the business context, data is being generated at an exponential rate by several technological applications. To uncover hidden patterns in this omnipresent data and gain knowledge for business intelligence is becoming extremely crucial.

Data mining is an ongoing process: It starts with data, then through analysis, informs or inspires action, which, in turn, creates data that begets more data mining. The practical consequence is that organizations who want to excel at using their data to improve their business do not view data mining as a sideshow (Linoff & Berry, 2011). Instead, their business strategy must include collecting data, analyzing data for long-term benefit, and acting on the results. To extract meaningful knowledge, data miners have at their disposal a plethora of methods that develop according to two types of analysis. The descriptive analysis aims to identify regularities in data that are expressed in rules. The recognized rules must be

non-trivial in order to develop the knowledge and understanding of the phenomenon under study. In the realm of the descriptive analysis, the data mining tasks that are used more frequently are clustering, association analysis, sequence discovery, and summarization. The predictive analysis aims to estimate the probability that an event occurs based on the rules that emerged in the descriptive phase. The main predictive data mining tasks are classification, regression, and time series analysis. The predictive analytics include a series of techniques that fall within machine learning, such as neural networks and ensemble learning. Machine learning is a branch of computer science that provides data analysis techniques that allow the use of available data to predict future behaviors, trends, or events.

Data mining methods can be classified according to various criteria. Firstly, algorithms can be distinguished according to the presence or absence of a target variable. Supervised learning techniques contemplate the existence in the dataset of a target attribute. The attribute, concerning each record (e.g., associated with each customer), qualifies the class to which the statistical unit belongs to, based on explanatory characteristics. These supervised algorithms perform predictive tasks. Unsupervised learning techniques are applied to data in which the target attribute is absent. Therefore, the analysis is aimed at identifying recurrences and homogeneity in the statistical units of the dataset. The unsupervised algorithms perform descriptive tasks.

Another criterion that can be used to differentiate data mining techniques concerns the objectives that a firm wants to pursue. The most used data mining tasks in CRM are (Berry & Linoff, 2011):

- Classification. Classification derives a model that determines the class of an object based on its attributes. For building a classification model, one needs a target variable, characterized by a set of a discrete and limited number of classes (often only binary), and a set of exploratory variables. Classification techniques aim at identifying a model that describes the relationship between the explanatory variables and the target variable so that

it can assign each record to its class as accurately as possible. Classification is one of the most common learning models in data mining (Ahmed, 2004; Linoff & Berry, 2011; Carrier & Povel, 2003). Classification models can be built through different types of algorithms like logistic regression, decision trees, random forests, among others. The most popular way to apply this kind of analysis is customer churn prediction.

- Regression. Unlike the previous case, which considers targets as discrete modalities, in the regression, the target variable is quantitative and assumes continuous values. The aim is to predict the value of the target variable for each observed statistical unit, based on the available explanatory variables.
- Time series analysis. A time series is a sequential set of data points, measured typically over consecutive and uniform times. The analysis of time series is associated with the discovery of useful patterns and rules in the structure of time series and forecasting of the future values of the observed phenomena.
- Association Analysis. Association aims to establish relationships between items that exist together in a given record (Jiao, Zhang, & Helander, 2006). It is a type of unsupervised data mining that finds patterns in the data. Whether the patterns are meaningful is left to human interpretation. Typical examples of adoption of this analysis in the CRM context are market basket analysis and product cross-selling programs.
- Clustering. It aims to identify homogeneous subgroups (i.e., clusters) within a heterogeneous population (Linoff & Berry, 2011). It is different from classification because at the time the algorithm starts, there are no predefined clusters. The records are grouped base on self-similarity. Thus, the nature of the task is unsupervised, and, similarly to the association analysis, it is left to human interpretation to determine what meaning give to the resulting clusters. Moreover, clustering can be a preliminary step for supervised data mining because the information on which cluster a customer belongs to

can be a useful input to predictive models. The typical application of clustering in CRM is customer segmentation.

4. Literature Review Methodology

4.1. Questions Formulation

With this literature review, we aim to answer some questions which we believe to be essential for understanding how data mining has been used to address the CRM dimensions. The questions we asked ourselves are:

Q1: Through which and how many studies was the use of data mining in CRM processes investigated?

To answer this question, we have identified the type of study (i.e., descriptive and empirical) conducted and the number of studies conducted per year over the period considered. This information allows us to understand the academics' interest in the use of data mining in the context of CRM and how this interest has evolved.

Q2: Which data mining techniques have been used most in empirical research overall and across the CRM processes?

To answer this question, we have identified for each article the algorithms that have been used. This information was subsequently treated from two different points of view. On the one hand, we calculated how many models in total had been developed for each data mining technique. On the other hand, we calculated the number of times a given technique has been used across studies. If one study used the same technique several times (for example, because the models were built on different algorithms), we counted it as unique. For example, if a study reported the use of two ensemble methods (e.g., random forest and AdaBoost), the use of this method (e.g., ensemble method) was counted only once. This allows us to get an estimate of the use of a method parceling out its multiple uses in the same study. Therefore,

this estimate allows us to understand the popularity of the different methods among the studies conducted.

Q3: What objectives had to be achieved by using data mining techniques?

To answer this question, we recorded and categorized the main objective of each study. This information allows us to understand the motivation that drove researchers to apply data mining in the context of CRM.

Q4: What are the characteristics of customer data on which data mining techniques have been applied?

To answer this question, we have recorded for each empirical study the type of dataset used (i.e., public and private) and the type of industry from which these data come. For datasets from private companies, we have recorded their geographical origin. In general, this information allows us to understand in which contexts the use of data mining is most frequent.

Q5: What customer information is relevant to predict customers' behavior in each CRM process considered?

To answer this question, we recorded for all the studies where predictive models were built and in which the results were reported, the type of predictors that improved the predictive performance of data mining models. We believe that this information is extremely relevant for two reasons. On the one hand, they can provide practitioners with suggestions on the type of information that would be useful to obtain for predicting customers' behavior accurately. On the other hand, they allow us to understand customer behavior from a theoretical point of view better.

4.2. Literature Search: Sources and Selection Criteria

Since the research in analytical CRM and data mining can go beyond the limits of business disciplines, the relevant materials are scattered across various online journal and conference

proceedings databases. To provide a comprehensive bibliography of the literature, we searched through Web of Science, Scopus, and IEEE Transaction. The literature search was conducted separately for each step of the CRM dimensions considered and was based on the keywords enlisted in Table 3.1.

Table 3.1. Results of the literature search on the three databases.

Topic	Keywords	Number of unique results
Acquisition	"customer acquisition" *AND* ("data mining" *OR* "machine learning")	130
Cross-Sell	("cross sell*" *OR* "cross-sell*") *AND* ("data mining" *OR* "machine learning")	136
Churn	"customer churn" *AND* ("data mining" *OR* "machine learning")	757
Win-Back	("customer win-back" *OR* "customer win back" *OR* "customer switch back") *AND* ("data mining" *OR* "machine learning")	0
	Total	**1023**

We considered all the manuscripts, either journal papers or conference proceedings, published from 2000 to 2018. These searches produced 1023 manuscripts. The full text of each work was reviewed to eliminate those that were not related to the application of data mining techniques in the four lifecycle steps considered. The selection criteria were as follows:

- Only those manuscripts that dealt with the use of data mining on the four CRM dimensions described above were considered.
- We considered only those studies where customers data from a firm's database were used. Therefore, no manuscripts have been considered if their studies were based on non-customer samples (e.g., students) or synthetic data.
- We discarded all the manuscripts in which only self-report methods (e.g. interview, questionnaire) were used. However, we retained manuscripts in which self-report measures were combined with customer data from companies' databases but discarded

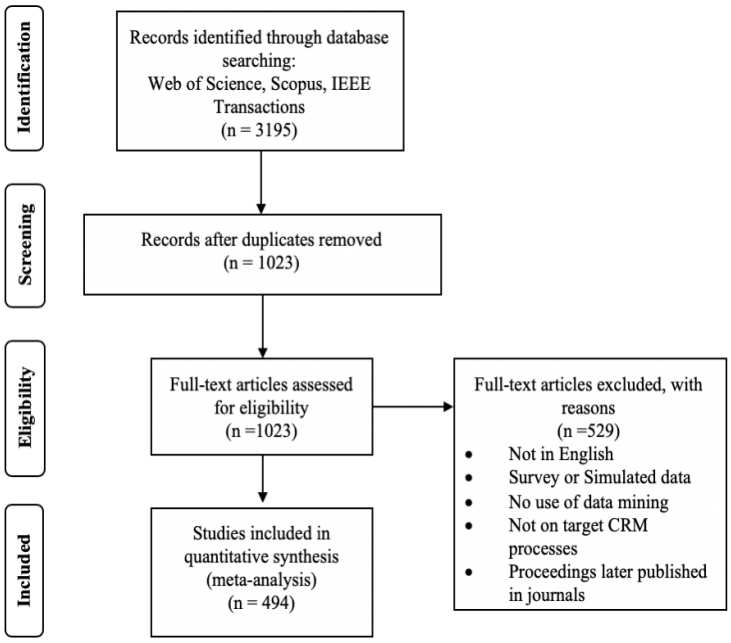
them if the dependent variable was measured through the self-report measures. Customer behavior (e.g., purchase behavior) may be considered as more valid than responses given in a survey (e.g., willingness to purchase).

- We considered only manuscripts written in English.
- Differently from other reviews (e.g., Ngai, Xiu & Chau, 2009), we have also considered conference proceedings as suitable materials but complying with specific conditions. Conference proceedings which contents have been published afterwards in scientific journals were discarded. Moreover, if there were two or more similar conference proceedings, only the most recent one was considered.

Each work was carefully and thoroughly reviewed, and a series of information was retained for making a comprehensive overview of the use of data mining in analytical CRM.

In the end, 494 articles match the including criteria (see Figure 3.1).

Figure 3.1. PRISMA Flow Diagram (adapted from Moher, et al., 2009)



Due to space reasons, the full list of the published works (References List.pdf) and the raw data are stored at the following link

https://drive.google.com/drive/folders/1TEhpoV8uVPldbuZvjNoEoFniLgHBojfe?usp=sharing.

5. Results

5.1. Distribution of Works by CRM Processes and Year of Publication

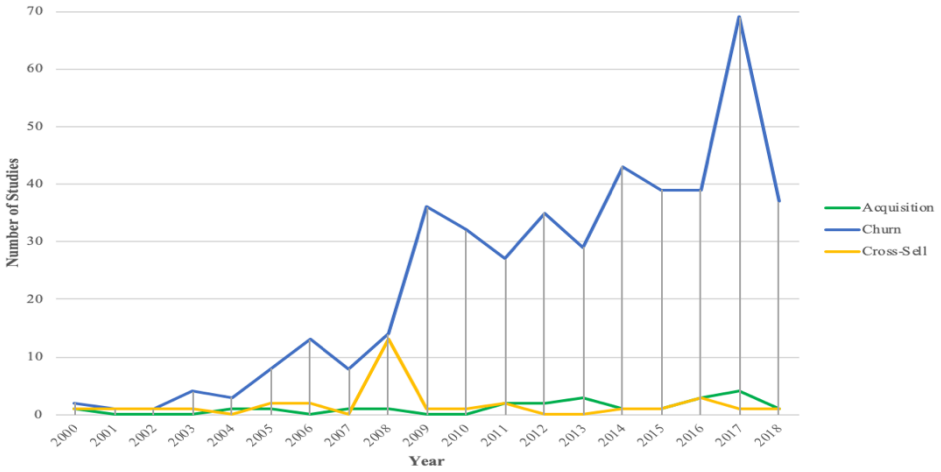
The distribution of the published material across the 4 CRM processes we selected is depicted in Table 2. The most substantial part of the works (89 %, 440 out of 494) addressed the churn topic. Less attention has been dedicated to the other three topics: 4% (22 out of 494) of the manuscripts addressed the customer acquisition topic, and the 6% (32 out of 494) the cross-sell topic. Surprisingly, no manuscript addressed the use of data mining for customer win-back. From Table 3.2, we can see that the literature search reported no results. This finding represents a considerable gap in the literature.

Table 3.2. Number of publications by topic and type.

	Number of results
Acquisition	**22**
Book Chapter	1
Conference Proceeding	6
Research Article	15
Churn	**440**
Book Chapter	3
Conference Proceeding	226
Research Article	211
Cross-sell	**32**
Book Chapter	1
Conference Proceeding	14
Research Article	17
Total	**494**

The distribution of studies by year of publication is shown in Figure 3.2. Whereas acquisition and cross-sell gained low and stable attention throughout the 18 years considered, the attention addressed to the churn topic increased significantly over the same period. However, in the last year churn topic has experienced a decrease of 21% when compared with 2017.

Figure 3.2. Year of publication.



Across topics (see Table 3.3), the majority of the manuscripts are empirical (467 out of 494). Some of the publications describe different aspects of the use of data mining in CRM (24 out of 494). Fourteen of these are literature reviews on the use of data mining for predicting churn behavior. However, eight of these manuscripts revised the use of data mining for predicting customer churn in the telecommunication industry. All the manuscripts included in such reviews were matched with the results of our literature search. Since all manuscripts reported in past reviews had a match in our literature search results, no further work has been included in this review. From now on, all the presented results will concern only the empirical manuscripts.

Table 3.3. Distribution of results by type of study.

	Type of Study	Number of records
Acquisition	Descriptive	1
	Empirical	21
Cross-sell	Descriptive	0
	Empirical	32
Churn	Descriptive	23
	Empirical	414

5.2. Distribution of Works by Data Mining Tasks and Techniques

Among the 494 studies, a total of 1724 models have been built and evaluated, and they vary widely in the type of algorithms used. Table 3.4 shows the distribution of empirical studies by data mining techniques from a general perspective and across the CRM topics.

Table 3.4. Distribution of data mining techniques across topic.

Topic	Data Mining Task	Data Mining Model	Model Type	Number of Models	Number of Studies
Acquisition	Descriptive	Clustering	M - Cluster	1	1
	Preditive	Classification	Decision tree	10	7
			Ensemble	9	8
			Evolutionary Learning	1	1
			Hybrid Algorithm	2	2
			Neural Network	4	4
			Regression	14	11
			SVM	3	3
Cross-Selling	Descriptive	Association Analysis	Association Rules	1	1
			Hybrid Algorithm	1	1
		Clustering	Neural Network	1	1
	Descriptive and Predictive	Association Analysis and Classification	Hybrid Algorithm	2	2
	Descriptive and Predictive	Clustering and Classification	Hybrid Algorithm	1	1
	Predictive	Classification	Bayesian	1	1
			Decision tree	14	11
			Ensemble	24	11
			Evolutionary Learning	2	2
			Hybrid Algorithm	5	5
			Instance Based	1	1
			Markov Logic	3	1
			Neural Network	7	7
			Pattern Recognition	1	1
			Regression	14	11
			Regularization	1	1
			Rule Induction	1	1
			SVM	5	2

Churn	Descriptive	Association Analysis	Pattern Discovery	5	3
	Descriptive	Clustering	Clustering	5	3
			Fuzzy Clustering	1	1
			Hybrid Algorithm	3	1
	Descriptive and Predictive	Clustering and Classification	Hybrid Algorithm	44	19
			Hybrid Evolutionary Learning	2	1
			Hybrid Fuzzy Rule Based System	2	1
		Classification	Auto-classification	1	1
			Bayesian	96	74
			Clustering	4	1
			Decision tree	267	213
			Deep Learning	21	10
			Distance Based	1	1
			Ensemble	357	152
			Ensemble (Deep Learning)	1	1
			Ensemble (Transfer Learning)	11	4
			Evolutionary Learning	15	14
			Factorization Machine	1	1
			Fuzzy Rule Based System	40	17
			Gaussian Models	1	1
			Hybrid Algorithm	28	21
			Hybrid Fuzzy Rule Based System	1	1
			Hybrid Swarm Intelligence	1	1
			Inductive Algorithm	5	4
			Instance Based	41	38
			Linear Classifier	11	8
			Markov Logic	5	5
			Neural Network	161	148
			Pattern Discovery	1	1
			Polynomial Model	2	1

	Probabilistic model	1	1
	Regression	191	169
	Regularization	5	3
	Relational Classifier	10	3
	Rule Induction	48	26
	Spatial Model	1	1
	Spreading Activation	1	1
	Statistical Model	1	1
	Stochastic model	3	3
	Survival Analysis	3	3
	SVM	173	130
	Swarm Intelligence	11	8
	Transfer Learning	4	4
Classification and Survival Analysis	Hybrid Algorithm	2	2
Survival Analysis	Ensemble	1	1
	Regression	6	5
	Total	*1724*	

In all of the three domains, the vast majority of the studies are predictive (1655), with a particular concentration of classification models (1646). The prevalence of classification models occurs in all three topics. The remaining portion of predictive models relates to survival models (7) and the combination of classification models and survival analysis (2). For example, Bahmani and colleagues (2013) proposed a hybrid model that first used neural networks to learn hidden relationships. Then, they implemented the Cox proportional hazard regression to predict customer behavior.

Eighteen models have a descriptive nature with the use of techniques such as clustering and association analysis. Descriptive models are used more frequently to address customer acquisition and cross-sell compared to customer churn. Descriptive models are used more frequently to address customer acquisition (2%) and cross-sell (4%) compared to customer churn (0.08%).

Finally, there are 55 models in which descriptive models (i.e., association analysis and clustering) have been combined with predictive models, especially classification models. For example, Yang and colleagues (Yang, Wu, Zhang, & Lu, 2008) combined decision tree and association rule to discover cross-selling opportunities for a telecommunication service. Within customer churn, there are several examples in which clustering models have been combined with classification models. For example, Bose and Chen (2009) used two-stage hybrid models consisting of unsupervised clustering techniques (e.g., K-Means, Self-Organizing Maps, Fuzzy C Means) and boosted C5.0 decision tree. The results indicated that the hybrid models showed better performance compared to benchmark models where no clustering was used.

Regarding the classification techniques, the most used algorithms are the ensemble models (402; 23%). Following are the decision trees (291; 17%), the regression models (219; 13%), the support vector machines (181; 11%), and the neural networks (173; 9%).

We get similar results if we consider the number of manuscripts in which these algorithms were used. In this case, we considered an algorithm as a unique value. Thus, if in a manuscript were used three different decision trees algorithms, we counted it as one. Decision trees were implemented in 232 (49%) manuscripts, followed by 192 (40%) manuscripts that used regression models, 173 (37%) that used ensemble methods, 160 (34%) used neural networks and 135 manuscripts (29%) used the support vector machines. The discrepancy between the number of ensembles built overall (e.g., 402 ensemble models) and the number of manuscripts in which this type of algorithm was used (e.g., 173 manuscripts) is largely due to the presence of manuscripts where many ensemble models were compared within the same manuscript. For example, Abbasimehr and colleagues (Abbasimehr, Setak, & Tarokh, 2014) compared 16 different ensemble models.

In general, the use of data mining techniques has changed over time. After a period of low activity between 2000 and 2007, there was a general increase in the number of models built

between 2007 and 2008. After 2008 there was an increase in usage of all the algorithms. However, each algorithm shows a specific trend of usage that we will describe below.

We can see from Table 3.4 how the studies conducted on customer churn show a greater variety of the implemented techniques compared to the other topics. This could be traced back to the fact that much of the research on churn aimed at the development or testing of new techniques capable of improving the prediction of customers' behavior, such as hybrid algorithms (in which clustering methods and classification methods are combined), deep learning (Kim, Choi, Lee, & Rhee, 2017), systems based on fuzzy logic (Abbasimehr, Setak, & Soroor, 2013), Markov logic (Prinzie & Van den Poel, 2006b), or rough set theory (Amin et al., 2017).

If the reader intends to obtain more detailed information on the specific algorithms, the performance metrics, and the results obtained from each study, we refer to the supplementary materials (Techniques.csv).

Now, we provide brief descriptions of the commonly used data mining techniques and their use over time.

Ensemble models. An ensemble model for classification is a composite model, made up of a combination of classifiers (Han, Kamber, & Pei, 2012). The main idea behind this method is that each classifier makes its predictions, which are then combined with an ensemble method, such as bagging, boosting, majority voting, or stacking. One of the earliest studies to use an ensemble model in CRM was conducted by Mozer and colleagues (Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000). They explored the performance of the Adaptive Boosting ensembles based on neural network and decision tree in predicting churn. After this work, ensemble models were incrementally used: As shown in Figure 3.3, the number of ensemble models has considerably increased over time, skyrocketing in 2017. The increased use of ensemble techniques could be due to the fact that the combination of two or more learners often provides better predictive performance than a single classifier (Ahmed, Afzal,

Majeed, & Khan, 2017). Of the most popular ensemble models, the most used technique is random forest (83 models). Followed by bagging (44 models), boosting (30 models), AdaBoost (28 models), stacking ensembles (26 models), and majority voting ensembles (18 models). The other part of ensemble models is made up of "unconventional" ensemble models. Some of these models are modified versions of well-known ensemble algorithms such as the decision jungle (Semrl & Matei, 2017), or adjusted real AdaBoost (Liu, Qiao, & Xu, 2011). Some other works tested the performance of an ensemble of various ensemble models. For example, Zhang and his colleagues (Zhang et al., 2008) built an ensemble model as a group of ERTree, RankBoost and Bagging ensembles.

Decision trees. The most popular type of predictive model is the decision tree. Decision tree development usually consists of two phases, tree building, and tree pruning. The tree-building phase consists of recursively partitioning the data according to the values of the features involved. A root is followed by internal nodes, each node is labeled with an "if-then" question, and an arc associated with each node covers all the possible answers (Chen, Hsu, & Chou, 2003). The pruning phase involves selecting and removing the branches that contain the largest estimated error rate. Tree pruning can enhance the predictive performance of the decision tree while reducing its complexity. The most widely used decision tree algorithm is C4.5 (81 models), followed by CART (34 models), and by C5.0 (33 models). Compared to the other algorithms, decision trees show a more stable trend of usage over the years (Figure 3.3), probably because this type of algorithm is often used as a benchmark model.

Regression. Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables when the focus is on the relationship between a target variable and one or more predictors. Not only regression tells whether or not a relationship exists, but it estimates the strength of the relationship between the predictor and the target outcome. This characteristic makes its results easy to interpret, which is an important advantage over "black-box" methods, such as

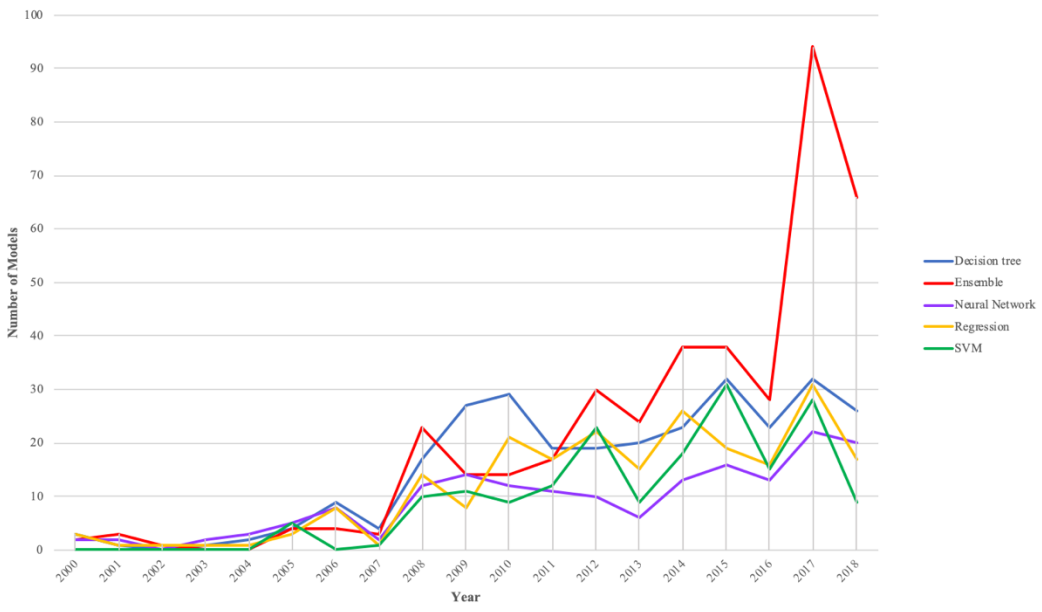
the neural networks. There are several regression techniques used in the literature to classify customer behaviors, but the most used technique is logistic regression (180 models). Like decision trees, logistic regression is mostly used as a benchmark against which to compare the performance of other algorithms. However, logistic regression was also used as a classifier in several ensemble models (Olle & Cai, 2014; Yabas & Cankaya, 2013). Similarly to the trend showed by decision tree methods, regression methods were used quite steadily over the years without showing a peak of major activity.

Support Vector Machine. A Support Vector Machine (SVM; Cortes and Vapnik, 1995) is a universal classifier used for two-class problems. It builds a hyperplane that better separates training instances while keeping the minimum distance between them. The SVM classifier provides good performance under certain conditions like a large number of data samples, few missing values, and nonlinear data. The SVM classifier performs better when there is an equal distribution of the target variable (Cortes & Vapnik, 1995). In the revised manuscripts, SVM plays a special role. SVM has not been used only for predicting behavior, but also for fulfilling other functions within the data mining process, that is data preprocessing. For example, Cao e Shao (2008) introduced and experimentally evaluated SVM-recursive feature elimination algorithm. This algorithm demonstrated to be able to identify key attributes of customer churn, rule out the redundant attributes, and reduce data dimensionality. In another work, SVM was used as a synthetic data generator for data resampling purposes (Khalid, Farquad, & Kamakshi Prasad, 2017). The use of the SVMs had a later onset than the other techniques. The first works reporting its use were published in 2005. Over time, the use of SVMs was unstable (see Figure 3.3).

Neural Networks. Neural networks represent a brain metaphor for information processing. These models are biologically inspired by how the human brain functions. Neural networks showed to be an up-and-coming algorithm in many business applications due to their ability to "learn" from the data, their nonparametric nature (i.e., no rigid assumptions), and their

ability to generalize. They can be used to model complex relationships between inputs and outputs or to find patterns in data. The most used neural network algorithm is the multilayer perceptron (34 models), followed by the back-propagation neural network (11 models), and by the radial basis function network (5 models). Neural networks were used mainly during the first years of activity considered (from 2003 and 2005). Later, the technique was used less than other techniques (see Figure 3.3). Neural networks are a class of powerful and flexible techniques applied to prediction, estimation, and classification problems. Despite these positive characteristics, neural networks' main disadvantage is that its behavior remains uninterpretable. When a neural network produces a solution, it does not give a clue as to why and how. This disadvantage could be the reason for the decrease in its use over time, as long as business insight can be retrieved when data mining models are explicable.

Figure 3.3. Data mining techniques usage over the period.



5.3. Distribution of Articles by Journal and Conferences Topics and Aims

Manuscripts published in journals are distributed across 138 journals. We grouped journals into two main categories according to the classification made by Scimago (Table 3.5): computer science-related journals (e.g., computer science, engineering, mathematics), and business-related journals (e.g., business, management, economics, decision sciences; social

science). Those journals that did not fit into one of the two groups have been classified in the "Other" class (e.g., life science, medicine).

Table 3.5. Distribution of results by journals and conference proceedings fields

Journals

	Field	**Number of Results**
Acquisition	Business Related	7
	Computer Science Related	7
	Other	1
Cross-Sell	Business Related	5
	Computer Science Related	12
Churn	Business Related	46
	Computer Science Related	159
	Other	5

Conference Proceedings

	Field	**Number of Results**
Acquisition	Business Related	1
	Computer Science Related	5
Cross-Sell	Business Related	1
	Computer Science Related	13
Churn	Business Related	28
	Computer Science Related	192
	Other	4

Of the total number of journal articles (242), 178 (71%: 7 on acquisition, 159 on churn and 12 on cross-sell) have been published on computer sciences related journals, 58 (27%: 7 on acquisition, 5 on cross-sell, and 46 on churn) on business-related journals, and the left 5 articles (2%: 5 on churn) fall into the "Other" category. The journal that published most of the works is Expert System with Applications (49 out of 242: 3 on acquisition, 2 on cross-sell, and 44 on churn), followed by Decision Support System (10 out of 242: 3 on acquisition and 7 on churn) and the European Journal of Operational Research (7 out of 242: one on cross-sell, and 6 on churn).

Regarding the conference proceeding topics, we used the same classification criteria. Of the total number of conference proceedings (244), 210 (of which five on acquisition, 13 on

cross-sell, and 192 on churn) were presented at conferences mostly related to computer sciences topics and 30 (of which one on acquisition, one on cross-sell, and 19 on churn) at conferences on business-related topics.

These results suggest that, in general, the aims that the authors intended to achieve with their studies concerned either methodological improvements in implementing data mining in the context of CRM (287 on 494) or improving the understanding of CRM dimensions through the use of data mining techniques (164 out of 494). In Table 3.6, we provide an overview of the objectives covered in the literature. If the reader intends to obtain more detailed information, we refer to the supplementary materials (Aims.csv).

Table 3.6. Distribution of results by papers aims

	Aims	**Number of Results**
Acquisition	Data Augmentation	7
	Identify Prospects	7
	Test Data Mining Method	5
	Compare Data Mining Method	1
Cross-Sell	Predict Cross-Sell Outcome (Propensity, Next Product to Buy)	20
	Test Data Mining Method	11
	Data Augmentation	1
Churn	Test Data Mining Method	197
	Compare Data Mining Method	85
	Predict who churn	78
	Data Augmentation	47
	Factors affecting customer behavior	11
	Use predicted churn rate to build management model	6

Concerning the manuscripts which objective was directed to methodological improvements, the primary aim of these studies was to enhance the prediction performance. Various authors achieved this goal by improving data preprocessing methods or by powering the prediction methods. Over the past 18 years, various methods have been developed for data preprocessing that have been tested in the context of CRM, such as the data reduction (Sato,

Huang, Lefait, Kechadi, & Buckley, 2009), variable selection (Mitrović, Baesens, Lemahieu, & De Weerdt, 2018), and resampling methods (Sundarkumar, Ravi, & Siddeshwar, 2016). However, in the same period, more works were published in which the improvement of predictive performance was achieved by improving the techniques and methods used in the prediction phase. This includes testing the prediction performance of existing or modified versions of data mining techniques (Sivasankar & Vijaya, 2017), the development of new techniques, such as ensemble models (De Caigny, Coussement, & De Bock, 2018), or the development of new performance metrics (Óskarsdóttir, Baesens, & Vanthienen, 2018).

Another commonly pursued methodological objective is the comparison of the predictive performance of various data mining techniques in order to establish which leads to better results in terms of predictive performance. On the one hand, some authors compared the performance of different algorithms to verify which one would have given the best predictive performance. On the other hand, some authors compared the prediction performance of hyperparameters tuning of a same algorithm in order to establish which parameters values can improve predictive performance (Gajowniczek, Orłowski, & Zabkowski, 2016).

However, we found that the predictive techniques were not the only ones to be compared. In fact, there are works in which data preprocessing methods have been compared to test their impact on model performance. Over the nearly twenty years considered, feature selection methods (Li, Wang, & Chen, 2016), data transformation methods (Moeyersoms & Martens, 2015), data representation methods (Mitrović, Singh, Baesens, Lemahieu, & De Weerdt, 2017), resampling methods (Zhu, Baesens, Backiel, & vanden Broucke, 2017), resampling rates (Aditsania, Adiwijaya, & Saonard, 2017), and time windows lengths (Ballings & Van den Poel, 2012) have been examined.

The studies that addressed the use of data mining in CRM from a business point of view can be further divided into two categories. Some manuscripts addressed CRM dimensions for predicting customer behavior in order to drive marketing decisions. In the case of customer

acquisition, models were created to predict the likelihood of prospects being acquired by the company and to use the results to assist sales representatives (D'Haen & Van den Poel, 2013). In the case of cross-selling, some models were built to establish the propensity to accept the cross-sell proposal put forward by the company to improve the customer's portfolio return and the company's profitability concurrently (Ali, Akcay, Sayman, Yilman & Ozcelik, 2017). Some other models were developed to predict what product is more likely to be purchased by the customer (i.e., next product to buy models; Prinzie & Van den Poel, 2006a). In the case of churn, business-related objectives are more varied. In general, the most pursued aims concern, on the one hand, the prediction of churn of specific types of customers (e.g., first time users, Chou & Chuang, 2018); or expert users (Adaji & Vassileva, 2015) or according different definitions of churners (e.g., financial and commercial churners, Burez & Van den Poel, 2008; influential churners, Droftina, Štular, & Košir, 2015). On the other hand, there are many studies that aimed to apply data mining to predict churn behavior in specific industries for the first time (e.g., the prediction of churn in an online health community; Wang, Zhao, & Street, 2017).

The other business objective concerns the development of data mining models to establish if using additional and untested information about customers can improve the performance of the model. For example, in the case of B-to-B customer acquisition, it was verified the positive effect of including various customers' information extracted by Facebook (Meire, Ballings, & Van den Poel, 2017). In cross-sell, Mau and colleagues (Mau, Pletikosa, & Wagner, 2018) found that the search keys entered by customers on an e-commerce web site improved the predictive performance of the cross-sell model. In the case of customer churn, it has been shown, for example, how the information on the structure and interactions within the customer network has a positive impact on the identification of high-risk churners (Mitrović et al., 2017).

Finally, a small portion of articles has instead made use of data mining techniques to identify and understand which factors can affect customer behavior and its prediction (Guo & Qin, 2015).

5.4. Distribution of Articles by Dataset Type, Industries and Dataset Geographical Provenance

Overall, 650 datasets were analyzed in the 467 empirical studies. Most of these datasets were gathered by authors privately from firms (460), while other studies used publicly available datasets (190). The number of datasets analyzed exceeds the number of studies because many of them have been conducted on more than one dataset, especially when the aim was to develop new and more effective data mining methods. In such cases, the efficacy of the new methods would have been compared both across methods and across data, thus increasing the generalizability of the results.

Publicly available datasets can be downloaded for free from several open repositories (e.g., UCI Machine Learning Irvine). The use of open-source datasets is frequent in CRM studies, probably because they are easy to retrieve compared to private data. Secondly, open-source data may not require excessive preprocessing efforts since they are published ready to be analyzed. Finally, using public datasets can facilitate the direct comparison of results between different studies. From the revised manuscripts, we observed the presence of 22 different datasets related to churn (e.g., the dataset provided by the Teradata Center at Duke University), three datasets on cross-selling (e.g., the dataset of used for the 2007 PKDD competition), and one on customer acquisition (i.e., the one used in the UC San Diego data mining contest). Most of these datasets come from the telecommunications industry.

The other type of dataset, the datasets provided by companies, show a much more varied situation (see Table 3.7).

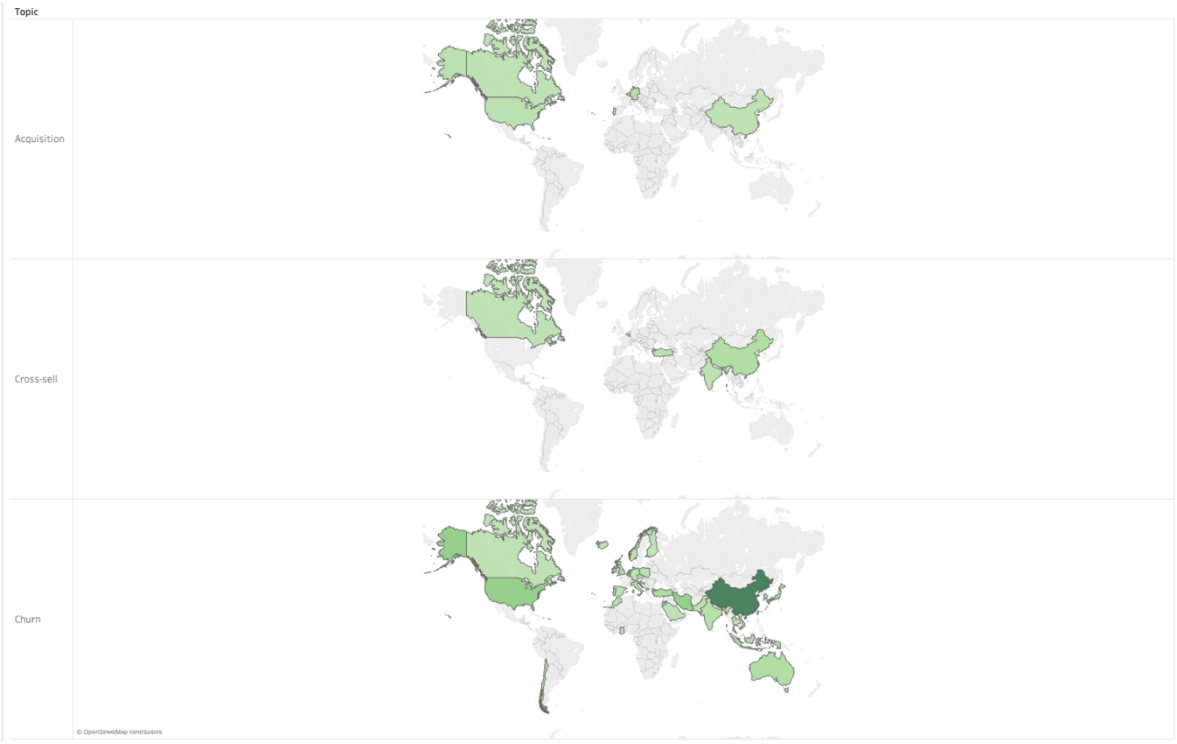
Table 3.7. Type of datasets used and their industry.

		Industry Macro Category	Number of Dataset
Private Dataset	**Acquisition**	Business & Finance	5
		Goods and services	5
		Automotive	4
		Telecommunication & Computing	3
		Energy & Natural Resources	1
		Food & Beverage	1
	Cross-Sell	Business & Finance	9
		Telecommunication & Computing	5
		Goods and services	4
	Churn	Telecommunication & Computing	254
		Business & Finance	66
		Goods and services	92
		Healtcare	4
		Energy & Natural Resources	3
		Automotive	2
		Manufacturing and Construction	2
Open Source Dataset	**Acquisition**	Telecommunication & Computing	1
	Cross-Sell	Business & Finance	10
		Goods and services	1
	Churn	Telecommunication & Computing	157
		Business & Finance	16
		Goods and services	5

The telecommunications sector is still the one on which most models have been developed, especially in the context of churn. However, in private datasets, we can note a large presence of the business and finance sector, which includes both the banking and insurance sectors, the services and products sector, which includes commercial categories such as entertainment and media, retail, and publishing industries. The presence of such a large number of private datasets suggests that many companies worldwide are increasingly willing to ground their decision-making processes on the use of data mining techniques. We can see from Figure 3.4 how data mining projects are geographically spread. The states in which data mining models

have been developed more frequently are China (53), Belgium (20), and the United States (14).

Figure 3.4. Distribution of dataset geographical provenance across topics.



5.5. The Predictors of Customers' Behaviors

The essential part of data mining is data. The dropping costs of data warehousing and the exponential increase in computational power contributed to the fact that plenty of organizations started to acquire data from their customers. Customer databases are created and processed to get more insights into customers' behaviors, which should help to improve marketing strategies.

In the literature considered, some authors have tested the effects of using specific customers' information on the predictive performance of the model. The rationale is to compare a model without the new information and a model that includes it. If this information is an important predictor, then the predictive performance of the augmented model should be higher than the model without that information. The fundamental contribution of these works lies in the identification of information relating to customers that are predictive of the

customers' behavior and which, consequently, could be useful if not necessary to collect. In Table 8, we have collected the main results obtained from these studies.

Table 8. Recapitulation of the main findings on the predictors of customers' behavior.

Authors	Industry	Main Findings
		Acquisition
Baecke & Van den Poel, 2012	Goods and services (various categories)	For publicly consumed durable goods neighborhood effects can be identified. However, for more exclusive brands, incorporating spatial information will not always result in major predictive improvements (e.g., luxury products)
Baecke & Van den Poel, 2013	Automotive	In a model that already includes socio-demographic and lifestyle variables, spatial interdependence provides extra predictive values
Baecke & Van den Poel, 2011	Newspaper subscription	Spending pleasure variables, a composite measure of purchasing behavior and attitude, can significantly improve the model predictive performance
D'Haen, Van den Poel & Thorleuchter , 2013	B2B mail order	Web data had a higher predictive performance compared to commercial data, but the combination of both data types rendered the best results.
D'Haen, Van den Poel, Thorleuchter & Benoit, 2016	Energy retailer	Added value of web crawling data and expert knowledge compared to a basic decision support system.
Meire, Ballings & Van den Poel, 2017	Food & Beverage	Facebook is the most informative data source to qualify prospects, and is complementary with the other data sources (e.g., commercial data and web data)
Thorleuchter, Van den Poel & Prinzie, 2012	B2B mail order	Using information of existing customers' websites helps a B-to-B acquisition manager to identify profitable customers with a higher precision.
		Cross-Selling
Ali, Akcay, Sayman, Yilman & Ozcelik, 2017	Banking	Previously used products by the customer and recent product returns can be used for predicting the customer's propensity to accept an offer from a company.
Mau, Pletikosa & Wagner, 2018	Insurance	Enriching traditional customer data with online quotes yields a valuable approach to predicting purchase behavior.
Moon & Russell, 2008	Insurance	Superiority of purchase history over demographics. The joint-space map (measure of customer similarity) obtained from purchase information represents clusters of (unknown) variables that determine purchase behavior (e.g., word-of-mouth, network externalities, lifestyle).
Prinzie & Van den Poel, 2006	Household Appliences	The results indicate a serious loss in predictive accuracy when dropping features on the number of (different) appliances acquired per product category, the order of acquisition of home appliances, and the time until a first acquisition within a product category or between repeated acquisition in a product category.

		Churn
Backiel, Baesens & Claeskens, 2016	Mobile telecommunication	Using network features (e.g., churn neighbors) can improve performance over local features (e.g., account information, service usage) while retaining high interpretability and usability.
Backiel, Verbinnen, Baesens & Claeskens, 2015	Mobile telecommunication	Including social network information improves classification model performance
Ballings, Van den Poel & Verhagen, 2012	Newspaper subscription	Pictorial- stimulus choice data significantly increases AUC of models with administrative (i.e., current subscription) and operational data (i.e., entire customer history).
Baras, Ronen & Yom-Tov, 2014	Mobile telecommunication	Including various measures of social ties capture different calling and texting patterns that entails significant improvement in the accuracy of prediction.
Benoit & Van den Poel, 2012	Business & Finance	The predictive power of the churn model is improved by adding the social network-based variables. Including network structure measures (i.e. degree, betweenness centrality and density in kinship network) increase predictive accuracy, but contextual network-based variables turn out to have the highest impact on discriminating churners from non-churners.
Chen, Gel, Lyubchich & Winship, 2018	Banking	Models trained on both individual (e.g., age, number of transactions) and kinship network (e.g., individual features aggregated within a family, family size, presence of churners in family) features are more accurate than models trained exclusively on individual features
Coussement & Van den Poel, 2008	Newspaper subscription	Adding unstructured, textual information in call center emails into a conventional churn-prediction model resulted in a significant increase in predictive performance
Coussement & Van den Poel, 2009	Newspaper subscription	Adding emotions expressed in client/company emails increases the predictive performance of an extended RFM churn model.
Das, Elsner, Nandi & Rajiv, 2015	News web site	Our models based on Web news content, on-line user activity and sentiment features improved predictability over baseline models that utilize only transaction metadata
Dasgupta, et al., 2008	Mobile telecommunication	Good prediction accuracy can be achieved by using a simple, diffusion-process that exploits social influences affecting churn
Dierkes, Bichler & Krishnan, 2011	Telecommunication	Information on the churn of network neighbors (e.g., word-of-mouth) has a significant positive impact on the predictive accuracy and in particular the sensitivity of churn models

Gamulin, Štular & Tomažič, 2015	Telecommunication	For each subscriber, they observed three social network parameters: The number of neighbors that have churned, the number of calls to these neighbors, and the duration of these calls for different time periods. The results indicate that using only one or two of these parameters yields results that are comparable or better than the complex models with large amounts of individual and/or social network input parameters.
Gruszczyński & Arabas, 2011	Telecommunication	The improvement of prediction gained after including network variables is relatively small, it must be however noted that performance of base model is high and making it better is not simple.
Hashmi & Sheikh, 2012	Telecommunication	The addition of social attributes (e.g., count of offnet in the community, customers' role in the community) resulted in significant improvement in the accuracy of the prediction models
Huang, Kechadi & Buckley, 2009	Internet service provider	Henley segmentation and service usage are efficient for customer churn prediction in the broadband service field
Huang, Kechadi & Buckley, 2012	Land line telecommunication	The new feature/variable set (e.g., information on grants, service order, Henley segmentation) is more efficient than the existing ones (e.g., frequency of use, sphere of influence, and minutes of use)
Kawale, Pal & Srivastava, 2009	MMORPG games	Combining social influence and player engagement factors has shown to improve prediction accuracy
Kim, Jun & Lee, 2014	Telecommunication	Combining network variable with traditional personal improved the churn prediction model compared with the traditional machine learning approach that handles personal information stored in companies
Kirui, Hong, Cheruiyot, & Kirui, 2013	Mobile telecommunication	Improved prediction when using the features derived from call details (e.g., call pattern description, and call pattern changes description features), customer profiles, and contract-related information
Larivière & Van den Poel, 2005	Banking	Variables related to the salespeople (e.g., selling tendency and the number of customers served by the salesperson) show the highest importance measures.
Lee & Jo, 2010	Mobile telecommunication	Customers are more likely to transfer to a competing company when (1) their ARPU of the previous month was rather high (2) they pay their bill using JIRO and (3) they call contact center very few
Lee, Kim & Lee, 2017	Mobile telecommunication	Words exposed on online news articles as advertising have an impact on customer churn prediction.
Liao, Chen, Liu & Chiu, 2015	Virtual world games	Combining monetary consumption, activity energy and social neighbor analyses improves model predictive ability.

Migueis, et al., 2012	Retail	Results reveal the relevance of the inclusion of a products' sequence likelihood in partial churn prediction models
Modani, Dey, Gupta & Godbole, 2013	Pre-paid and post-paid telecommunication	Social networking behavior can provide significant insights, especially for prepaid customers where the profile data is frequently missing or is unreliable.
Ngonmang, Viennet & Tchuente, 2012	Social Network	Using attributes computed from the local community (e.g., community size, neighbor size) around the user allows to build a robust model to predict churn
Nie, Rowe, Zhang, Tian & Shi, 2011	Banking	Demographic information makes little contribution to the churn prediction. The credit card information and the transaction information which relate to behavior, work very well in the model
Padmanabhan, et al., 2011	Shipping service	Service quality factors, in combination with customer demographics and behaviors, do correlate with instances of customer churn
Phadke, et al., 2013	Telecommunication	Many of the socially relevant predictors such as accumulated influence and number of calls to churners are also some of the top variables, which suggests that social predictors play an important role in churn prediction
Pudipeddi, Akoglu & Tong, 2014	Q&A web site	The temporal features (e.g., time gap between account creation and first post) provides the best accuracy. Models learned using three other feature categories, namely knowledge level (e.g., mean reputation of user), content (e.g., average length of the answer), and frequency (e.g., number of answers) rank the next best
Pushpa & Shobha, 2013	Telecommunication	Social network analysis placed a vital role in visualizing the communities, finding the central player who influences the other customers to churn
Radosavljevik, et al., 2010	Pre-paid telecommunication	While adding customer experience management (e.g., service quality) parameters did not influence the predictability of churn, one variation on the sample and especially a particular change in the outcome definition had a substantial influence.
Rehman & Ali, 2014	Pre-paid and post-paid telecommunication	Social network analysis extracts relationships between different subscribers to improve the results produced by the traditional learning algorithms at individual subscriber level
Rowe, 2016	Social Network	Model in which the information about how the user has evolved throughout his lifecycle to date was added, demonstrate a superior performance
Subramanya & Somani, 2017	E-retail	Implicit features obtained through mining of clickstream/web logs and marketing campaign, etc. act as significant features along with conventional data from sales history. These implicit features establish customer behavior and experience, and hence can be used as features to find customer churn
Tang, et al., 2014	Insurance	Derived information based on the value of financial policies, macroeconomic conditions, and investment related policies can help our understanding of customer attrition behavior and give better predictions

Verbeke, Martens & Baesens, 2014	Telecommunication	A significant impact of social network effects, including non-Markovian effects, on the performance of a customer churn prediction model is found
Vo, et al., 2018	Superannuation	Unstructured data retrieved from call log text mining (e.g., semantic information, word importance, word embedding) contains vital information which improves the accuracy of churn prediction by at least 5% on different cus- tomer datasets.
Wei & Chiu, 2002	Mobile telecommunication	The proposed call-behavior-based (e.g., frequency of use, sphere of influence, and minutes of use) churn-prediction technique exhibited satisfactory predictive effectiveness when more recent call details were employed for the churn prediction model construction.
Zhang, Liang, Li, Zheng & Berry, 2011	Telecommunication	Only customer service usage information is used to predict whether customers will churn or not. This method predicts customer churn according to the customer's behavior and is independent of other customers' information
Zhang, Liu, Yang, Shi & Wang, 2010	Mobile telecommunication	Incorporating network attributes into predicting models can greatly improve the prediction accuracy. In particular, the churn behavior for some customers could only be distinguished by the network attributes. Thus,
Zhang, Qi, Shu & Li, 2006	Fixed line telecommunication	Contract length is the most predictive variable. Payment type and several variables of amount and structure of monthly service fees (e.g., average total fee over the first six months, proportion of international IP call fee) are also effective predictors for chum prediction of the investigated provider's subscribers.
Zhang, Zhu, Xu & Wan, 2012	Mobile telecommunication	Traditional classification models that incorporate interpersonal influence (e.g., neighbor composition, similarity, homophily) can greatly improve prediction accuracy

5.5.1. Customer acquisition

If people had never had any contact with the company, no information is available to target prospects efficiently. This unavailability creates an inevitable void within the commercial databases that companies intend to fill in some way. To solve this issue, companies can enhance their databases with commercially available databases sold by external data vendors (Baecke & Van den Poel, 2011; Lix, Berger, & Magliozzi, 1995). However, this is not the only option available. These data can be replaced or used in conjunction with other useful data sources.

Databases can be enhanced with data acquired through traditional methods such as surveys. For example, Baecke and Van den Poel (2011) administer a survey that measured consumer's spending pleasure for 26 product categories to a sample of consumers present in an external database. Afterward, they linked survey results to the external database by predicting the spending pleasure value for the left consumers. Then they used this variable to identify prospects and found that enhancing the data with product categories spending pleasure variables significantly the predictive performance.

Databases can be enriched with data retrieved from the web. Different studies have proved that adding web data retrieved from company's web sites (D'Haen, Van den Poel, & Thorleuchter, 2013; Thorleuchter, Van den Poel, & Prinzie, 2012) or company's Facebook pages (Meire et al., 2017) improves the predictive performance of the model, especially when used in combination with commercial data. Even though they are difficult data to collect and manage, they can improve the performance of acquisition models substantially. Thus, web data helps to identify profitable customers with higher accuracy (D'Haen, Van den Poel, & Thorleuchter, 2013).

Other works, instead, have proved that the inclusion of customer geographical information (e.g., customer location) can entail beneficial effects on model predictive performance. Even if a model already includes a large number of socio-demographic and lifestyle information, extra predictive value can be achieved by considering the geographical location of eligible prospects (Baecke & Van den Poel, 2013) and their geographical proximity (Baecke & Van den Poel, 2012). However, the impact of geographical information is influenced by its level of granularity (e.g., country, district), and by the product category and type (e.g., spatial interdependence best predicts purchasing behavior for durable public goods, followed by privately consumed durable goods, and finally by consumer packaged goods).

5.5.2. Cross-selling

Cross-selling activities have been associated with offering the right product to the right customers at the right time (Li, Sun, & Wilcox, 2005). If one assumes that product characteristics are static over time, the most important variables affecting the customers' buying decisions over time are their characteristics and buying behaviors. One of the main factors capable to improve the prediction of what product a customer would buy next is its product purchase history (Ali, Akçay, Sayman, Yılmaz, & Özçelik, 2016; Moon & Russell, 2008; Prinzie & Van den Poel, 2006a). More specifically, Prinzie and Van den Poel (2006a) found that the number of different products acquired per product category, the acquisition order, and the time until first acquisition within a product category or time between repeated acquisitions were the main contributors of the improvement of model predictive accuracy. Moon e Russell (2008) have found that purchase history can be used to define customers' similarity through a joint space map. The information on customers' similarity proved to be an important predictor of cross-selling behavior, probably because this information can represent other clusters of variables, such as word-of-mouth, network externalities, lifestyle, values, opinion or other psychological construts. Additionally, not only the product purchase history is relevant for predicting customers' behavior, but also the product's search history is important. For example, Mau and colleagues (Mau et al., 2018) demonstrate that enriching traditional customer data with online quotes from an insurer's website could help predict a customer's propensity to adjust or extend his/her products coverage in the near future.

5.5.3. Churn

In the considered literature on customer churn, there are several examples in which the use of commonly used customer data (e.g., socio-demographic, recency, frequency, monetary value) has been combined with other information capable of increasing the predictive capacity of predictive models. One good example is the inclusion of information relating to customers' social networks. The reason why this information is predictive is that customers do not act in isolation. Their actions are strongly influenced by those who act within their social network,

such as family, friends, work colleagues (Domingos & Richardson, 2001). In the context of churn, the social effect means that family members/friends who churn affect ones' churn propensity. As shown by Nitzan and Libai (2011), the exposure to a churning customer increases the likelihood of churn by 80 % after controlling for homophily, (i.e., user similarity). The interdependencies in a network can be measured through explicit links (e.g., communications between actors, family ties) or through implicit links (e.g., matching on demographic attributes, geographic links) (Hill, Provost, & Volinsky, 2006). Regarding the explicit links, numerous studies have been conducted in the telecommunications sector, where information on the customers' social networks has been built starting from the call detail records. In these studies authors have extracted information that described the characteristics of the social network, such as the number of neighbors who churned (Gamulin, Štular, & Tomažič, 2015), the number of contacts (e.g., calls, SMSs) a customer had with churners and non-churners (Backiel, Baesens, & Claeskens, 2016; Baras, Ronen, & Yom-Tov, 2014; Dierkes, Bichler, & Krishnan, 2011), the duration of calls made to and received from churners and non-churners (Dasgupta et al., 2008; Gamulin et al., 2015), the length of calls made with a customer of a competitor company (Modani, Dey, Gupta, & Godbole, 2013; Phadke, Uzunalioglu, Mendiratta, Kushnir, & Doran, 2013). Other studies have investigated how network-related information can be combined effectively with customer data. For example, Backiel and colleagues (Backiel, Verbinnen, Baesens, & Claeskens, 2015) identified an effective way to combine commercial and network features through a combined model. They built a hybrid classifier that uses as input variables the output of two separate classifiers, one based solely on local features and the other a relational network learner. The use of explicit links to recreate customers' social network is not limited to telecommunications but has also been exploited in other industries such as online games (Kawale, Pal, & Srivastava, 2009; Liao, Chen, Liu, & Chiu, 2015) and social network (Ngonmang, Viennet, & Tchuente, 2012).

Regarding the implicit links, it has been shown how it is possible to extract information relating to the family network of each customer from the socio-demographic information and then use this information as a predictor. The inclusion of network information, especially those that capture the contextual characteristics of the social group (e.g., average length of relationship of the network) to which customers belong, have a greater impact on the prediction accuracy than the information relating to the individuals (e.g., customer length of relationship) (Benoit & Van den Poel, 2012; Chen, Gel, Lyubchich, & Winship, 2018). Regardless of the way through network information is extracted or the data mining techniques used, the commonly reported finding is that the model in which network information is included shows better predictive performance than a model in which only commonly customer data is used, and in some cases, churners are recognizable only through network variables (Zhang, Liu, Yang, Shi, & Wang, 2010). The fact that churn is a socially influenced behavior opens new directions of psychological investigation for predicting and explaining customers' behavior. Using information on the customers' social networks would allow psychologists to investigate what and how social factors influence churn behavior.

Generally, revised studies on churn make use of structured data to predict customer churn. Structured data is easy to collect, treat, and analyze, which makes it an efficient approach within the CRM field. However, the information from structured data is typically accounted for only 20% of business insights (Vo et al., 2018). The remaining 80% lie within unstructured data from daily customer interactions, such as emails, calls, chats, and social media interactions. In the revised literature, we found examples in which unstructured data was included in predictive models resulting in an improvement in their predictive performance. For example, Coussement and Van den Poel (Coussement & Van den Poel, 2008) extracted from the emails sent by customers a weighted and aggregated term-by-email matrix that represented the terms used by each customer and the importance of such terms. The same authors (Coussement & Van den Poel, 2009) obtained equally promising results by

extracting the emotions communicated by customers through emails sent to the company through sentiment analysis. Instead, by analyzing the calls between customers and company, Vo and colleagues (2018) extracted information relating to semantic information, the importance of the words used, and word embedding. The use of this data improved model performance of 5%. Conducting further studies on such type of data (i.e., customer-generated data) would be of great importance for psychology because it would allow to investigate the emotions and/or the semantic dimensions that are associated with customer's churn behavior. Other studies have highlighted the importance of additional information such as the sequence of purchase of products (Miguéis, Van den Poel, Camanho, & Falcão e Cunha, 2012), variables related to the salespeople (Van den Poel & Larivière, 2005), the words exposed on news articles as advertising (Lee, Kim, & Lee, 2017), predicted customer lifecycle trajectories based on social and lexical dynamics (Rowe, 2016), clickstream and weblogs mining (Subramanya & Somani, 2017).

6. Conclusion

The application of data mining techniques in CRM is an emerging trend across industries and worldwide. It has attracted the attention of practitioners and academics. This review has identified 494 works related to customer acquisition, cross-sell, and churn published in the last 18 years. We aimed to give a research summary on the application of data mining in the CRM domain and techniques which are most often used, the characteristics of the studies, the datasets that have been used, and the major findings. Although the review might not be exhaustive, it provides useful insight both from academic and business perspectives.

Research on the application of data mining in CRM will increase significantly in the future based on past publication rates and the increasing interest in the area. In the last 18 years, customer churn has been the most studied phenomenon, followed by cross-selling and customer acquisition. For churn publications, 2017 was the most flourishing year. For the cross-sell 2008 was the year of greatest activity. Concerning the acquisition studies, they did

not show peaks of activity during the period considered. The fact that customer acquisition is addressed less than the other two dimensions may be due to the fact that such a process is more complex and more expansive than cross-selling and customer churn. Moreover, companies have limited internal information on prospects, which makes it difficult to predict their responses to acquisition activities (Meire et al., 2017; Thomas, 2001). In general, all the revised manuscripts can provide insights to organizational policymakers on the common data mining practices used in acquiring, cross-selling, and retaining customers.

The lack of studies on the prediction of customer win-back is a significant gap in the literature. This absence could be the result of several factors, such as the actual lack of data to feed a data mining model. Although a company owns the data of lost customers, this information may not be sufficient to determine if a customer will be reacquired. For example, as demonstrated by Kumar and colleagues (Kumar, Bhagwat, & Zhang, 2015), the reasons that led to churn are good predictors of customer reacquisition, and this information is often difficult to obtain. Unless a major incident led a customer to reach out and had the chance to explain why he decided to leave, marketers are left with very little information. The lack of attention towards customer win-back could also be the result of a little attention of the businesspeople. Even though lost customers, if won back, can be profitable to a company and that win-back initiatives are worth the time and effort, most of the companies do not get involved in these initiatives, probably because it is too expensive. Moreover, the challenge of trying to win customers back is that customers can be lost for different reasons. Some may have left because they found a better offer from a competitor, while others did not need the service anymore. Others may have been unsatisfied and frustrated by inconveniences. To get the best from win-back activities, a company should find solutions for each of these reasons, which require efforts and capabilities.

Classification models are the most commonly applied model in the considered CRM dimensions. This finding is not surprising as classification models predict behaviors that are

dichotomous (e.g., become customer vs. not become customer, cross-sold vs. not cross-sold, and churn vs. not churn). However, there are several attempts to hybridize classification models with descriptive models, such as clustering techniques. Several studies showed that this type of model has a good predictive ability (Bose & Chen, 2009; Fathian, Hoseinpoor & Minaei-Bidgoli, 2016).

The most used data mining classification techniques are ensemble models (especially random forest), decision trees (especially C4.5), and regression (especially logistic regression). The fact that these techniques are used more than "black-box" techniques (e.g., neural networks) can be explained by the fact that they make it easier to interpret the relationships between predictors and target behavior and use such information to direct marketing actions.

Since increasing the predictive performance of data mining models is a desirable outcome, a large portion of the literature has been dedicated to the development and test of new predictive techniques that have often proved to be more efficient in predicting behavior than existing algorithms. For example, there are works in which new ensemble of base learners (e.g., LogitLeaf Model, De Caigny et al., 2018), modified ensemble classification methods (e.g., RealAdaBoost, GentleAdaBoost, ModestAdaBoost, Shao, Li, & Liu, 2007), classification methods based on different logics (e.g., fuzzy logic, Azeem & Usman, 2018), or different mathematical approaches (e.g., rough set theory, Amin et al., 2017) have been tested. Improvements in models' performance were achieved not only by testing new algorithms but also through the data preprocessing methods. There are works where new methods of feature selection. There are works where new methods of features selection (e.g., feature selection by hybrid genetic algorithm with particle swarm optimization, Kamalakannan & Mayilvaghnan, 2018), or new resampling techniques (resampling by SVM synthetic data generator, Khalid, Farquad, & Kamakshi Prasad, 2017) have been examined.

Another way to improve model performance is by testing the predictive power of new and underinvestigated sets of predictors derived from internal or external sources within a firm. The customer database can be seen as the foundation of CRM, which will be used as input for the data mining techniques. The omission of relevant variables can lead to incorrect interpretations and poor predictions. In other words, if the quality of the data is inferior, even the best data mining techniques will still result in mediocre performance (Petrison, Blattberg, & Wang, 1993; Verhoef, Spring, Hoekstra, & Leeflang, 2003). As a result, companies constantly try to augment their database through data collection as well as through the acquisition of commercially available external data. Many studies aimed to find out whether new data sources can increase the model's predictive performance. For example, we have seen that information on customers' social network (e.g., family, friends) is valuable information for the prediction of churn (Backiel, Baesens & Claeskens, 2016). Companies' information retrieved from web pages increments the capacity of the data mining model in identifying good prospects (D'Haen, Van den Poel & Thorleuchter, 2013). The number of different products acquired per product category, the acquisition order, and the time until first acquisition within a product category or time between repeated acquisitions were the main contributors to the improvement of cross-selling model predictive performance (Prinzie & Van den Poel, 2006). The results obtained in these studies can be useful for two reasons. On the one hand, knowing what information proved to predict behavior accurately could lead companies to collect that information and use it in their data mining projects. On the other hand, these types of studies have identified which information is predictive of a given behavioral outcome, thus allowing a better understanding of customers' behaviors.

The revised works that showed the relevant predictors of customer behavior allow us to understand whether particular information is linked to behavior. However, it does not tell us why this relationship exists. Therefore, to better understand consumer behavior, it would be beneficial to interpret predictive relationships to derive possible explanations of the

underlying psychological motivations and guiding new research questions. Although many works have tried to verify whether a given piece of information or set of information can improve the quality of the predictions, none of them has considered outlining possible explanations of such predictions. Deriving psychological knowledge would be valuable not only because it offers the possibility of creating new research questions, but also because a psychological understanding of the predictive relationships may have positive impacts on data-driven business strategies. In the next chapter, I will try to fill this gap using customers' data to understand the psychological underpinnings of customer churn. First, we will build a series of churn predictive models and, by interpreting the predictions of the model with the best predictive performance, I will provide a posteriori psychological explanations of such relationships.

CHAPTER IV

WILL THEY STAY OR WILL THEY GO? THE PREDICTION AND UNDERSTANDING OF CUSTOMER CHURN BEHAVIOR THROUGH DATA MINING TECHNIQUES

1. Introduction

Customer churn (i.e., customers that stop using a company's services) is one of the most important problems for every firm and one of the principal challenges many firms face worldwide (Neslin, Gupta, Kamakura, Lu, & Mason, 2006). Until a few years ago, the energy market in Italy was monopolistic and the Italian electricity provider did not worry much about customer retention. Since the liberalization of the European energy markets[6], new energy companies have entered the competition and for customers, it has become easier and faster to change one's electricity provider. The direct result of these changes is an increasing number of energy customers who decide to switch (i.e., to churn) their service provider. For electricity providers, the cost of replacing lost customers can be very high and may lead to damaging consequences to their financial health. Companies are aware of the necessity to identify which customers will churn and have begun to consider churn analysis as an indispensable part of their strategic decision-making and planning processes. Traditional marketing strategies focused primarily on customer acquisition might be too expensive. On average, attracting new customers costs five to six times more than retaining existing ones (Athanassopoulos, 2000).

[6] Over the last several decades, the European Union energy markets have witnessed an important liberalisation, related to the European Commission's promulgation of directives favoring market liberalization in 1996, 2003, and 2009.

Moreover, long-term customers generate more profits, become less costly over time, tend to be less sensitive to competitor's marketing campaigns, and may provide new referrals through positive word-of-mouth (Colgate, Stewart, & Kinsella, 1996; Ganesh, Arnold, & Reynolds, 2000; Verbeke, Martens, Mues, & Baesens, 2011; Zeithaml, Berry, & Parasuraman, 1996).

Energy suppliers can use the power of analytics in many ways to restrict increasing customer churn: Through the information the companies hold on each of their clients, they can estimate the likelihood of their future churn (Ganesh et al., 2000; Neslin et al., 2006). Practitioners can build churn prediction models using various types of information, such as data related to the customer's account (e.g., whether the customer is residential, commercial or industrial) and contract, socio-demographic information (e.g., sex, age), and behaviors (e.g., complaints, contacts). Through these models, utilities can learn which customers have a high probability of leaving and the characteristics of these customers. They could then use this information to tailor proactive offers and targeted retention campaigns (Keaveney & Parthasarathy, 2001). This approach would enable companies to focus their efforts on high churn-risk customers, which would save money by not providing incentives to customers who do not need them (Dalvi, Khandge, Deomore, Bankar, & Kanade, 2016). Therefore, the accurate identification of those customers more prone to leaving could reduce churn rates by helping suppliers proactively build a profitable and lasting relationship with each customer (Yan, Miller, Mozer, & Wolniewicz, 2001). Moreover, the ability to predict whether a customer will leave and to thus intervene at the right time could be essential for preventing problems and providing a higher quality of customer service (Ahmed & Maheswari, 2016).

Aside from having central importance in the business world, the investigation of churn behavior has also gained considerable relevance in the academic world. Researchers from different fields (e.g., statistics, marketing and business, information technology) could draw new and fundamental knowledge from studies on churn prediction. From the psychologist's perspective, the investigation of churn could contribute to a better understanding of

consumers' behavior, since churn represents not only an economic loss but, more importantly, consumer's behavior.

In the last 20 years, churn prediction has been widely studied in different domains such as telecommunication (for a review see Mahajan, Misra, & Mahajan, 2015), insurance (Bolancé, Guillen, & Padilla-Barreto, 2016; Holtrop, Wieringa, Gijsenberg, & Verhoef, 2017), banking (Prasad & Madhavi, 2012; Sundarkumar & Ravi, 2015), online gaming (Milošević, Živić, & Andjelković, 2017; Tamassia et al., 2016), social networks (Long et al., 2012), subscription services (Burez & Van den Poel, 2007; Coussement & Van den Poel, 2008), and online community platforms (Adaji & Vassileva, 2015; Qin, Cunningham, & Salter-Townshend, 2016). Despite its relevance, the prediction of churn in the electricity market has not been sufficiently studied. In Chapter 3, we have seen that of the nearly 450 works on customer churn published in the last 18 years, only two used a dataset from an energy company (De Caigny, Coussement, & De Bock, 2018; Moeyersoms & Martens, 2015). However, the focus of both papers was not directed to the peculiarities of churn in the energy market but to other aims (i.e., the former on a new ensemble algorithm for classification and the latter on how to handle high-cardinality variables).

The primary aim of the present study is to develop a customer churn prediction model for an Italian energy supplier. We will use data mining and machine learning methodologies and techniques. Data mining techniques search for compelling and useful information without demanding a priori hypotheses. This characteristic is a result of a paradigm shift from fitting data to preconceived theories of the marketplace to using data to frame theories (Erevelles, Fukawa, & Swayne, 2016). In other words, scientific inquiry requires less reliance on existing knowledge and a higher focus on what is still unknown (Sammut & Sartawi, 2012).

The second aim of the present research is to understand the meaning of the predictive relationships highlighted by the statistical model and to shed some light on the psychology behind churn behavior. Our data-driven approach to the predictive model has the potential to

discover the hidden value of information regarding consumers that has never been examined in relation to churn before, possibly because of the theory-driven approach that typically characterizes a large share of the psychological research. At the same time, modelling may help confirm the value of other consumer characteristics that have been related to churn by theory-driven models of consumer behavior.

In conclusion, psychologists can use churn prediction models to better understand the reasons for this behavior: In addition to predicting the future behavior of consumers, a model could tell researchers what features consistently characterized churners. As we decided to use a data mining methodology, we adhered to its characteristics. Therefore, we did not start by stating an *a priori* hypothesis and then testing it. We decided to "let our data talk" by themselves and derive theoretical explanations *a posteriori*. We reported all theoretical explanations in the Discussion section.

1.1. Predictive Churn Modelling

The term "churn" refers to the phenomenon in which a consumer stops using a firm's products or services. Van den Poel and Lariviere (Van den Poel & Larivière, 2004) identified two different types of churn based on the reasons that cause the failure of the consumer-company relationship (Morita, Lee, & Mowday, 1993). Voluntary churn occurs when consumers purposely close their contract and move to another provider. Involuntary churn happens when the customers are discharged from the company (e.g., because they are consistently late with payments) or when they close the contract without the aim of switching to a competitor (e.g., because of financial problems, death, or relocation). Generally, the type of churn researchers wants to predict is the voluntary type, because it typically occurs due to factors over which companies might have control. In this study, we opt to focus on reasons for failure that are not beyond the company's control. It is essential to keep in mind what type of churn we wish to predict, and we must ensure that our data reflect this choice. To

predict the voluntary churn behavior, we need to identify and remove those customers that fall within the other category of churners.

Buckinx and Van den Poel (2005) further classify churn into two other groups based on the type of behavior that characterizes consumers leaving service. Total churn happens when there is an official cancellation of the contract. Hidden churn occurs when the contract is still active, but the consumer is not actively using the service for an extended period. Partial churn happens when the customer uses some of the services offered by the company alongside similar services provided by a competitor. As the electricity market is a contractual setting, churn prediction is the forecast of whether the consumer will officially cancel the contract and move to a competitor over a future relatively short period.

Since customer churn is not a random event that occurs out of nowhere, researchers can discover hidden relationships and patterns from consumers' data that can be predictive of future behavior by building prediction models (Rygielski, Wang, & Yen, 2002). The main purpose of such model is to forecast what is likely to happen in the future based on what happened in the past (Blattberg, Kim, & Neslin, 2008). This approach has become possible because most companies now own databases containing a large variety of potentially valuable information about their consumers' past behavior.

Predictive modelling involves two stages: the training stage and the test stage. The training stage aims at building (and, in some cases, fine-tuning) a model that connects the target behavior to historical customer information by using a training dataset (i.e., a randomly selected subset of the entire original dataset of consumers). This dataset contains a set of independent variables and one criterion variable. The independent variables describe the customers' profiles and their past behavior within a specified period. The criterion variable reflects a subsequent behavior, which is the target of our prediction. The training stage aims to create a model that best captures the relationships between the independent variables and the criterion variable. In the test phase, we deploy that model with the test set (i.e., a new subset

of the original dataset, which was not used to train the model), and we assess the quality of its predictions using one or several of the available performance metrics. Of note, researchers can build different models that are based on different algorithms. They can then compare the models' performance to find and select the model that best suits the data at hand and, as a consequence, makes better predictions.

Performing both the training and the test of the predictive model is necessary to evaluate the prediction of its performance. The standard goal in modelling is to build a model that can capably generalize to new observations similar, but not identical, to the ones that contributed to its development. If we want this approach to be successful, we need to prove the performance of the model on a portion of data that the model has never seen before. The predictions made on the training data could be biased because they could capture idiosyncratic patterns of that specific dataset. In this case, if we were to use this model to predict new cases, it would fail. In technical terms, the model would be overfitting the data. Overfitting happens when the random fluctuations in the training data are picked up and learned as general concepts by the model, impairing the model's ability to make accurate predictions with new datasets.

An important decision regards how much data we should allot to the training and how much to the test set. On the one hand, we would like to use as much data as possible in our training set, so that the model has more examples to learn from. On the other hand, we would like to have a large test set to minimize the variance of our estimate of the model's predictive performance. If the test set is too small, the estimate of the performance could be too strongly influenced by chance. There is no strong consensus on the best proportion of observations for splitting the original dataset. In the literature on churn prediction, we found a variety of splitting decisions ranging from .6 : .4 to .9 : .1. When making this choice, the most important information we should consider is the number of observations in the original dataset: The more data we have, the lower the proportion that should be allotted to the training set.

The prediction of churn is a binary classification task, a type of supervised learning task used for classifying each consumer into one of the criterion categories (i.e., churner and non-churner). We can conceptualize binary classification problems as follows. Let $y_i = \{0, 1\}$ be an observation of the response variable **Y**, which provides information about customer status. When the customer remains active, the y value is equal to 0, and when the customer churns the observed value is equal to 1. Moreover, let X_i be a vector of *k* independent variables that could be associated with the response variable. The aim is to find the subset of independent variables that best predict future churn behavior. To develop our binary classification model, we have at our disposal various data mining techniques: logistic regression (Coussement, Lessmann, & Verstraeten, 2017; Coussement & Van den Poel, 2009), decision tree (Azeem, Usman, & Fong, 2017; Prasad & Madhavi, 2012), neural network (Adwan, Faris, Jaradat, Harfoushi, & Ghatasheh, 2014; Backiel, Baesens, & Claeskens, 2016), support vector machine (Bolancé et al., 2016; Coussement & Van den Poel, 2008), and ensemble methods such as random forest and gradient boosting (Azeem et al., 2017; Jayaswal, Prasad, Tomar, & Agarwal, 2016). There is, at present, no consensus on which technique performs best overall. No reviews on this aspect have been published thus far, and widely varying methodologies and experimental characteristics make direct comparison difficult to carry out, especially when private companies provide the data (Vandecruys et al., 2008).

Considering the literature on churn prediction, we were able to identify two types of research direction. On the one hand, many studies focus on the statistical performance of different data mining techniques. On the other hand, to the best of our knowledge, there are no studies that examine prediction models to extract psychological knowledge. The only extant studies discuss the practical, business and managerial implications of the best-built model (Gordini & Veglio, 2017; Lima, Mues, & Baesens, 2009; Verbeke et al., 2011). Our study is intended to intertwine and develop these two directions. First, we build and test different models using three different algorithms (decision trees, random forest, and logistic

regression). Second, we discuss the practical and, more importantly, the psychological *a posteriori* implications of the model with the best performance.

1.2. Predictive Modeling and Psychology

Many psychologists have recently recognized the power of leveraging data mining and machine learning in psychological research. The main difference between these fields is mostly in their general goals. On the one hand, machine learning focuses more on predicting events as accurately as possible. The nature of the relationship between predictors and criterion is not as relevant as the existence of that relationship. If a variable increases the accuracy of the prediction, in machine learning, the researcher retains it with little or no consideration of what kind of relationship bound that variable and the criterion. What matters is the accuracy of prediction. This approach is why some of the machine learning techniques are sometimes described as "black box" approaches that can produce accurate predictions but are virtually impossible to understand. For example, neural networks are a powerful method, but when researchers in machine learning use them, they only know what the independent variables and the network's predictions are. Everything in between may occur without the researchers' awareness.

Moreover, there is little emphasis on the statistical significance of the predictors. This lack of emphasis is probably because of the effect of the sample size on the estimate of the significance level. In fact, conclusions based on statistical inferences can be ineffective at best and misleading at worst, because with very large datasets, even minuscule effects can become statistically significant. On the other hand, psychologists' primary goal is to explain the relationships among the variables at hand: They develop hypotheses and test them with statistical methods. They care about the nature of the relationship between predictors and criterion, and as such, they carefully examine and interpret the regression coefficients. Trying to join both perspectives means that the question is no longer whether relationships are

"significant" (in large samples, they nearly always are), but whether they are interesting and what the practical implications of the results are.

However, we should not draw a dividing line between these approaches. On the one hand, theoretically driven hypotheses might suggest what variables to measure and include in the model. On the other hand, data-driven predictive models could provide valuable evidence in favour of or against existing theories and suggest new variables to consider in theoretical explanations. Thus, the primary goals of these two approaches might be different, but combining them may lead to more profitable results. The use of machine learning in psychology research should be seen as an opportunity. According to Yarkoni and Westfall (2017), researchers who value interpretability can directly benefit from machine learning approaches in different ways. For example, not all the machine learning algorithms are black boxes. Decision trees, for instance, generate a set of '"if-then" rules that are easy to understand and interpret. Moreover, machine learning concepts and techniques can often increase the efficiency and reproducibility of research.

1.3. The Factors of Customers' Churn

In this section, I want to summarize the main findings on the factors that can affect customers' churn behavior.

1.3.1. Price

Price is considered the top-churning factor in comparison with service quality and loyalty programs (Keaveney, 1995; Lee & Murphy, 2005). While pricing may be the sole factor affecting churn behavior, very often, it is combined with other factors. Pricing includes any rates, fees, surcharges, penalties, or promotional deals. Customers tend to churn not only because the price is high, but also when the price increase is deemed unfair. In such a case, customers have a reference point against which prices are compared. The reference point may be based on past experience with the provider or a decision on what is acceptable for the

value of service rendered. With the different providers competing for the same business, competitive pricing comparison is another way in which a consumer decides if the prices are high. This shows that a change in price can push the consumer to switch the service. As a consequence, pricing strategies should be used to control loyalty (Martensen & Grønholdt, 2010).

1.3.2. Perceived Quality and Value

Perceived quality and value are highly interconnected. Perceived quality is a function of overall quality, reliability, and the extent to which a product or service meets the customer's needs and expectations. On the other hand, perceived value refers to the customer's assessment of the product/service price given its quality. Both perceived quality and value have been shown to be drivers of customer churn behavior.

In a study of bank churn behavior, Gerrard and Cunningham (2004) found that service failure was the major reason that caused bank customers to churn to a competing bank. Michel (2001) explained service failure as the service that fails to live up the customer expectation (Ahmed et al., 2010; Spreng, Harrell, & Mackoy, 1995). Service failure is a driver of churn not only for those customers who experienced it. It can also have an impact on churn behavior of those who heard about service failure because of the spread of negative word of mouth (Boroumand, 2006).

According to Ranganathan and colleagues (Ranganathan, Seo, & Babad, 2006), consumers who are more service-oriented and seek quality service are more likely to churn between the companies. Customers try to avail of quality services and can easily be attracted by the service provider who promises to give quality services to them.

Srivastava and Sharma (Srivastava & Sharma, 2013) stated that in the global market environment, challenging companies are continuously on the lookout for the task that provides superior customer-perceived value and image in order to gain customer loyalty.

1.3.3. Service Encounter Failure

Charkravarty and colleagues (Chakravarty, Feinberg, & Rhee, 2004) found that the propensity to churn was affected by whether or not the customer had reported problems with the bank. Such an effect is enhanced if the service provider fails to address the customer complainants appropriately. According to Sidhu (2005), customers may churn the brand if the provider is reluctant to respond properly or gives a negative response. These failures are related to the human factor in the organization (Awwad & Neimat, 2010). Employees who treat their customers in an impolite and uncaring way create displeasure. Furthermore, unknowledgeable employees can be the reason to push consumers to churn the brand (Tax, Brown, & Chandrashekaran, 1998).

1.3.4. Attraction by Competitors

According to Jones and colleagues (Jones, Mothersbaugh, & Beatty, 2000), customers are aware of the competing alternatives of the brand and they compare them. If a company provides better or differentiated services that are difficult for the competitors to meet, and if there are few competitors in the market, customers tend to remain loyal to the brand (Bendapudi & Berry, 1997). As the competition increases, companies are more focused on distinguishing themselves from other companies in order to attract customers (Sidhu, 2005). Akbar and Parvez (2009) stated that customers switch to other service providers if the competitors are providing more benefit as compared to their existing service provider.

1.3.5. Change in Technology

New services and technology help the company to retain the customers, and this will help them to generate higher revenue. Companies which are not up to the mark with the technology will ultimately be losing their customers (Awwad & Neimat, 2010). Churn behavior moves along with the advancement of technology probably because technology changes the need of consumers (Aamir, Ikram, & Zaman, 2010).

1.3.6. Switching costs

Customers want to maintain relationships with service providers for one of two reasons: constraints (they "have to" stay in the relationship) or loyalty (they "want to" stay in the relationship) (Bendapudi & Berry, 1997). Switching costs are the factors that act as constraints preventing customers from freely switching to other service providers. In the service industry, loyalty points and membership card programs are the major components of switching costs, because all the membership benefits and accumulated points may be lost when service contracts are terminated, or customers switch their service providers.

In a study of 306 subscribers in the Korean mobile service market, Kim and colleagues (Kim, Park, & Jeong, 2004) found that among factors constituting switching costs, loss of loyalty points had both a direct effect and an adjustment effect on customer loyalty. Because current loyalty points were lost as they churned, even dissatisfied customers might show a high level of "false" loyalty (Gerpott, Rams, & Schindler, 2001).

The goal of membership card programs is to increase the rate of customer retention by providing benefits to their members. If the benefit is not available from other providers, it works as a switching barrier from the perspective of customer retention. Additionally, members in loyalty reward programs may overlook or discount negative evaluations of the company against competitors in terms of product, quality, and price (Bolton, Kannan, & Bramlett, 2000). Therefore, such membership card programs establish switching costs for customers, and thus they may inhibit customer churn.

1.3.7. Demographic factors

Past studies have shown that some household demographic factors can impact churn behavior. For example, higher education levels have been found to increase churn behavior in the electricity market (Ek & Söderholm, 2008; McDaniel & Groothuis, 2012). Men tend to be more positive towards churn than women (Gamble, Juliusson, & Gärling, 2009), which might

be interpreted as males being somewhat more likely to favor competition than females (McDaniel and Groothuis 2012). Income is usually identified as an important determinant in churn decisions, but previous work has not reached a consistent verdict. Some studies have found that higher-income households have more positive views of churn than lower-income households and are also more likely to be active in the market (e.g., Ek and Soderholm 2008; Gamble et al. 2009). However, if we consider opportunity costs (i.e., the loss of other alternatives when one alternative is chosen), conclusions about the influence of income on churn may shift or even flip. For example, the value of an individual's time may differ between high-income and low-income customers: High-income customers, by placing a greater value on their time, are less prone to churn than low-income customers (Waddams Price, Webster, & Zhu, 2013).

2. Empirical Research

2.1. Data

This study is based on data from 81836 consumers of the Italian division of an energy company who were active on December 31st, 2016 and held one residential contract for electricity supply. Moreover, we considered only those consumers who buy electricity for a single residential unit.

We considered as non-churners those consumers who had an open contract on December 31st, 2016 and remained active in the subsequent nine months from January 1st to September 30th, 2017 (n = 74915). We considered as churners (n = 6899) those consumers who, in the same period, cancelled their contracts and moved to a competitor. Finally, consumers who, in the same period, incurred any form of involuntary churn (n = 14968) were discarded from the modelling.

We considered various types of potential predictors. These predictors consisted of demographic (e.g., age, regional area), account (e.g., length of the contract in months),

behavioral (e.g., the number of contacts the consumer had with the company call-centres) and socio-economic information (e.g., whether the customer lives in a wealthy or a struggling area). Some of these predictors were categorical; others were continuous. Table 4.1 provides the complete list of the predictors. Some of them were characterized by stability in time (e.g., sex), whereas other variables were measured monthly by the company (e.g., the number of contacts that a consumer had with the company). For the latter type of variables, we considered data from December 1st, 2016, to December 31st, 2016.

Table 4.1. Variables Description.

Variable	Explanation (Categories)	Variable Type	% Missing
Contract starts with a transfer	Transferring the registration of the contract from the former tenant (No; Yes)	Categorical	0%
Acquisition Channel	How has the customer been acquired? (Agency; Call Centre; Counter; Intern; Tele-selling; Web)	Categorical	0%
Customer type	Has the customer a power only or dual (electricity and gas) contract? (Electricity only; Electricity and Gas)	Categorical	0%
Regional area	Geographical area of residence (North West; North East; Central; South; Islands)	Categorical	0%
Age	Customer's age	Continuous	0%
Sex	Customer's sex (Male; Female)	Categorical	0%
Cameo_ITAG	Customer's socio-economic status	Continuous	6,8%
Presence of adults over 60	The probability that the customer lives in an area with lower or higher presence of adults over 60 years old	Continuous	6,8%
Presence of children	The probability that the customer lives in an area with lower or higher presence of children	Continuous	6,8%

Household size	The probability that the customer lives in an area with single-person household or more than 4 persons	Continuous	6,8%
Education	The probability that the customer lives in an area with little or high level of education	Continuous	6,8%
Building age	The probability that the customer lives in an area with older or newer properties	Continuous	6,8%
Loyal Customer	If the customer has subscripted to a loyalty program (Not subscribed; Subscribed)	Categorical	0%
Length of the contract	The length of the contract in months	Continuous	0%
Payment method	The method by which the customer pay the bills (Bank Transfer; RID)	Categorical	0%
Online Billing	If the customer has an online access to his/her bills (No; Yes)	Categorical	0%
Critical requests	Whether a customer addressed one or more critical requests over the previous month (No; Yes)	Categorical	0%
Change offer	Whether a customer changes the offer over the previous month (No; Yes)	Categorical	0%
Number of contacts	Number of contacts with the company over the previous month	Continuous	0%
Retention proposal	Whether a customer has received a retention proposal over the previous month (No; Yes)	Categorical	0%
Cross sell proposal	Whether a customer has received a cross sell proposal over the previous month (No; Yes)	Categorical	0%
Digital customer	Does the customer has used his/her webpage on the company internet site? (No; Yes)	Categorical	0%
Number of previously churn	How many times the customer churned on previous contract?	Continuous	0%
Churn	Did the customer churn? (No; Yes)	Categorical	0%

2.2. Data Preprocessing

Data pre-processing is a long but necessary process that precedes the model-building and evaluation phases. It is a fundamental part of the process we used in this research, the KDD process (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) which is composed of into six phases: business understanding and goal definition, data acquisition, data preprocessing, data transformation, data mining, model evaluation and interpretation. These phases help organizations and researchers understand the knowledge-discovery process and provide a useful road map to follow while carrying out the project.

In the preprocessing stage, the dataset was thoroughly examined to establish its completeness and correctness. We removed inconsistent values, handled missing values and outliers, removed redundant variables, and transformed some variables to obtain more valuable information about consumers.

We eliminated some of the variables of the initial dataset for various reasons. We excluded from the dataset variables with unique values because they do not contribute to predictive modelling. We discarded all the variables that identified the consumers (e.g., consumer ID number), as they do not provide any valuable information to be used for making predictions. We also excluded high-cardinality categorical variables because of their large number of classes and the lack of knowledge on how to categorize those classes.

We handled predictor variables with missing values according to the algorithm used. Specifically, as decision trees can handle missing values directly, we did not impute missing values. In contrast, as logistic regression requires the data to be complete, we imputed missing values with the median value because only continuous variables had missing values.

To assess the quality of a predictive model, we needed to test its performance in predicting the behavior of interest with new dataset that was not used for developing the model. Therefore, we split our original dataset into a training set, containing 70% of the data, and a testing set, containing the remaining 30%. We applied a random stratified sampling to the

criterion variable, which ensured that the percentage of churners and non-churners was equal in both datasets (Table 4.2).

Table 4.2. Percentage of churner in samples.

	Total Sample	Training Sample	SMOTE Training Sample	Test Sample
N of Churner (%)	6899 (8.4%)	4848 (8.5%)	24240 (45.5 %)	2050 (8.4%)
N of Non Churner (%)	74915 (91.6%)	52421 (91.5%)	29088 (54.5 %)	22494 (91.6%)
Total	81813	57269	53328	24544

2.3. Class Imbalance

High class imbalance occurs in real-world domains where the decision system is aimed at detecting a rare event (Kotsiantis, Kanellopoulos, & Pintelas, 2006). The imbalance occurs in the case of churn prediction, where the number of non-churners is far higher than the number of churners. The skewed distribution of data poses great challenges in data mining algorithms (Neslin et al., 2006; see also Weiss, 2004).

Many methods have been developed to solve the problem of class imbalance (He & Garcia, 2009; Weiss, 2010). These methods can be classified into two main categories: algorithm-level and data-level approaches. Algorithm-level approaches are designed to improve the learning task by fine-tuning algorithms. Data-level approaches modify the data distribution through the resampling of the data at a pre-processing stage. The basic idea is to eliminate or minimize rarity by altering the distribution of original training examples using a specific mechanism (Burez & Van den Poel, 2009). Data-level and algorithm-level approaches have acquired great importance because many studies have demonstrated that a balanced dataset improves the overall performance for several classifiers compared to an imbalanced dataset (Babu & Ananthanarayanan, 2016; Zhu, Baesens, Backiel, & vanden Broucke, 2017). However, a consensus on which is the best approach to address the problem has never been achieved.

The basic resampling methods are oversampling and undersampling. Oversampling methods increase the number of minority-class instances, while undersampling eliminates majority examples. Both methods decrease the overall level of imbalance by making the rare class less rare. However, each method introduces its own problematic consequences that can perhaps hamper learning. Undersampling removes instances that may be important and valuable in the learning phase. With the oversampling method, we replicate minority instances by making exact copies of existing ones, and, consequently, the likelihood of overfitting may increase.

To overcome the risk of overfitting related to oversampling methods, Chawla and his colleagues introduced a novel technique, which generates synthetic examples by operating in "variable space" rather than "data space", called the Synthetic Minority Oversampling Technique (SMOTE; Chawla, Bowyer, Hall, & Kegelmeyer, 2002). While random oversampling replicates minority instances from the existing dataset, SMOTE oversamples the minority class by using the K-nearest neighbor graph. At first, SMOTE randomly selects one or more nearest neighbors of a minority-class instance and produces new instances based on the linear interpolations between the original examples and randomly selected nearest neighbors afterwards.

Our dependent variable has a highly skewed distribution: the percentage of churners (the class we aim to predict) is much lower than the percentage of non-churners (8% for churners vs. 92% for non-churners). To make the two classes more balanced, we used the SMOTE resampling method, which resulted in a dataset with 53328 instances, 45% of which were churners and the remaining 54% of which were non-churners (see Table 4.2 for details).

2.4. Techniques

2.4.1. Decision tree

Tree-based methods are a set of nonparametric methods that have significantly increased in popularity because they closely resemble human reasoning and are easy to understand

(Kotsiantis, 2013). A decision tree is a diagram of decision-making rules used to classify instances or to make predictions. It consists of branches and leaves in which the instances are classified. The decision tree classifies each instance by posing a series of questions about input variables. Each node contains a question, and every node points to one child node for each possible answer to its question. An instance is sorted into a class by following the path from the root node to a leaf according to the answers that apply to the instance under consideration.

Various algorithms for inducing decision trees have been proposed over the years and differ in the methods used for selecting splitting attributes and splitting criteria. In our study, we used the CART and C5.0 algorithms.

2.4.1.1. CART

The Classification and Regression Tree (CART; Breiman, 1984) is one of the most popular decision trees in the machine learning community (Jiawei, Kamber, & Pei, 2012). The CART builds binary decision trees by splitting the records at each node according to a specific function. It initially grows the largest tree possible and subsequently prunes it to a size that has the lowest estimate of error. The aim is to remove unreliable branches from the tree to improve its predictive performance. The CART addresses both continuous and categorical predictor variables. It handles missing values by using a series of surrogate splits, which are splits into other variables that substitute for the preferred split when the latter is inapplicable due to missing values. To split a node, the CART uses three types of impurity measures: the sum of squared deviations, the Gini index, and entropy.

2.4.1.2. C5.0.

The C5.0 algorithm (Quinlan, 2004) is the updated and improved version, regarding computation time and memory, of the ID3 and C4.5 algorithms. C5.0 handles both continuous and categorical variables and works by splitting the sample based on the feature that provides the maximum information gain, an entropy-based measure of node impurity. The sample

subset that is obtained from the former split is split afterwards. The process continues until the sample subset cannot be split any further (Patil, Lathi, & Chitre, 2012). C5.0 uses pessimistic global post-pruning to remove additional branches and replaces them with leaves to improve the accuracy of classification (Rokach & Maimon, 2015). It can also handle missing values: When it encounters a missing value, it changes the gain function such that any missing value is sent to every child node with weights proportional to the number of non-missing observations in those nodes.

2.4.2. Random forest.

Despite their ease of use and interpretation, decision trees may lack robustness, and sometimes they can offer suboptimal predictive performance (Hastie, Tibshirani, & Friedman, 2009). To avoid these drawbacks, Breiman (2001) has developed an algorithm that grows many trees (hundreds or even thousands) instead of producing only one, resulting in what is known as the random forest. The random forest algorithm has many strengths. First, it need not to adhere to certain assumptions (e.g., a linear relationship between the criterion and the predictors). Second, the outcomes of the classifier are very robust to outliers and noise (Breiman, 2001). Moreover, random forests are easy to implement because there are only two parameters to be set, namely, *m*, the number of randomly chosen predictors (which is equal to the square root of the total number of predictors), and the total number of trees to be grown. In addition to its potential, the random forest algorithm also has some drawbacks. The major one is the lack of interpretability (the "black box issue"). However, even though it is not easy to determine how each variable affects the prediction (e.g., we are not able to assess the effect size of one specific predictor over the response variable), researchers can create a partial dependence plot that shows how the criterion changes as the predictor changes.

2.4.3. Logistic regression.

Logistic regression is also a popular and well-known technique for predicting dichotomous dependent variables. According to many researchers (Buckinx & Van den Poel, 2005; Gordini

& Veglio, 2017), logistic regression is very appealing for various reasons. First, logit modelling is well-known, conceptually simple and frequently used in marketing, especially at the level of the individual consumer (Neslin et al., 2006). Second, compared to other techniques (e.g., neural networks), logit is easier to interpret. Third, as many comparison studies have proved, it provides good and robust results in general (Neslin et al., 2006). Last but not least, several authors (Levin & Zahavi, 1998) have shown that logit modelling may even outperform more sophisticated methods.

2.5. Evaluation Metric

To assess the performance of both classification methods, we used the area under the Receiving Operating Characteristics curve (AUC). The ROC graph is produced by plotting the rate of the true positive versus the rate of false positive on the y-axis and x-axis, respectively. The AUC is used as a measure of the quality of the classification models, and its values fall between .5 and 1. A random classifier has an AUC value equal to 0.5, whereas the AUC value of a perfect classifier is equal to 1: in other words, the higher the value, the better the performance. This measure is frequently used in customer prediction studies. The greater robustness of its evaluation compared with other methods is the main advantage of this measure: It accounts for the overall performance of a classifier by considering all possible cut-offs on the ROC curve. AUCs are intuitively clear and easy to interpret. Furthermore, the measure is independent of the prior probability of class distributions, which makes the AUC resistant to class imbalance (Oommen, Baise, & Vogel, 2011).

2.6. Tools

We used IBM SPSS 24 (IBM Corporation, 2016) for manipulating and cleaning the data and R (R Development Core Team, 2017) for the modelling and testing phases. More specifically, for building CART trees, we used the rpart package (Therneau, Atkinson, & Ripley, 2017); for modelling C5.0 trees, we used the C50 package (Kuhn, Weston, & Coulter, 2015); and for producing the random forest, we used the RandomForest package (Liaw &

Wiener, 2002). To evaluate model performance, we utilised caret (Kuhn et al., 2017) and AUC (Ballings & Van den Poel, 2013) packages. To resample training data, we used the SMOTE function from the DMwR package (Torgo, 2010).

3. Results

In this study, we aimed to find a churn model for the Italian electricity market. To attain this primary objective, we trained and tested eight different models according to the algorithm used (CART, C5.0, random forest, or logistic regression) and the type of dataset being modelled (balanced vs. unbalanced): Four models based on the three algorithms were trained and tested on the unbalanced dataset, whereas four other models, using the same algorithms, were trained and tested on the balanced dataset. Thus, we first trained each model (and fine-tuned when possible) on the training set. Then, we checked its predictive performance on the test set.

Table 4.3. AUC performance on the training and test set.

Modello	Train	Test
CART Unbalance	0.51	0.51
C5.0 Unbalance	0.51	0.51
CART + SMOTE	0.76	0.56
C5.0 + SMOTE	0.73	0.56
Logistic Regression Unbalance	0.67	0.68
Logistic Regression + SMOTE	0.76	0.61
Random Forest Unbalance	0.99	0.66
Random Forest + SMOTE	0.99	0.63

We report in Table 4.3 the AUC values of the prediction performance on the training set and the test set. The AUC values in the test set range from AUC = .51 (in the CART and C5.0 decision trees on the unbalanced dataset) to AUC = .68 (in the logistic regression on the unbalanced dataset). In the training phase, the models with the best performance were the random forests (AUC = .99). However, the logistic regression model on unbalanced data showed the best overall performance in the test phase.

Therefore, we decided to focus on the latter and interpret, from a psychological point of view, the strongest relationships between the predictors and the criterion.

Before interpreting the logistic model, we decided to refine its main effects by determining whether the relationships between the continuous variables are linear or curvilinear (Hosmer, Lemeshow, & Sturdivant, 2013). To select to which variables added a non-linear effect, we observed the partial dependence plots derived from the random forest model, which depicts the functional relationship between the independent variable and its predictions. The variables that seemed to have a non-linear relationship were the consumers' age and the length of the contract, and both relationships took a U-shape. Therefore, we explored the nature of these relationships between the two predictors and churn using the sequential orthogonal polynomial. The polynomial terms were calculated from the standardized values to avoid multicollinearity (Aiken & West, 1991). We subsequently performed likelihood ratio tests to check whether the model with quadratic or higher-order polynomials would fit the data better than the linear model. As displayed in Table 4.4, the quartic model is better than all the models with lower power and the linear model.

Table 4.4. Likelihood Test Comparisons.

Model	**LogLikelihood**	**Df**	χ^2	***p***
Linear	-15771			
Quadratic	-15743	2	56.49	<.001
Cubic	-15739	2	6.61	.037
Quartic	-15731	2	17.16	< .001
Quintic	-15729	2	3.80	0.150

As the regression coefficients are the logarithmic function of the odds, and thus they are not directly interpretable, we used the exponential function to calculate the odds ratio for each predictor included in the model. Odds ratio values range from 0 to infinity. For dummy predictors (dichotomous and categorical predictors), the odds ratio represents the odds that an outcome will occur given a particular condition, compared to the odds of the outcome

occurring within the reference category. For continuous predictors, this index represents how a unit increase in the value of the predictor affects the odds of the outcome. The odds ratio is a statistic that can be employed as an index of effect size to demonstrate an increased risk for churn: By comparing the odds ratio of different predictors, one is able to determine the strength of the relationship between predictors and outcome (Charry, Coussement, Demoulin, & Heuvinck, 2016). For a correct interpretation of the odds ratio, the continuous variables must be standardize. Thus, we rescaled each continuous variable using the z-score formula.

Table 4.5. Odds Values from Logistic Regression.

Variables: Reference level	**Level of Categorical Variable**	**Odds Ratio Value**
Contract started with a transfer: No	Yes	0.91
Acquisition Channel: Intern	Agency	2.05
Acquisition Channel: Intern	Call Center	1.02
Acquisition Channel: Intern	Teleselling	1.22
Acquisition Channel: Intern	Web	1.18
Acquisition Channel: Intern	Counter	0.45
Customer type: Only electricity	Electricity and Gas	0.91
Regional area: North-West	North-East	1.02
Regional area: North-West	Center	1.14
Regional area: North-West	South	1.28
Regional area: North-West	Islands	1.03
Age		0.87
Age^2		1.00
Age^3		0.99
Age^4		1.00
Sex: Male	Female	0.99
CAMEO_ITAG: Comfortable Families	HomeComfort	1.07

CAMEO_ITAG: Comfortable Families	MiddleClassCommunities	1.04
CAMEO_ITAG: Comfortable Families	ModestMeans	1.27
CAMEO_ITAG: Comfortable Families	ModestSuburbia	1.00
CAMEO_ITAG: Comfortable Families	ProfessionalFamilies	1.05
CAMEO_ITAG: Comfortable Families	ProvincialCommunities	0.95
CAMEO_ITAG: Comfortable Families	StretchedFamilies	1.26
CAMEO_ITAG: Comfortable Families	StrugglingSociety	1.10
CAMEO_ITAG: Comfortable Families	WealthyHouseholds	0.98
Presence of adults over 60		1.01
Presence of children		1.01
Household size		0.99
Education		0.99
Building age		0.99
Loyal Customer: Non-Loyal	Loyal	0.80
Length of the contract		0.94
Length of the contract^2		1.01
Length of the contract^3		0.99
Length of the contract^4		1.01
Payment Method: RID	Postal Bulletin	1.08
Online Billing: No	Yes	0.84
Number of critical requests		1.14
Change offer: No	Yes	0.45
Number of contacts		1.10
Retention proposal: No	Yes	1.51
Cross sell proposal: No	Yes	0.69

Digital customer: No	Yes	0.96
Number of previous churn		1.46

In Table 4.5 we have reported the odds ratio for each predictor for the logistic regression model on unbalanced data. For the categorical variables, we report the predictor's reference level ("Variables" column) and a row for each of its target levels. The predictors most strongly associated with an increase in the likelihood of churn were regional area (specifically, South Italians churn more), type of acquisition channel (more specifically, those customers that subscribe to a contract through in-person contact in a company office), having only the power commodity (as compared to having both power and gas), the number of critical requests (the more the customer complaints, the higher the probability of churn), and having already churned on previous contracts with the energy supplier. The features that decrease the likelihood of churn are starting the relationship with a transfer of contract, the subscription to a loyalty programme, having received cross-sell offers, and receiving the bill online.

We can consider a classification model as robust when the performance on the test set is close to the performance obtained on the training set (Vercellis, 2009). In this sense, not only does the model we selected demonstrate better performance, it is also the one that demonstrates most robustness. In fact, if we compare the performance between the performance of the trained model with the performance of the tested model, we can see how close the AUC values are (only .01 difference).

Another important characteristic that emerges from the results is the stability of the variables chosen by different models. Despite their different statistical nature, the decision tree, the random forest, and logistic regression have all considered as crucial and have discarded nearly all the same predictors. This consistency demonstrates the stability of the conclusions we gathered from our different classification models, and, thus, the robustness of

those variables when predicting the outcome of interest. The predictors that recurrently appear in the models and that we considered are the acquisition channel, length of the contract, consumer's age, the subscription to the loyalty programme, receiving the bills online, and regional area of residency.

Finally, we found an unexpected result, namely, that contrary to the results typically reported in the literature, both the logistic regression model and the random forest had a better performance on the unbalanced dataset than on the balanced dataset. However, these results emerged only with the logistic regression and random forest; in contrast, the CART model and the C5.0 model demonstrated better performance on the balanced dataset. Although these results are not common, they are not new in the class imbalance literature. For example, Zhu and his colleagues (2017) built several churn models using nine different resampling methods and compared their performances, finding that most of the sampling methods (including SMOTE) did not improve the performance of logistic regression on an imbalanced dataset.

4. Conclusions

With this study, we first aimed to build a prediction model able to predict whether a customer would churn from an energy provider. To reach our first goal, we trained and tested eight models that differed according to the algorithm (CART, C5.0, random forest, and logistic regression) and the dataset (balanced vs. unbalanced) used. We built these models using various types of information on consumers: Socio-demographic (e.g., age), behavioral (e.g., number of contacts), account (e.g., customer type), and socio-economic information (e.g., whether the customer lives in a wealthy or a struggling area). After having built the models, we selected the one that had attained the highest AUC value, that is, the model with the best predictive performance in the test phase: the logistic regression model built on the unbalanced dataset (AUC = .68). The selected model has also demonstrated superior robustness in comparison to the other models because its performance on the training set (AUC = .67) is close to its performance on the test set. Moreover, the predictors that were

retained and discarded by the models are nearly the same, indicating that the conclusions we can derive from the classification models are similar to one another.

The selection of the logistic regression model allowed us to achieve the second aim of this study, which was to provide psychological explanations and interpretations through the discussion of the discovered relationships between the predictors and churn behavior. To achieve this aim, we explicate these relationships here in the discussion.

One variable that has proved to be important in churn prediction is acquisition channel. Consumers acquired by an external agency or by an external teleselling service proved more prone to churn, and the only acquisition channel that decreased the probability of future churn is the counter, i.e., a physical place provided by the company where the customer can go and, among other things, request to subscribe to a contract. The method of acquiring new customers is important in shaping and developing the relationship between the consumers and the company (Blattberg, Getz, & Thomas, 2001), both in terms of customer lifetime values and loyalty. At least three different types of psychological processes might explain the relationship between acquisition channel and churn: The first is based on the differences between channels in terms of characteristics of the offer, the second on their differences in terms of types of motivation that drive the subscription to an energy contract, and the third on their differences in terms of characteristics of the consumer.

According to the first explanation, the specific characteristics of the offers that characterize different acquisition channels may affect the length of the relationship. For example, Bolton and colleagues (Bolton, Lemon, & Verhoef, 2004) found that acquisition channels that focus more on price (rather than on brand image or service quality) will create lower customer loyalty, whereas acquisition channels that focus on establishing of economic or social bonds with customers will create more loyal customers. For instance, customers acquired through very attractive offers are also inclined to switch when they receive attractive offers from competitors (Kamakura, Wedel, de Rosa, & Mazzon, 2003). Moreover, using the Internet may

have an important effect on loyalty: Verhoef and Donkers (2005) found that the customers of an insurance company who were acquired using the company's website were more loyal. According to this theoretical explanation, the fact that agencies and teleselling have a greater negative impact on consumer loyalty could be due how those external services acquire new customers. For example, a company may ask an external agency to acquire customers for it by providing a list of prospective customers to be contacted. Later, a different company asks to do the same, and it could be that some of the prospects' names on the list overlap with those provided by the former company. Therefore, it could be that the same customer acquired by the former company is asked to join the latter by offering a new attractive contract, thereby causing churn.

According to the second explanation, the different effects of the acquisition channel on churn may be due to the different role taken by the consumer in the acquisition process. We can classify acquisition channels into two overriding categories. On the one side, there are passive channels, where the company makes an offer to the prospective consumer (e.g., agency and teleselling); on the other side, there are active channels, where the consumer actively searches for a contract with the company (e.g., counter, call centre, and web site). Active and passive channels differ in the degree of control of the consumer over his/her intention and decision. In passive channels, the action of changing the service provider is triggered by an external factor. In active channels, customers are autonomously motivated to initiate a process leading to the change of service provider. Therefore, the level of self-determination related to the choice of changing the electricity provider could decrease the likelihood of churn: The more the customer feels him/herself to be the driver of his/her own choices, the less he/she churns in the future. Self-Determination Theory (Ryan & Deci, 2000) assumes the existence of different types of extrinsic motivations, according to the volitional degree and the degree of internalization of the motivation that underlies the action. Using this theory as a reference frame, the acceptance of an offer could be motivated by an extrinsic

motivation, characterized by a low level of volitional willing and internalization. An extrinsic motivation could also motivate the request for a subscription, but this situation would be characterized by a higher level of volition and a more internalized motivation. This initial motivation could, in turn, impact on the future customer's churn. Behaviors associated with higher degrees of external control are less satisfying, less consistent and produce less optimism in the long run than behaviors motivated by more internal control (Ryan, 1995; Wang, Pyun, Kim, & Chatzisarantis, 2009). Moreover, the more the behavior is self-determined, the higher the probability is of increasing the behavioral effectiveness and its persistence (Roth, Assor, Niemiec, Deci, & Ryan, 2009; Ryan, Rigby, & King, 1993). In this sense, a choice based on more internal motivations might entail a higher level of satisfaction, which in turn could be a protective factor against future churn.

According to the third explanation, the specific characteristics of the customers using certain acquisition channels might affect their loyalty. Customers attracted by different channels might vary in their sociodemographic, psychographic, and other characteristics. For example, in the early days of the Internet, the user population consisted mainly of young, well-educated, high-income people. These customers are known to be less deal-prone and more loyal than other customers (Blattberg & Neslin, 1990).

Transfer of contract was associated with a reduced probability of future churn. When making a transfer of contract, consumers moving into a new house buy the electricity from the same provider that supplied it to the former tenant. The transfer of contract from the former tenant and the initiation of a contract with a different supplier might be the outcome of different types of decision-making processes. When they approach decision-making, consumers tend to fit into one of two categories: satisficers and maximisers (Schwartz et al., 2002). Satisficers consider a limited range of alternatives, with the purpose of finding an option that satisfies the given criteria, i.e., an option that is considered satisfactory or "good enough." On the other hand, maximisers aim at the best possible option and seek information

about as many alternatives as possible before making their choice. When selecting an electricity provider, maximisers may be involved in a longer and more careful search for the best service. For example, they may want to achieve the best price on the market or to find a company that shares the same values as the consumer (e.g., green energy). However, maximisers appear to regret their decisions more often than satisficers and to be unhappy (Peng et al., 2018; Schwartz et al., 2002), they experience more post-decisional dissonance than satisficers (Iyengar, Wells, & Schwartz, 2006), are more prone to change their initial decisions (Chowdhury, Ratneshwar, & Mohanty, 2009), thereby expressing considerably low brand loyalty (Lai, 2011), and tend to be fixed on realized and unrealized options (Iyengar et al., 2006). The lower level of satisfaction of maximisers with their choices and their preference to seek numerous options could increase their sensitivity to competitors' campaigns and their proneness to searching for alternatives, even after having signed a contract with the provider. As a consequence, the probability of churn of this type of consumer might be high. Satisficers, in contrast, who prefer to select the "good-enough alternative" rather than to evaluate all possible solutions, may be less affected by external triggers, such as unfavorable price increases or competing offers. In sum, the choice to transfer the contract from the previous tenant might signal a satisficing decision-making strategy, which might be associated with higher passive loyalty and correspondingly lower levels of churn.

According to Neslin and his colleagues (2006), there are two main approaches of customer retention strategies: the untargeted and the targeted strategy. Untargeted strategies focus on strengthening loyalty through service improvement (e.g., cross-selling, change of the service offer) and mass advertising. Targeted strategies rely on giving special incentives, such as reduced pricing or customized solutions, to those customers who are at higher risk of churn.

Concerning untargeted strategies, the model shows that having received a cross-sell proposal in the last month decreases the probability of future churn. Cross-selling is a way to

establish stronger relationships with customers (Kamakura, Ramaswami, & Srivastava, 1991) and reduce the probability of churn. As a consumer acquires additional services or products from a company, the number of points at which consumer and company connect increases, leading to a higher switching cost for the consumer. However, the effect of cross-selling on consumer churn might be mediated by other characteristics of the consumers. Indeed, those customers receiving a cross-selling proposal might be chosen by the company not at random but based on characteristics that are the real reason why their level of churn is lower. Moreover, the factor decreasing the risk of churn could be the fact of having purchased the product and not the proposal *per se*. The other untargeted strategy, which is the change of the offer related to the service, has a positive impact on retention. Moreover, the odds ratio for this strategy suggests that the change of the offer is a more powerful untargeted strategy than other cross-sell incentives (e.g., selling lightbulbs or the air conditioners) for strengthening the relationship with consumers.

Concerning the targeted strategies, retention offers have a positive impact on churn: Those customers that have received a retention offer are more likely to churn. The effect that a company wants to achieve from a retention campaign is to keep their customers. However, retention campaigns might end with a loss. We propose that two different reasons could explain the failure of retention strategies. First, and assuming that the company has correctly identified potential churners, the offer the company puts forward might not be sufficient to convince customers to stay. In this case, the retention strategy may not be concerning a salient aspect of the service from the standpoint of the customer. For example, if the customer is price-sensitive (e.g., he/she wants to pay less) and the company advances a proposal that implies a price increase, the customer might become unsatisfied and thus decide to leave the company. Thus, instead of keeping customers, the company has, in this case, pushed them away by causing a drawback effect. Second, the company may have addressed the offer to those customers who did not need it. As we stated in the introduction, one of the benefits that

come from the building of churn prediction modelling is that businesses can use it to identify actual potential churners and address appropriate retention offers only to them. When a company relies more on “gut feelings” and less on predictive modelling, it could be possible that they may have identified false churners (e.g., they have considered a customer a potential churner who was not) who did not need company’s attention. The effect we found is rather unexpected. However, it allows us to pay attention to a phenomenon that has been understudied so far. This effect could prompt further examination of the drivers behind the failure of retention campaigns.

Another variable that predicts customer churn is the geographical area of residency. South Italians churn more frequently than Italians from other regions. This result could be due to the company’s differing presence in Italian regions and different availability of services in those areas. The energy provider was established in Italy quite recently and is characterized by a stronger presence in the northern areas of the country. Therefore, since the company has not fully extended its services in southern Italy, it could be possible that the southern customers are prone to be less loyal.

If a customer has already churned previously on a contract with the same company, the probability of churn again increases. As we stated in the introduction, market deregulation has increased competition, giving the customer the power to choose the best provider. As a result, electricity retailers face a unique problem: customers tend to churn repeatedly, continuously migrating from one provider to another. The repetitiveness of churn behavior may be due to different reasons. First, the customer decides to voluntarily return to the previous company after experience with another company. Second, the customer is reacquired by the company through the formulation and implementation of win-back strategies. Customer win-back is the process that companies use to revitalize their relationships with customers who have defected (Stauss & Friege, 1999). When companies assess reacquisition opportunities, these opportunities are evaluated based on the likelihood of reacquisition and the value of the

customer in terms of Second Lifetime Value (SLTV; Stauss & Friege, 1999). However, after reacquisition, companies should understand this renewed relationship in order to fortify customer retention strategies and contain the looming risk of churn. The key to comprehending the second life of a customer is to consider a customer's prior experience with the company. Based on the reasons for the first-time churn, companies can identify room for improvement and effectively engage in win-back dialogue with lost customers. If the company succeeds in identifying the underlying causes of the previous churn, customers may not churn for the same reason in the future. However, this approach does not apply to every reacquired customer. There are customers whose return is more likely to be encouraged by the incentives included in the win-back offer (e.g., cost-sensitive customers). These customers may churn again for any reason, including the one that caused their first-time churn. Nevertheless, both types of customers are still at risk of defection in their second life, whether for the same or other reasons. Therefore, understanding second-time churn remains a crucial aspect of managing reacquired customers: Companies need to know when and why a returning customer will churn again and how this propensity to repeat churn changes over time (Kumar, Leszkiewicz, & Herbst, 2018).

Critical requests (e.g., complaints) are also related to the probability of churn. Specifically, the more the customer complaints, the higher the probability of churn. This effect is in line with the studies in the literature, according to which the propensity to churn is affected by whether or not the customer has complained with the bank (Chakravarty, 2004). A psychological explanation of the effect may lie in the fact that one function of complaining is the emotional release from the distress and frustration (Alicke et al., 1992) elicited by a bad experience with the service (e.g., a disservice in the provision of energy or problems with invoicing). In general, dissatisfaction is the attitude resulting from disconfirmation of customer's expectancies and standards, and complaining is a behavioral expression of that dissatisfaction (Kowalski, 1996). Although experiencing adverse events increases the

customer's dissatisfaction level, it does not necessarily influence the customer's complaining behavior: It seems that there is a complaint threshold that must be reached before someone decides to complain, and this threshold is more personal than objective. In the presence of a discrepancy between actual and desired states, the person's dissatisfaction threshold is lowered, and the subjective experience of negative emotions increases. Motivated to reduce the perceived discrepancy that caused the feeling of dissatisfaction, the person evaluates the utility of complaining. If the perceived utility is high and, therefore, the complaining threshold is low, the person will complain. Otherwise, if the perceived utility of complaining is perceived to be low, the individual's complaining threshold will increase and complaining will not occur (Kowalski, 1996). Therefore, those customers that contact the company to complain must be treated carefully by the company, because they not only had a bad experience with the service but also perceived negative emotional states towards it. In fact, we can argue that it is not the complaint, *per se*, that drives churn behavior but the negative emotions and the dissatisfaction that the complaint carries with it. When customers express their dissatisfaction with the service by complaining, this does not necessarily mean that these customers will churn, because that behavior could depend on how those complaints are recognized and treated by the company (Coussement & Van den Poel, 2009). As a matter of fact, well-managed complaint management programs can substantially reduce customer churn (Fornell & Wernerfelt, 1987).

The logistic regression model evidenced receiving bills online, through one's personal web page on the company's website, as a relevant predictor. Specifically, those customers who decided to receive their bills online were less prone to churn. The growing use of the Internet channel among customers is hypothesized to have an impact on customer retention. There are two possible explanations for potential differences in customer retention between users of Internet channels and users of traditional channels. The first explanation claims that it is the usage *in sè* of Internet channels that cause changes in customer's retention. Some studies have

supported the existence of a positive linkage between the use of the Internet and retention rate (Hitt & Frei, 2002; Mols, 1998). Different authors have explained that this effect might occur because the online environment offers more opportunities for personalized marketing, greater flexibility and convenience to the customer (Srinivasan, Anderson, & Ponnavolu, 2002; Wind & Rangaswamy, 2001) and create additional switching costs (Hitt & Frei, 2002). In contrast, some authors did not support the occurrence of a positive linkage, arguing that in the Internet era, the competition is only a few clicks away. The Internet has given customers the opportunity to compare competing offerings with minimal or no costs, thus creating tougher competition that reduces customer retention (Ansari, Mela, & Neslin, 2008). The second explanation for potential differences in customer retention between Internet users and users of traditional channels claims that the effect is due to the self-selection effect (Heckman, 1990). Customers are usually offered the free choice to select a channel through which to interact with the company or to receive the company's services (e.g., receiving bills online). Therefore, it could be that it is not the usage of Internet channels that has an impact on the retention. Instead, there may be certain specific characteristics of the customers that might affect their intrinsic preference for Internet channels over traditional channels (Degeratu, Rangaswamy, & Wu, 2000) and, as a consequence, have a direct impact on customer retention. The results of our model leave room for this second explanation since the company provides the Internet service, but it is ultimately the customers who decide whether or not to use it: By deciding to use the service, customers might self-select themselves on churn-relevant characteristics.

Another variable that has proved to be important in churn prediction is being enrolled in a loyalty programme. Specifically, being enrolled in such a programme decreases the probability of future churn. Companies can come up with different loyalty programs to retain their customers by offering to those customers benefits and more personalized services. For example, the energy retailer in this study offers easy-to-read bills, locked energy prices for a

given period and a personable agent. Subscription to a loyalty programme might increase the perception of contractual switching costs (Klemperer, 1987) and therefore decrease the proneness to churn. There are numerous reasons that could explain this phenomenon. For example, the awareness of contractual switching cost may increase because all the membership benefits accrued by the consumer will be lost once he/she switches service providers. As shown in the literature review in the introduction, the higher the switching cost, the more loyal the customer (Caruana, 2003). However, another possibility is that membership to the loyalty programme, per se, has no impact on churn, and the observed relationship is spurious, due to some a priori characteristics of loyal consumers that decrease the likelihood of churn. For instance, consumers who more strongly self-identify with the brand might have a higher probability to sign for a loyalty programme. It is also possible that marketers do not offer the loyalty programme to each customer indiscriminately but instead choose to which consumers to propose it according to their behavior or characteristics. If so, the common characteristics of loyalty members might be the reasons behind their lower churn rate.

The logistic regression analysis has offered insights into the importance of the contract's length. Our model shows that the longer the lifetime of the relationship between the customer and the company, the less likely the churn behavior. That effect is commonly found in the literature on churn prediction, and in different contractual industries. It is unlikely that this effect is related to contractual characteristics (e.g., having a contract with fixed length or the presence of penalty costs whenever the customer decides to close the contract before its expiration). In fact, as a result of the liberalization of the market, all electricity customers can choose to move to a competitor whenever they decide to do so, without any costs. Therefore, the length of the contract could be considered as an indicator of intrinsic loyalty, that type of commitment that goes beyond the subscription to a loyalty programme and taps into a consumer's thoughts, feelings and emotions, reaching beyond just logic and cost/benefit

analysis. In this sense, the more loyal the customer is, the longer he/she wants to stay with the company. From a managerial point of view, this means that a company's retention efforts should be more addressed to short-term customers to increase their intrinsic loyalty and, as a consequence, make them more profitable.

Some of the customers had both power and gas contracts with the same provider, while others had only a power contract. The results indicated that customers who bought both energy and gas from the same provider showed a lower probability of churn. This result is in line with the findings of studies conducted in different markets. In retailing, customers who buy multiple product categories from a firm tend to have longer profitable lifetime duration (Reinartz & Kumar, 2003), and in the insurance market customers with two home insurance policies are less likely to leave (Günther, Tvete, Aas, Sandnes, & Borgan, 2014). Subscribing to a contract of purchase for two services (gas and electricity) could be assimilated to a kind of re-purchase behavior. The effect of repurchasing on customer churn has been extensively proved: The more customers buy, the higher their loyalty (Meyer-Waarden, 2007). When the consumer re-purchases, the relationship between the consumer and the company becomes closer. If the consumer-company relationship reaches a high level of closeness, the loyalty of the consumer becomes stronger (Mishra, Sinha, & Koul, 2017). Furthermore, and similarly to the effect of the loyalty programme, buying both commodities from the same company could also increase the perception of the cost related to switching behavior. However, unlike the subscription to the loyalty programme, perception of increased cost is more related to the time and effort expenditure when searching for a new provider (set-up switching cost; Burnham, Frels, & Mahajan, 2003). In this sense, switching both commodities to another provider entails higher switching cost than the switching of only one service. As a consequence, customers that have both electricity and gas commodities with the same company are more prone to be loyal.

Conclusions

This research makes several contributions to our knowledge of consumers' churn behavior in the electricity market, a market that is currently undergoing substantial changes and competitiveness. The literature related to the use of machine learning in psychological research is still scarce and deserving of research efforts. We discussed how knowing the characteristics and behaviors that can affect churn behavior might be useful not only from a managerial point of view but also from a psychological point of view. The relationships we identified between the predictors and the behavioral outcome help shed light on the psychological drivers of such behavior. Therefore, psychologists can derive useful insights for future experimental research from this kind of study. The joint use of these fields in the research projects, with the aim to establish a fruitful combination of them, is increasingly gaining attention.

By comparing the models' performance, we demonstrated that the logistic regression technique outperforms the other models. Moreover, the logistic regression has shown to be robust: In fact, the difference between the AUCs on the training set and the test set is narrow. Moreover, it should be noted that despite their statistical nature, the logistic regression model and the tree-based models identified and used nearly all the same predictors. We can consider the stability of the models' selected variables as an indicator of the robustness of the predictors we utilized in the models.

Although this research makes contributions to our knowledge of consumers' churn behavior, there are limitations and future research extensions that need to be noted. First, we acknowledge that there is a broader range of factors that may influence customer churn. Certain factors, such as educational background, marital status, income level, or the usage of the personal web account were not included in this study due to their unavailability. In future research, we could consider these variables and make use of different machine learning algorithms to obtain better predictions of the outcome.

This study used data from a large dataset from one Italian electricity retailer, and through that, we were able to understand better consumer churn behavior. A possible advance of this research could pertain to the examination of multiple retailers in order to observe whether the variables that predicted churn behavior in this study can also predict it in a different context.

Contrary to other studies that made use of explicit measures (e.g., interview, questionnaires), the effects we found in the model are derived from the actual behavior of electricity consumers. Responses to explicit measures can be considered as verbal expressions of the attitudes, behaviors, and opinions that may be influenced by many factors, in addition to the evaluative associations in one's memory, for example, self-presentation concerns, question comprehension, use of appropriate standards of comparison, and the fallibility to retrieve the response from explicit memory (Perugini, Richetin, & Zogmaister, 2010). In this sense, the method we used to retrieve psychological explanations can be assumed as an implicit measure, as we did not ask the customers to describe their feelings towards or evaluations of the company's services. Instead, we used their actual behaviors, predicted the outcome and inferred psychological relationships between predictors and outcome. Moreover, the use of consumers' actual behavior has permitted us to circumvent the weakness of explicit measurement, and therefore, has increased the ecological validity of our study.

Finally, the study we conducted is correlational. With this model, we have identified several relationships between the predictors and the outcome that could be causal (e.g., subscription to the loyalty programme). To disentangle the causality of the effects we found, researchers could conduct new experimental studies, for example, to understand if and what aspects of the loyalty programme have a direct effect on the retention rate. In this case, we can conduct an experimental study on a sample randomly selected from the energy retailer's customer base, casually ask some of them to subscribe to a loyalty programme and predict the effect of the subscription and its incentives on churn probability. That is an example of how

analyzing a large dataset and using data mining can feed and support psychological experimental research.

CHAPTER V

CONCLUSION, CONTRIBUTIONS, AND FUTURE RESEARCH

The general aim of this dissertation is to understand how Big Data and data mining can be fruitfully used in psychological research, while the specific aim is to investigate their use in consumer psychology to understand customers' churn, or in other words to put light on the correlates of customers' stopping doing business with a company. The availability of large quantities of data, new generations of technology, and breakthrough algorithmic architectures have changed the scale and aims by which academics can empirically address critical questions (Wenzel & Van Quaquebeke, 2018). In recent times, Big Data has acquired an increasingly important and central role also in the psychological community. The use of Big Data provides unprecedented opportunities and challenges for understanding human behavior and cognition. The Big Data approach extracts analyzable features from raw information about people and derive quantitative answers to substantive research questions (Chen & Wojcik, 2016).

In the following sections, I provide a recapitulation of the findings of this dissertation, and I discuss both theoretical and practical implications. I end by discussing the limitations of the conducted researches and by identifying directions for future research.

1. Recapitulation of findings

In Chapter II, I provide technical knowledge to those psychologists who want to challenge themselves with Big Data. Although psychologists often show a sense of excitement when talking about the opportunities that Big Data entail, this enthusiasm has not yet let to extensive use of Big Data in the psychological community. According to Paxton and Griffiths

(2017), there are three main reasons why the application of Big Data is still limited in psychology: imagination, cultural, and skills gaps. Bridging the first two will take some work to adjust the psychology field's idea of the possible scope of data beyond data generated through traditional methods and the opportunities of Big Data-driven researches. Another way to bridge these gaps lies in providing psychologists with concrete examples of the use of Big Data in psychological research. If more psychologists start to conduct research with Big Data and show its effectiveness in increasing psychological knowledge, then other psychologists may be encouraged to conduct this type of research. The best way to bridge the third gap is by providing psychologists with knowledge on the tools, methods, and techniques useful to conduct Big Data researches. In Chapter II, I aim to make my contribution to bridging the skill gap. By taking the Knowledge Discovery from Database (KDD) steps as the *fil rouge*, I show where to find data suitable for psychological investigations, the methods to preprocess these data, the techniques through which analyze it, and the programming languages through which all these steps can be implemented. In this chapter, I present many procedures and techniques that I think are essential for psychologists who want to conduct Big Data projects.

Moving from the general question to a specific research topic, in this dissertation I deal with the comprehension of the psychological underpinnings of consumer behavior by applying data mining and machine learning on customer data. The use of such methods on customer data, albeit it is new to the psychological literature, is not new in the CRM literature. Data mining techniques in CRM assist businesses in finding and selecting the relevant information that can then be used to get a holistic view of the customer lifecycle. Thus, in Chapter III, I aim to understand how data mining has been used in the CRM literature. Specifically, in Chapter III, I present a literature review on the use of data mining in customer acquisition, cross-selling, churn, and win-back processes. I discuss the characteristics of the studies (e.g., study type, aims), the data mining techniques being used, the industries where data were retrieved, and the predictors that showed to be related to

customer behaviors of nearly 500 works. The results show how the application of data mining in CRM has continuously increased over time, especially for customer churn. However, the findings evidenced that the use of data mining on the win-back process is entirely lacking.

The high number of researches on churn has guided the choice of analyzing customer churn behavior in the empirical study presented in Chapter IV. Indeed, I reasoned that the knowledge created on customer churn in the last 18 years would have allowed me to understand better both the data mining methods, the characteristics of customer churn behavior, and its predictors. Although important, knowing what the predictors of churn behavior is just a part of the understanding of the phenomenon. The other part is trying to understand why customers churn. However, none of the revised works considered the psychological motivations that may have caused customer churn, even though data-driven CRM processes may benefit from psychological explanations of customers' behavior. In this sense, the empirical research presented in Chapter IV is the first in which psychological explanations have been extracted from a churn prediction model.

Regarding the data mining techniques, the review shows that classification models are the most applied models in the considered CRM processes. Overall, the most used data mining classification techniques are ensemble models (especially Random Forest), decision trees (especially C4.5), and regression (especially Logistic Regression). The fact that these techniques are used more than "black-box" techniques (e.g., neural networks) can be explained by the fact that interpretable models make it easier to understand the relationships between predictors and target behavior and use such information to direct marketing actions. Another main finding concerns those studies that investigated the predictive power of additional customers' data. The results show that information on customers' social network (e.g., family, friends) is a valuable information for the prediction of churn (Backiel, Baesens & Claeskens, 2016). The number of different products acquired per product category, the acquisition order, the time until first acquisition within a product category, or time between

repeated purchases were the main contributors to the improvement of cross-selling model predictive performance (Prinzie & Van den Poel, 2006). Companies' information retrieved from web pages increments the capacity of the data mining model in identifying good prospects (D'Haen, Van den Poel & Thorleuchter, 2013).

The results of the review do not only show the state of the art of applying data mining to customer data for directing CRM processes. They have been of great help in the development of the empirical research presented in this dissertation. In a certain sense, the literature considered in the review has been the teaching materials on which I built my knowledge on the use of data mining techniques on large customer datasets. I learned how to preprocess such data, what techniques to use for building the predictive model, and for testing the quality of predictions. In addition, they have allowed me to understand which information about customers is more predictive of customer behaviors, thus guiding the selection of theoretically relevant variables and creating expectations on the possible results. All this information guided the way through which the empirical research was conducted.

In Chapter IV, I apply data mining and machine learning techniques to customers' data from an energy retailer to understand the psychological underpinnings of customers' churn behavior. By comparing the models' performance, I demonstrate that the logistic regression technique outperforms the other models. Moreover, the logistic regression shows to be robust and it identified and used nearly all the same predictors selected by the tree-based models The aim of the study is not limited to the construction of a predictive model of churn behavior, but also to interpret the relationships between predictors and the behavioral outcome to derive psychological explanations of such relationships. To interpret the predictive relationships, I calculate the odds ratios because they quantify the direction and the strength of the predictive relationships. The features that increase the likelihood of churn are the regional area, the type acquisition channel, the type of the contract, and whether the consumer has already churned on previous contracts. Furthermore, the features that decrease the likelihood of future churn

are the subscription to a loyalty program, the length of the contract, and having received cross-sell offers. As mentioned in the second chapter, the KDD process is circular and iterative. Therefore, the achieved results and the interpretations provided are not the end of the process, but a new beginning. The psychological interpretations will allow (me or other psychologists) to continue this process by creating new research hypotheses and new experiments through which testing the causality of the relationships that emerged from the data. Thus, these results can be a starting point for new experimental research, either traditional (e.g., in a lab setting) or based on Big Data.

2. Theoretical and Practical Contributions

In this dissertation, I showed that Big Data and data mining techniques can be used in psychological research to advance our knowledge of people's behaviors and open to new research opportunities.

In Chapter II, I give my contribution to bridging the gap of psychologists' technical skills by introducing the reader to relevant tools and techniques to find and analyze large datasets. Thus, the contributions of the chapter are principally practical. Nevertheless, throughout the chapter, I also have discussed some relevant methodological issues and highlighted some related pitfalls that must be considered when applying data mining and machine learning techniques in psychological research. I am convinced that data mining and machine learning concepts, such as feature extraction techniques, out of sample validation, data mining techniques, and methods of interpretable machine learning, will positively contribute to the generalizability and robustness of psychological studies. Moreover, data mining and machine learning methods provide ways that may positively contribute to solving the replication crisis that is affecting the psychology field.

In Chapter III, I contribute to the literature on CRM by providing a comprehensive literature review on the use of data mining and machine learning in customer acquisition,

cross-selling, customer churn, and customer win-back processes. This is the first identifiable academic literature review on the use of data mining methods and techniques in CRM that covers an extensive period (2000-2018). Moreover, it is the first literature review that comprises both journal articles and conference proceedings (e.g., Ngai, et al., 2009). This choice permitted us to provide a more comprehensive overview of the knowledge produced in the last 20 years. The review findings have important implications. Indeed, it provides a roadmap to guide future research and facilitate knowledge accumulation and creation concerning the application of data mining techniques in CRM. The findings may be useful not only for academics of the business field but also for those psychologists who want to use customer data to derive psychological knowledge, as I did. Moreover, by making the raw data available on the Open Science Framework [7], I permit the reader, either academic or practitioner, to examine in detail the contents that may be of major interest. For example, researchers who want to know which customers' information proved to be predictive of customer churn, can find in the "Aims.csv" all the works that aimed to identify the predictors that affect churn behavior.

The research reported in Chapter IV makes several contributions, both theoretical and practical.

I provide evidence of the usefulness of using data mining and machine learning techniques in the context of psychological research. In the Introduction, I stated that one of the reasons why using such techniques in psychological research is that a data-driven approach can shed some light on novel (ir)regularities, which can generate new hypotheses and, hence, new experiments. In Chapter IV, I fulfill this function. The relationships I found open to a wide range of research opportunities probe new research questions, and hence new research opportunities. For example, for understanding if and what aspects of the loyalty program have

[7] https://drive.google.com/open?id=1TEhpoV8uVPldbuZvjNoEoFniLgHBojfe

a direct effect on the retention rate, we can conduct an experimental study on a random sample extracted from the energy retailer's customer base. Then, casually ask half of them to subscribe to a loyalty program and predict the effect of the subscription and its incentives on churn probability. Thus, data-driven research will not replace the traditional ways we do psychology. Data-driven research provides us with opportunities to see how different aspects of the environment are related, but they cannot tell us what factors cause specific behaviors. To do that, psychology needs to continue doing the kind of experimentation that has been central to the field for the last century. Nonetheless, data-driven does have great potential to be an important tool for understanding people's behavior. Indeed, modeling customers' data allowed me to recognize the hidden value of consumer's characteristics that have never been considered before, possibly because of the theory-driven approach that typically characterizes psychological research.

The empirical study shows another advantage of using large datasets. The possibility to analyze a large number of customers' data has allowed me to study a phenomenon that otherwise would be challenging to study with traditional research. Indeed, customer churn is a behavior that occurs in small proportions (e.g., in the analyzed data, only 8% of customers churned). When you have something that is a relatively rare outcome, like churn, doing research in a traditional way, gathering assessments, following people over time, can be very costly and very difficult. Consider a researcher who wants to study the changes in depression levels in women who developed breast cancer. For doing that, she should evaluate a large number of women for depression, wait years to see if any developed cancer, and then re-evaluate them. Thus, psychologists that embrace the use of Big Data in their research have the capability to capture rare events more easily and in greater quantity than those who prefer to use traditional research.

The study advances our knowledge of the psychological underpinnings of consumers' churn behavior. Although the level of performance of the model was modest, we have

achieved our primary objective as long as the predictive model enabled us to reach a first psychological understanding of customer churn behavior. For instance, having received a cross-sell proposal decreases the probability of churn. One possible explanation is that, as a consumer acquires additional services or products from a company, the number of points at which consumer and company connect increases, leading to a higher switching cost for the consumer. However, the effect of cross-selling on consumer churn might be mediated by other characteristics of the consumers. Indeed, those customers receiving a cross-selling proposal might be chosen by the company not at random but based on characteristics that are the real reason why their level of churn is lower. For this reason, a subsequent study in which cross-selling is manipulated in an experimental design, guided by the present research, would have the valuable result of allowing to go from prediction to explanation, which would have not only theoretical value, but also practical one, as we could be more confident in the robustness of the model. Indeed, combining new research with the theoretical insights from the prediction model may result in better, more robust models and prediction (Mohebbi, et al., 2011). Thus, prediction without explanation may negatively affect the predictions and the trust in the conclusions that can be derived from a model. For instance, Ginsberg and colleagues (2010) used machine learning to choose 45 Google search terms from 50 million queries and developed a prediction model that can accurately predict flu pandemics faster than the official disease control and prevention agency. However, researchers later found that the model completely missed nonseasonal influenza, suggesting that it predicted seasonality rather than the actual flu trend (Lazer, et al., 2014). The failure of Big Data use in this case stresses the importance of using theoretical insights to guide the research design. If predictors were chosen based on theoretical relevance, seasonality would have been included in the model because seasonality is strongly associated with flu pandemics.

Other than contributing to the psychological literature, these findings also contribute to the CRM literature as long as there are no applications of data mining techniques for predicting

customer churn in the energy sector. As the literature review evidenced, of the 460 works on customer churn published in the last 18 years, only two used customers' data from energy retailers (De Caigny, Coussement, & De Bock, 2018; Moeyersoms & Martens, 2015). However, the focus of both papers was not directed to the peculiarities of churn in the energy market but to other aims (i.e., the former on a new ensemble algorithm for classification and the latter on how to handle high-cardinality variables).

3. Limitations and future research

The study conducted in Chapter IV is correlational. While correlational research can suggest that there is a relationship between two variables, it cannot prove that one variable causes a change in another variable. In other words, correlation does not equal causation. Thus, correlational nature limits the type of considerations we can derive from the results. However, I consider the findings of the research as a starting point from which develop new experimental research either in laboratory settings or, as I would like to do, in "natural" settings. For example, to disentangle the causality of the effects that I found, I would conduct controlled experiments on random samples of customers from the energy retailer customer base. As stated repeatedly in this dissertation, data-driven research can complement traditional approaches by serving as a proving ground for theories developed in rigorously controlled experiments, and by directing new research opportunities.

Moreover, the combination of information about "what a customer does" with a deeper understanding of "why a customer does it" offers opportunities to not only understand better the psychological underpinnings of customers' behavior but also to boost the effectiveness of data-driven marketing campaigns.

Although I provide an extensive overview of the tools and techniques to conduct Big Data projects, I do not claim that it is exhaustive. Indeed, I described what I learned and what was useful in my experience for conducting researches during my Ph.D. This explains why I have

dedicated ample space to the description of predictive algorithms and the model construction phase. Thus, Chapter I can be particularly helpful for those psychologists who aim to predict people's behaviors (e.g., churn), psychological states (e.g., emotions), or psychological traits (e.g., personality). However, it could also be helpful to provide information on other applications of Big Data in psychological research. For instance, the data preprocessing section may be integrated with information on how to treat unstructured data (e.g., tweets), and the data mining section may include information on text mining or topic modeling. Moreover, the guide I presented provides technical information on the methods for preprocessing data and the techniques for analyzing them. To further boost its effectiveness, the guide can be complemented with additional materials on which psychologists can put the hands and start learning by doing, such as examples of coding scripts in R and datasets. In this sense, the open data culture gives a helping hand. Indeed, more and more researchers are willing to share their research data. Data that can be used by psychologists to get their hands dirty and to conduct their research on these data.

A far-reaching development would aim at the bridging of the imagination and the cultural gaps. Although trying to bridge the skills gap contributes to the dissolution of the other two, it is not enough. The psychology community should make more efforts to raise the profile of Big Data researches. Departments may help by developing courses at the intersection of Big Data and traditional research, eventually by teaming up with computer science departments. Most importantly, researchers who are already actively engaged in Big Data projects should consider ways to contribute to the community change. For example, through teaching workshops, participating in conference panels and online activities (e.g., social media and blogs), and joining the communities-focused initiatives (e.g., Data on the Mind; Paxton & Griffiths, 2017).

4. Concluding Words

The Big Data era is happening, bringing with it massive, multimodal, temporal data. Big data approach carries the promise of improving some predicaments in psychology traditional research zeitgeist. I hope I have been able to convince you that Big Data and data mining can offer useful practices for psychology research. Big Data provides structure to and extracts analyzable features from real-world behaviors and derives quantitative answers to substantive research questions. Even though Big Data projects require considerable efforts in learning the necessary skills, for those who are willing to take such a commitment, innovative analyses of unexpected or previously untapped data sources can offer fresh ways to develop, test, and extend psychological theories.

REFERENCES

2018 Kaggle Machine Learning & Data Science Survey. (2018). Retrieved July 26, 2019, from https://www.kaggle.com/kaggle/kaggle-survey-2018

Aamir, M., Ikram, W., & Zaman, K. (2010). Customers' Switching in Mobile Phone Service Providers... - Google Scholar. International Journal of Business Management and Economic Research, 1(1), 34–40. Retrieved from https://scholar-google-it.proxy.unimib.it/scholar?hl=it&as_sdt=0%2C5&q=Customers'+Switching+in+Mobile+Phone+Service+Providers+in+Pakistan&btnG=

Abbasimehr, H., Setak, M., & Soroor, J. (2013). A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques. *International Journal of Production Research, 51*(4), 1279–1294. https://doi.org/10.1080/00207543.2012.707342

Abbasimehr, H., Setak, M., & Tarokh, M. J. (2014). A comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction. *International Arab Journal of Information Technology, 11*(6), 599–606.

Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G., & Choudhury, T. (2016). Automatic detection of social rhythms in bipolar disorder. Journal of the American Medical Informatics Association, 23(3), 538–543. https://doi.org/10.1093/jamia/ocv200

Adaji, I., & Vassileva, J. (2015). Predicting Churn of Expert Respondents in Social Networks Using Data Mining Techniques: A Case Study of Stack Overflow. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/ICMLA.2015.120

Aditsania, A., Adiwijaya, & Saonard, A. L. (2017). Handling imbalanced data in churn prediction using ADASYN and backpropagation algorithm. In *Proceeding - 2017 3rd International Conference on Science in Information Technology* (Vol. 2018-Janua, pp. 533–536). https://doi.org/10.1109/ICSITech.2017.8257170

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist, 73*(7), 899–917. https://doi.org/10.1037/amp0000190

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). Predicting Customer Churn in Telecom Industry using Multilayer Preceptron Neural Networks: Modeling and Analysis Omar. *Life Science Journal, 11*(3), 75–81. https://doi.org/10.7537/marslsj110314.11

Aggarwal, C. C. (2015). *Data mining: the textbook*. New York, USA: Springer-Verlag.

Aghaei, M., Dimiccoli, M., Canton Ferrer, C., & Radeva, P. (2018). Towards social pattern characterization in egocentric photo-streams. *Computer Vision and Image Understanding, 171*, 104–117. https://doi.org/10.1016/J.CVIU.2018.05.001

Ahmed, A. A. Q., & Maheswari, D. (2016). Churn prediction on huge telecom data using hybrid firefly based classification. *Egyptian Informatics Journal, 18*(3), 2–7. https://doi.org/10.1016/j.eij.2017.02.002

Ahmed, I., Nawaz, M. M., Usman, A., Shautkat, M. Z., Ahmed, N., & Wasim-ul-Rehman. (2010). A mediation of customer satisfaction relationship between service quality and repurchase intentions for the telecom sector in pakistan: a case study of university students. African Journal of Business Management, 4(16), 3457–3462. Retrieved from https://academicjournals.org/journal/AJBM/article-abstract/B4E44C421081

Ahmed, M., Afzal, H., Majeed, A., & Khan, B. (2017). A Survey of Evolution in Predictive Models and Impacting Factors in Customer Churn. *Advances in Data Science and Adaptive Analysis, 09*(03), 1750007. https://doi.org/10.1142/s2424922x17500073

Ahmed, M., Afzal, H., Siddiqi, I., Amjad, M. F., & Khurshid, K. (2018). Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecom industry. *Neural Computing and Applications, 8*. https://doi.org/10.1007/s00521-018-3678-8

Ahmed, S. R. (2004). Applications of data mining in retail business. In *International Conference on Information Technology: Coding Computing, ITCC*. https://doi.org/10.1109/ITCC.2004.1286695

Aiken, L. S., West, S. G., & Reno, R. R. (1991). Multiple regression: Testing and interpreting interactions. Sage.

Akbar, M. M., & Parvez, N. (2009). Impact of Service Quality, Trust, and Customer Satisfaction on Customers Loyalty. ABAC Journal, 29(1), 24–38. Retrieved from http://www.assumptionjournal.au.edu/index.php/abacjournal/article/view/526

Ali, Ö. G., Akçay, Y., Sayman, S., Yılmaz, E., & Özçelik, M. H. (2017). Cross-Selling Investment Products with a Win-Win Perspective in Portfolio Optimization. *Operations Research, 65*(1), 55–74. https://doi.org/10.1287/opre.2016.1556

Alicke, M. D., Braun, J. C., Glor, J. E., Klotz, M. L., Magee, J., Sederhoim, H., & Siegel, R. (1992). Complaining Behavior in Social Interaction. *Personality and Social Psychology Bulletin, 18*(3), 286–295. https://doi.org/10.1177/0146167292183004

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing, 237*, 242–254. https://doi.org/10.1016/j.neucom.2016.12.009

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Retrieved July 11, 2019, from https://www.wired.com/2008/06/pb-theory/

Ansari, A., Mela, C. F., & Neslin, S. A. (2008). Customer Channel Migration. *Journal of Marketing Research, 45*(1), 60–76. https://doi.org/10.1509/jmkr.45.1.60

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics, 15*(2), 94–101. https://doi.org/10.1016/j.aci.2018.05.004

Apley, D. W., & Zhu, J. (2016). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. Retrieved from http://arxiv.org/abs/1612.08468

Artun, O., & Levin, D. (2015). Predictive marketing: Easy ways every marketer can use customer analytics and big data. John Wiley & Sons.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality, 27*(2), 108–119. https://doi.org/10.1002/per.1919

Athanassopoulos, A. D. (2000). Customer Satisfaction Cues to Support Market Segmentation and Explain Switching Behavior. *Journal of Business Research, 47*(3), 191–207. https://doi.org/http://dx.doi.org/10.1016/S0148-2963(98)00060-5

Awwad, S. M., & Neimat, A. B. (2010). Factors Affecting Switching Behavior of Mobile Service Users: The Case of Jordan. Journal of Economic and Administrative Sciences, 26(1), 27–51. https://doi.org/10.1108/10264116201000002

Aydin, S., Özer, G., & Arasil, Ö. (2005). Customer loyalty and the effect of switching costs as a moderator variable. A case in the Turkish mobile phone market. Marketing Intelligence and Planning, 23(1), 89–103. https://doi.org/10.1108/02634500510577492

Azeem, M., & Usman, M. (2018). A fuzzy based churn prediction and retention model for prepaid customers in telecom industry. *International Journal of Computational Intelligence Systems, 11*(1), 66. https://doi.org/10.2991/ijcis.11.1.6

Azeem, M., Usman, M., & Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. *Telecommunication Systems*, 1–12. https://doi.org/10.1007/s11235-017-0310-7

Babu, S., & Ananthanarayanan, N. R. (2016). Enhancing the performance of the classifiers for customer churn analysis in telecommunication data using EMOTE. *International Journal of Control Theory and Applications, 9*(34), 603–621. https://doi.org/10.1007/978-981-10-5520-1_43

Backiel, A., Baesens, B., & Claeskens, G. (2016). Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. *Journal of the Operational Research Society, 67*(9), 1135–1145. https://doi.org/10.1057/jors.2016.8

Backiel, A., Verbinnen, Y., Baesens, B., & Claeskens, G. (2015). Combining Local and Social Network Classifiers to Improve Churn Prediction. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 651–658). https://doi.org/10.1145/2808797.2808850

Baecke, P., & Van den Poel, D. (2011). Data augmentation by predicting spending pleasure using commercially available external data. *Journal of Intelligent Information Systems, 36*(3), 367–383. https://doi.org/10.1007/s10844-009-0111-x

Baecke, P., & Van den Poel, D. (2012). Including spatial interdependence in customer acquisition models: A cross-category comparison. *Expert Systems with Applications, 39*(15), 12105–12113. https://doi.org/10.1016/j.eswa.2012.04.008

Baecke, P., & Van den Poel, D. (2013). Improving customer acquisition models by incorporating spatial autocorrelation at different levels of granularity. *Journal of Intelligent Information Systems, 41*(1), 73–90. https://doi.org/10.1007/s10844-012-0225-4

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science, 7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications, 39*(18), 13517–13522. https://doi.org/10.1016/j.eswa.2012.07.006

Ballings, M., & Van den Poel, D. (2013). *Threshold independent performance measures for probabilistic classifiers*. Retrieved from https://cran.r-project.org/package=AUC

Ballings, M., Van den Poel, D., & Verhagen, E. (2012). Improving customer churn prediction by data augmentation using pictorial stimulus-choice data. In *Advances in Intelligent Systems and Computing* (pp. 217–226). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30864-2_21

Baras, D., Ronen, A., & Yom-Tov, E. (2014). The effect of social affinity and predictive horizon on churn prediction using diffusion modeling. *Social Network Analysis and Mining, 4*(1), 1–12. https://doi.org/10.1007/s13278-014-0232-2

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? Perspectives on *Psychological Science, 2*(4), 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Beaton, D., Dunlop, J., & Abdi, H. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. *Psychological Methods, 21*(4), 621–651. https://doi.org/10.1037/met0000053

Bellman, R. (1961). *Adaptive control processes: a guided tour*. Princeton, NJ, USA: Princeton University Press.

Bendapudi, N., & Berry, L. L. (1997). Customers' motivations for maintaining relationships with service providers. Journal of Retailing, 73(1), 15–37. https://doi.org/10.1016/S0022-4359(97)90013-0

Benoit, D. F., & Van den Poel, D. (2012). Improving customer retention in financial services using kinship network information. *Expert Systems with Applications, 39*(13), 11435–11442. https://doi.org/10.1016/j.eswa.2012.04.016

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Bertamini, M., & Munafò, M. R. (2012). Bite-Size Science and Its Undesired Side Effects. *Perspectives on Psychological Science, 7*(1), 67–71. https://doi.org/10.1177/1745691611429353

Biemer, P. (2010). Total Survey Error: Design, Implementation, and Evaluation. Public *Opinion Quarterly, 74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Biemer, P. (2014). Dropping the 's' from TSE: Applying the Paradigm to Big Data. In Paper presented at the 2014 *International Total Survey Error Workshop* (ITSEW 2014). Washington, DC: National Institute of Statistical Science. Retrieved from https://www.niss.org/sites/default/files/biemer_ITSEW2014_Presentation.pdf.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Blattberg, R. C., & Neslin, S. A. (1990). *Sales promotion: concepts, methods, and strategies*. Englewood Cliffs, N.J.: Prentice Hall. Retrieved from https://books.google.be/books?id=CjIPAQAAMAAJ

Blattberg, R. C., Getz, G., & Thomas, J. S. (2001). *Customer Equity: Building and Managing Relationships as Valuable Assets*. Harvard Business School Press. Retrieved from https://books.google.be/books?id=av9ysWUDsHAC

Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). Database Marketing: Analyzing and Managing Customers. *Springer Science & Business Media*, 607–633. https://doi.org/10.1007/978-0-387-72579-6

Bolancé, C., Guillen, M., & Padilla-Barreto, A. E. (2016). Predicting Probability of Customer Churn in Insurance. In R. León, M. J. Muñoz-Torres, & J. M. Moneva (Eds.), *Modeling and Simulation in Engineering, Economics and Management* (pp. 82–91). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40506-3_9

Bolton, R. N., Kannan, P. K., & Bramlett, M. D. (2000). Implications of Loyalty Program Membership and Service Experiences for Customer Retention and Value. Journal of the Academy of Marketing Science, 28(1), 95–108. https://doi.org/10.1177/0092070300281009

Bolton, R. N., Lemon, K. N., & Verhoef, P. C. (2004). The Theoretical Underpinnings of Customer Asset Management: A Framework and Propositions for Future Research. *Journal of the Academy of Marketing Science, 32*(3), 271–292. https://doi.org/10.1177/0092070304263341

Boroumand, L. (2006). Service failure and customer defection in online shops in Iran: customer-based view. Luleå University of Technology.

Bose, I., & Chen, X. (2009). Hybrid models using unsupervised clustering for prediction of customer churn. *Journal of Organizational Computing and Electronic Commerce, 19*(2), 133–151. https://doi.org/10.1080/10919390902821291

Bosl, W., Tierney, A., Tager-Flusberg, H., & Nelson, C. (2011). EEG complexity as a biomarker for autism spectrum disorder risk. *BMC Medicine, 9*(1), 18. https://doi.org/10.1186/1741-7015-9-18

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data. Information, *Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Boyd, D., & Ellison, N. B. (2008). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Bramer, M. A. (2013). *Principles of data mining*. Springer.

Braun, M. T., & Kuljanin, G. (2015). Big data and the challenge of construct validity. Industrial and Organizational Psychology, 8(4), 521–527. https://doi.org/10.1017/iop.2015.77

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123–140. https://doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall/CRC Texts in Statistical Science Series. Boca Raton, Florida (USA): Chapman & Hall/Crc. https://doi.org/10.1002/widm.8

Browne, M. W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology, 44*(1), 108–132. https://doi.org/10.1006/jmps.1999.1279

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*, 252–268. https://doi.org/10.1016/j.ejor.2003.12.010

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications, 32*, 277–288. https://doi.org/10.1016/j.eswa.2005.11.037

Burez, J., & Van den Poel, D. (2008). Separating financial from commercial customer churn : A modeling step towards resolving the conflict between the sales and credit department. *Expert Systems with Applications, 35*, 497–514. https://doi.org/10.1016/j.eswa.2007.07.036

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems With Applications, 36*, 4626–4636. https://doi.org/10.1016/j.eswa.2008.05.027

Burnham, T. A., Frels, J. K., & Mahajan, V. (2003). Consumer switching costs: A typology, antecedents, and consequences. Journal of the Academy of Marketing Science, 31(2), 109–126. https://doi.org/10.1177/0092070302250897

Butler, D. (2013). When Google got flu wrong. Nature, 494(7436), 155. https://doi.org/10.1038/494155a

Buttle, F. (2004). Customer Relationship Management. Concepts and Tools. Oxford, UK: Elsevier.

Buttle, F. A., & Ang, L. (2004). *Customer acquisition. An investigation of CRM competency*. https://doi.org/10.13140/RG.2.2.11410.96969

Buttle, F., & Maklan, S. (2019). Customer Relationship Management: Concepts and Technologies. Routledge.

Cao, K., & Shao, P. (2008). Customer Churn Prediction Based on SVM-RFE. In *International Seminar on Business and Information Management Customer* (pp. 306–309). https://doi.org/10.1109/ISBIM.2008.174

Carrier, C. G., & Povel, O. (2003). Characterising data mining software. Intelligent Data Analysis. Caruana, A. (2003). The impact of switching costs on customer loyalty: A study among corporate customers of mobile telephony. Journal of Targeting, Measurement and Analysis for Marketing, 12(3), 256–268. https://doi.org/10.1057/palgrave.jt.5740113

Cerda, P., & Varoquaux, G. (2019). Encoding high-cardinality string categorical variables. *ArXiv*. Retrieved from http://arxiv.org/abs/1907.01860

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The Promise and Perils of Wearable Sensors in Organizational Research. *Organizational Research Methods, 20*(1), 3–31. https://doi.org/10.1177/1094428115617004

Chakravarty, S., Feinberg, R., & Rhee, E. Y. (2004). Relationships and individuals' bank switching behavior. Journal of Economic Psychology, 25(4), 507–527. https://doi.org/10.1016/S0167-4870(03)00051-5

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112–130. https://doi.org/10.3758/s13428-013-0365-7

Charry, K., Coussement, K., Demoulin, N., & Heuvinck, N. (2016). *Marketing Research with IBM® SPSS Statistics. A Practical Guide* (2nd ed.). London, UK: Routledge.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods, 21*(4), 458–474. https://doi.org/10.1037/met0000111

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications, 19*(2). https://doi.org/10.1007/s11036-013-0489-0

Chen, P. P.-S., & Pin-Shan, P. (1976). The entity-relationship model-toward a unified view of data. *ACM Transactions on Database Systems, 1*(1), 9–36. https://doi.org/10.1145/320434.320440

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (pp. 785–794). New York, New York, USA: ACM Press. https://doi.org/10.1145/2939672.2939785

Chen, Y., Gel, Y. R., Lyubchich, V., & Winship, T. (2018). Deep Ensemble Classifiers and Peer Effects Analysis for Churn Forecasting in Retail Banking. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 373–385). Springer International Publishing. https://doi.org/10.1007/b97861

Cheung, M. W.-L., & Jak, S. (2016). Analyzing Big Data in Psychology: A Split/Analyze/Meta-Analyze Approach. *Frontiers in Psychology, 7*, 738. https://doi.org/10.3389/fpsyg.2016.00738

Chon, J., & Cha, H. (2011). LifeMap: A smartphone-based context provider for location-based services. IEEE Pervasive Computing, 10(2), 58–67. https://doi.org/10.1109/MPRV.2011.13

Chou, Y. C., & Chuang, H. H. C. (2018). A predictive investigation of first-time customer retention in online reservation services. *Service Business, 12*(4), 685–699. https://doi.org/10.1007/s11628-018-0371-z

Chow, P. I., Fua, K., Huang, Y., Bonelli, W., Xiong, H., Barnes, L. E., & Teachman, B. A. (2017). Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students. Journal of Medical Internet Research, 19(3), e62. https://doi.org/10.2196/jmir.6820

Chowdhury, T. G., Ratneshwar, S., & Mohanty, P. (2009). The time-harried shopper: Exploring the differences between maximizers and satisficers. *Marketing Letters, 20*(2), 155–167. https://doi.org/10.1007/s11002-008-9063-0

Christley, R. M. (2010). Power and Error: Increased Risk of False Positive Results in Underpowered Studies. *The Open Epidemiology Journal, 3*(1), 16–19. https://doi.org/10.2174/1874297101003010016

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Colgate, M., Stewart, K., & Kinsella, R. (1996). Customer defection: a study of the student market in Ireland. *International Journal of Bank Marketing, 14*(3), 23–29. https://doi.org/10.1108/02652329610113144

Cooley, T. (1879). A Treatise on the Law of Torts or the Wrongs Which Arise Independent of Contract. Retrieved from https://repository.law.umich.edu/books/11

Coppock, A. (2019). Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods, 7*(3), 613–628. https://doi.org/10.1017/psrm.2018.10

Cortes, C., & Vapnik, V. (1995). *Support-Vector Networks*. Machine Learning. https://doi.org/10.1023/A:1022627411411

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313–327. https://doi.org/10.1016/j.eswa.2006.09.038

Coussement, K., & Van den Poel, D. (2009). Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers. *Expert Systems with Applications, 36*, 6127–6134. https://doi.org/10.1016/j.eswa.2008.07.021

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems, 95*, 27–36. https://doi.org/10.1016/j.dss.2016.11.007

Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. In *Proceedings. Visualization '97* (pp. 235-244,). IEEE. https://doi.org/10.1109/VISUAL.1997.663888

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating "big data" and leveraging the potential of the geoweb. Cartography and Geographic Information Science, 40(2), 130–139. https://doi.org/10.1080/15230406.2013.777137

Crawford, K. (2013). The Hidden Biases in Big Data. Retrieved January 8, 2020, from https://hbr.org/2013/04/the-hidden-biases-in-big-data

Crouse, J. J., Moustafa, A. A., Bogaty, S. E. R., Hickie, I. B., & Hermens, D. F. (2018). Parcellating cognitive heterogeneity in early psychosis-spectrum illnesses: A cluster analysis. *Schizophrenia Research, 202*, 91–98. https://doi.org/10.1016/j.schres.2018.06.060

Crutzen, R., & Peters, G.-J. Y. (2017). Targeting Next Generations to Change the Common Practice of Underpowered Research. Frontiers in Psychology, 8, 1184. https://doi.org/10.3389/fpsyg.2017.01184

D'Haen, J., Van den Poel, D., Thorleuchter, D., & Benoit, D. F. (2016). Integrating expert knowledge and multilingual web crawling data in a lead qualification system. *Decision Support Systems, 82*, 69–78. https://doi.org/10.1016/j.dss.2015.12.002

D'Haen, Jeroen, & Van den Poel, D. (2013). Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework. *Industrial Marketing Management, 42*(4), 544–551. https://doi.org/10.1016/j.indmarman.2013.03.006

D'Haen, Jeroen, Van den Poel, D., & Thorleuchter, D. (2013). Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. *Expert Systems with Applications, 40*(6), 2007–2012. https://doi.org/10.1016/j.eswa.2012.10.023

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)* (pp. 1–4). https://doi.org/10.1109/CDAN.2016.7570883

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web - WWW '13* (pp. 307–318). New York, New York, USA: ACM Press. https://doi.org/10.1145/2488388.2488416

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., & Joshi, A. (2008). Social Ties and their Relevance to Churn in Mobile Telecom Networks. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology* (pp. 1–10). Nantes, France. https://doi.org/10.1145/1353343.1353424

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*(2), 760-772. https://doi.org/10.1016/j.ejor.2018.02.009

Degeratu, A. M., Rangaswamy, A., & Wu, J. (2000). Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes. International *Journal of Research in Marketing, 17*(1), 55–78. https://doi.org/10.1016/S0167-8116(00)00005-7

Deming, W. E. (1944). On errors in surveys. *American Sociological Review, 9*, 359–369. https://doi.org/10.2307/2085979

Deng, L., & Yu, D. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing, 7*(3–4), 197–387. https://doi.org/10.1561/2000000039

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks, 18*(4), 407–422. https://doi.org/10.1016/j.neunet.2005.03.007

Diebold, F. X. (2003). Big data dynamic factor models for macroeconomic measurement and forecasting. In M. Dewatripont, L. P. Hansen, & S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Societ*y (pp. 115–122). Cambridge University Press.

Dierkes, T., Bichler, M., & Krishnan, R. (2011). Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks. *Decision Support Systems, 51*(3), 361–371. https://doi.org/10.1016/j.dss.2011.01.002

Dietterich, T. G. (2000). *Ensemble Methods in Machine Learning* (pp. 1–15). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Domingos, P., & Richardson, M. (2001). Mining the network value of customers. In *International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/502512.502525

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

Donner, A. (1982). The exclusion of patients from a clinical trial. *Statistics in Medicine, 1*(3), 261–265. https://doi.org/10.1002/sim.4780010307

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. Retrieved from http://arxiv.org/abs/1702.08608

Droftina, U., Štular, M., & Košir, A. (2015). Predicting Influential Mobile-Subscriber Churners using Low-level User Features. *Automatika, 56*(4), 522–534. https://doi.org/10.1080/00051144.2015.11828665

Dunteman, G. H. (George H. (1989). *Principal components analysis*. Sage Publications.

Dwyer, F. R., Schurr, P. H., & Oh, S. (1987). Developing Buyer-Seller Relationships. *Journal of Marketing, 51*(2), 11. https://doi.org/10.2307/1251126

Eagle, N., & Pentland, A. S. (2009). Eigenbehaviors: Identifying structure in routine. Behavioral Ecology and Sociobiology, 63(7), 1057–1066. https://doi.org/10.1007/s00265-009-0739-0

Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications, 10*(1), 2319. https://doi.org/10.1038/s41467-019-10301-1

Ek, K., & Söderholm, P. (2008). Households' switching behavior between electricity suppliers in Sweden. Utilities Policy, 16(4), 254–261. https://doi.org/10.1016/j.jup.2008.04.005

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*(7), 1097–1126. https://doi.org/10.1037/0022-3514.37.7.1097

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research, 69*(2), 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001

Evenson, K. R., Goto, M. M., & Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *The International Journal of Behavioral Nutrition and Physical Activity, 12*, 159. https://doi.org/10.1186/s12966-015-0314-1

Fathian, M., Hoseinpoor, Y., & Minaei-Bidgoli, B. (2016). Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods. *Kybernetes, 45*(5), 732–743. https://doi.org/10.1108/K-07-2015-0172

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37–37. https://doi.org/10.1609/AIMAG.V17I3.1230

Fishbein, M., & Ajzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological Review, 81*(1), 59–74. https://doi.org/10.1037/h0035872

Floridi, L. (2012, December). Big data and their epistemological challenge. Philosophy and Technology. https://doi.org/10.1007/s13347-012-0093-4

Fornell, C., & Wernerfelt, B. (1987). Defensive Marketing Strategy by Customer Complaint Management: A Theoretical Analysis. *Journal of Marketing Research, 24*(4), 337–346. https://doi.org/10.2307/3151381

Fraley, R. C., & Marks, M. J. (2007). The null hypothesis significance-testing debate and its implications for personality research. In *Handbook of research methods in personality psychology*. (pp. 149–169). New York, NY, US: The Guilford Press.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine, 13*(3), 57–57. https://doi.org/10.1609/AIMAG.V13I3.1011

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences, 55*(1), 119–139. https://doi.org/10.1006/JCSS.1997.1504

Frické, M. (2015). Big data and its epistemology. Journal of the Association for Information Science and Technology, 66(4), 651–661. https://doi.org/10.1002/asi.23212

Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics*. Institute of Mathematical Statistics. https://doi.org/10.2307/2699986

Gajowniczek, K., Orłowski, A., & Zabkowski, T. (2016). Entropy based trees to support decision making for customer churn management. In *Proceedings of the 8th Polish Symposium of Physics in Economy and Social Sciences* (Vol. 129, pp. 971–979). https://doi.org/10.12693/APhysPolA.129.971

Gamble, A., Juliusson, E. A., & Gärling, T. (2009). Consumer attitudes towards switching supplier in three deregulated markets. Journal of Socio-Economics, 38(5), 814–819. https://doi.org/10.1016/j.socec.2009.05.002

Gamboa, J. C. B. (2017). Deep Learning for Time-Series Analysis. *ArXiv*. Retrieved from http://arxiv.org/abs/1701.01887

Gamulin, N., Štular, M., & Tomažič, S. (2015). Impact of Social Network to Churn in Mobile Network. *Automatika, 56*(3), 252–261. https://doi.org/10.7305/automatika.2015.12.742

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers. *Journal of Marketing, 64*(3), 65–87. https://doi.org/10.1509/jmkg.64.3.65.18028

General Data Protection Regulation. (2018). General Data Protection Regulation (GDPR) – Official Legal Text. Retrieved August 20, 2019, from https://gdpr-info.eu/

Gerpott, T. J., Rams, W., & Schindler, A. (2001). Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommunications Policy, 25(4), 249–269. https://doi.org/http://dx.doi.org/10.1016/S0308-5961(00)00097-5

Gerrard, P., & Cunningham, J. B. (2004). Consumer switching behavior in the Asian banking market. Journal of Services Marketing, 18(3), 215–223. https://doi.org/10.1108/08876040410536512

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can Psychological Traits Be Inferred From Spending? Evidence From Transaction Data. *Psychological Science, 30*(7), 1087–1096. https://doi.org/10.1177/0956797619849435

Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management, 62*, 100–107. https://doi.org/10.1016/j.indmarman.2016.08.003

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. Annual Review of Psychology, 66(1), 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology, 74*(5), 1337–1349. https://doi.org/10.1037//0022-3514.74.5.1337

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences, 33*(2–3), 94–95. https://doi.org/10.1017/S0140525X10000300

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. Annual *Review of Psychology, 60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Griffin, J., & Lowenstein, M. (2002). *Customer Winback: How to Recapture Lost Customers and Keep Them Loyal*. San Francisco, CA, USA: John Wiley & Sons.

Grolemund, G. (2014). Hands-on programming with R. O'Reilly Media.

Groves, W. (2013). Using Domain Knowledge to Systematically Guide Feature Selection. *Twenty-Third International Joint Conference on Artificial Intelligence*. Retrieved from https://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/viewPaper/6999

Gruszczyński, W., & Arabas, P. (2011). Application of social network to improve effectiveness of classifiers in churn modelling. In *Proceedings of the 2011 International Conference on Computational Aspects of Social Networks, CASoN'11* (pp. 217–222). https://doi.org/10.1109/CASON.2011.6085947

Günther, C. C., Tvete, I. F., Aas, K., Sandnes, G. I., & Borgan, Ø. (2014). Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, (1), 58–71. https://doi.org/10.1080/03461238.2011.636502

Guo, F., & Qin, H. L. (2015). The analysis of customer churns in e-commerce based on decision tree. In *2015 International Conference on Computer Science and Applications, CSA 2015* (pp. 199–203). IEEE. https://doi.org/10.1109/CSA.2015.74

Gupta, S. (2014). *Marketing Reading: Customer Management*. Retrieved October 17, 2019, from https://hbsp.harvard.edu/product/8162-PDF-ENG?itemFindingMethod=Collections

Guyon, I., Bitter, H.-M., Ahmed, Z., Brown, M., & Heller, J. (2005). Multivariate Non-Linear Feature Selection with Kernel Methods. In *Soft Computing for Information Processing and Analysis* (pp. 313–326). Berlin/Heidelberg: Springer-Verlag. https://doi.org/10.1007/3-540-32365-1_12

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques. Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. *Journal of the Royal Statistical Society, 160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Hao, T., Xing, G., & Zhou, G. (2013). ISleep: Unobtrusive sleep quality monitoring using smartphones. In SenSys 2013 - Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems. Association for Computing Machinery. https://doi.org/10.1145/2517351.2517359

Harari, G., Gosling, S. D., Wang, R., & Campbell, A. (2015). Capturing Situational Information with Smartphones and Mobile Sensing Methods. *European Journal of Personality, 29*, 509–511. https://doi.org/10.1002/per.2032

Hashmi, O. Z., & Sheikh, S. (2012). Impact of social attributes on predictive analytics in telecommunication industry. In *2012 15th International Multitopic Conference, INMIC 2012* (pp. 47–52). https://doi.org/10.1109/INMIC.2012.6511470

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics (2nd ed.). New York, USA: Springer-Verlag. https://doi.org/10.1007/978-0-387-84858-7

Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G. F., & Clermont, G. (2013). Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics, 46*(1), 47–55. https://doi.org/10.1016/J.JBI.2012.08.004

Hawkins, D. M. (1980). *Identification of outliers*. Chapman and Hall.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Heckman, J. J. (1990). Varieties of Selection Bias. *The American Economic Review, 80*(2), 313–318. https://doi.org/10.1126/science.151.3712.867-a

Hemminki, S., Nurmi, P., & Tarkoma, S. (2013). Accelerometer-based transportation mode detection on smartphones. In SenSys 2013 - Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems. Association for Computing Machinery. https://doi.org/10.1145/2517351.2517367

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature, 466*(7302). https://doi.org/10.1038/466029a

Hernandez, I., Newman, D. A., & Jeon, G. (2016). Twitter analysis: Methods for data management and a word count dictionary to measure city-level job satisfaction. In S. Tonidandel, E. B. King, & J. M. Cortina (Eds.), *Big Data at Work: The Data Science Revolution and Organizational Psychology*. New York, USA: Routledge. https://doi.org/10.4324/9781315780504

Hersh, E. D. (2013). Long-term effect of September 11 on the political behavior of victims' families and neighbors. *Proceedings of the National Academy of Sciences of the United States of America, 110*(52), 20959–20963. https://doi.org/10.1073/pnas.1315043110

Hey, T., Tansley, S., & Tolle, K. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.

Hill, S., Provost, F., & Volinsky, C. (2006). Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science, 21*(2), 256-276. https://doi.org/10.1214/088342306000000222

Hitt, L. M., & Frei, F. X. (2002). Do Better Customers Utilize Electronic Distribution Channels? The Case of PC Banking. *Management Science, 48*(6), 732–748. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=6991713&site=ehost-live&scope=site

Hollon, S. D., Cohen, Z. D., Singla, D. R., & Andrews, P. W. (2019). Recent Developments in the Treatment of Depression. *Behavior Therapy, 50*(2), 257–269. https://doi.org/10.1016/J.BETH.2019.01.002

Holtrop, N., Wieringa, J. E., Gijsenberg, M. J., & Verhoef, P. C. (2017). No future without the past? Predicting churn in the face of customer privacy. *International Journal of Research in Marketing, 34*(1), 154–172. https://doi.org/10.1016/j.ijresmar.2016.06.001

Horowitz, M. (2015). *Detailed NFL Play-by-Play Data 2015*. Retrieved October 23, 2019, from https://www.kaggle.com/maxhorowitz/nflplaybyplay2015

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. (3rd, Ed.), Wiley. Hoboken, New Jersey , USA: Wiley. https://doi.org/10.1002/9781118548387

Huang, B. Q., Kechadi, M., & Buckley, B. (2009). Customer Churn Prediction for Broadband Internet Services. Data Warehousing and Knowledge Discovery. *11th International Conference, DaWaK 2009*, 229–243.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications, 39*(1), 1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

Hwang, J.-N., Lay, S.-R., & Lippman, A. (1994). Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing, 42*(10), 2795–2810. https://doi.org/10.1109/78.324744

Hwong, Y.-L., Oliver, C., Van Kranendonk, M., Sammut, C., & Seroussi, Y. (2017). What makes you tick? The psychology of social media engagement in space science communication. *Computers in Human Behavior, 68*, 480–492. https://doi.org/10.1016/j.chb.2016.11.068

IBM Corporation. (2016). IBM SPSS Statistics for Machintosh. Armonk, New York.

Iyengar, S. S., Wells, R. E., & Schwartz, B. (2006). Doing better but feeling worse looking for the "Best" job undermines satisfaction. *Psychological Science, 17*(2), 143–150. https://doi.org/10.1111/j.1467-9280.2006.01677.x

Jacoby, J., & Morrin, M. (2015). Consumer Psychology. In International Encyclopedia of the

Social & Behavioral Sciences: Second Edition (pp. 738–743). Elsevier Inc.

https://doi.org/10.1016/B978-0-08-097086-8.22004-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014*). An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., … Usher, A. (2015). Big Data in Survey Research. *Public Opinion Quarterly, 79*, 839–880. https://doi.org/10.1093/poq/nfv039

Jayaswal, P., Prasad, B. R., Tomar, D., & Agarwal, S. (2016). An Ensemble Approach for Efficient Churn Prediction in Telecom Industry. *International Journal of Database Theory and Application, 9*(8), 211–232. https://doi.org/http://dx.doi.org/10.14257/ijdta.2016.9.8.21

Jeon, J., Yoon, D., Yang, S., & Kim, K. (2017). Extracting gamers' cognitive psychological features and improving performance of churn prediction from mobile games. In *2017 IEEE Conference on Computational Intelligence and Games*, CIG 2017 (pp. 150–153). https://doi.org/10.1109/CIG.2017.8080428

Jiao, J., Zhang, Y., & Helander, M. (2006). A Kansei mining system for affective design. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2005.07.020

Jiawei, H., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers Inc. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Jones, M. A., Mothersbaugh, D. L., & Beatty, S. E. (2000). Switching barriers and repurchase intentions in services. Journal of Retailing, 76(2), 259–274. https://doi.org/10.1016/S0022-4359(00)00024-5

Kamakura, W. A., Ramaswami, S. N., & Srivastava, R. K. (1991). Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *International Journal of Research in Marketing, 8*, 329–349. https://doi.org/10.1016/0167-8116(91)90030-B

Kamakura, W. A., Wedel, M., de Rosa, F., & Mazzon, J. A. (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in Marketing*. https://doi.org/10.1016/S0167-8116(02)00121-0

Kamalakannan, T., & Mayilvaghnan, P. (2018). Optimal customer relationship management in telecalling industry by using data mining and business intelligence. *International Journal of Engineering & Technology, 7*(1.1), 12–17. https://doi.org/10.14419/ijet.v7i1.1.8907

Kawale, J., Pal, A., & Srivastava, J. (2009). Churn Prediction in MMORPGs: A Social Influence Based Approach. In *2009 International Conference on Computational Science and Engineering* (pp. 423–428). IEEE. https://doi.org/10.1109/CSE.2009.80

Keaveney, S. M. (1995). Customer Switching Behavior in Service Industries: An Exploratory Study. Journal of Marketing, 59(2), 71–82. Retrieved from http://www.jstor.org/stable/1252074

Keaveney, S. M., & Parthasarathy, M. (2001). Customer Switching Behavior in Online Services: An Exploratory Study of the Role of Selected Attitudinal, Behavioral, and Demographic Factors. *Journal of the Academy of Marketing Science, 29*(4), 374–390. https://doi.org/10.1177/03079450094225

Keaveney, Susan M. (1995). Customer Switching Behavior in Service Industries: An Exploratory Study. *Journal of Marketing, 59*(2), 71–82. Retrieved from http://www.jstor.org/stable/1252074

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive Science: A New Paradigm for Biodiversity Studies. BioScience, 59(7), 613–620. https://doi.org/10.1525/bio.2009.59.7.12

Kessler, R. C., Hwang, I., Hoffmire, C. A., McCarthy, J. F., Petukhova, M. V., Rosellini, A. J., … Bossarte, R. M. (2017). Developing a practical suicide risk prediction model for targeting high-risk patients in the Veterans health Administration. *International Journal of Methods in Psychiatric Research*, 26(3), e1575. https://doi.org/10.1002/mpr.1575

Khalid, M. A. R., Farquad, M. A. H., & Kamakshi Prasad, V. (2017). Data Classification using Active Learning based Data Modification: An Application to Churn Prediction. In *International Conference on Current Trends in Computer, Electrical, Electronics and Communication* (pp. 529–533). IEEE. https://doi.org/10.1109/CTCEEC.2017.8454989

Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 2288–2296). Retrieved from https://dl.acm.org/citation.cfm?id=3157352

Kim, Jae-on., & Mueller, C. W. (1978). *Factor analysis: statistical methods and practical issues*. SAGE Publications.

Kim, Jinhwa, Won, C., & Bae, J. K. (2010). A knowledge integration model for the prediction of corporate dividends. *Expert Systems with Applications, 37*(2), 1344–1350. https://doi.org/10.1016/j.eswa.2009.06.035

Kim, K., Jun, C. H., & Lee, J. (2014). Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Systems with Applications, 41*(15), 6575–6584. https://doi.org/10.1016/j.eswa.2014.05.014

Kim, M. K., Park, M. C., & Jeong, D. H. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. Telecommunications Policy, 28(2), 145–159. https://doi.org/10.1016/j.telpol.2003.12.003

Kim, S., Choi, D., Lee, E., & Rhee, W. (2017). Churn prediction of mobile and online casual games using play log data. *PLoS ONE, 12*(7), 1–19. https://doi.org/10.1371/journal.pone.0180735

Kirui, C., Hong, L., Cheruiyot, W., & Kirui, H. (2013). Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining. *International Journal of Computer Science Issues, 10*(2), 165–172. https://doi.org/10.1108/IMDS-12-2015-0509

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data and Society, 1. https://doi.org/10.1177/2053951714528481

Klemperer, P. (1987). Markets with Consumer Switching Costs*. *The Quarterly Journal of Economics, 102*(2), 375–394. https://doi.org/10.2307/1885068

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text Mining in Organizational Research. *Organizational Research Methods, 21*(3), 733–765. https://doi.org/10.1177/1094428117722619

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conference on Artificial Intelligence*. Retrieved from https://www.researchgate.net/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection

Kohavi, R., & Provost, F. (1998). Glossary of Terms. *Machine Learning, 30*, 271–274. https://doi.org/10.1023/A:1017181826899

Kohavi, Ron, & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods, 21*(4), 493–506. https://doi.org/10.1037/met0000105

Kotler, P. T., Bowen, J. T., & Makens, J. (2014). *Marketing for Hospitality and Tourism* (6th ed.). Pearson.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review, 39*(4), 261-283. https://doi.org/10.1007/s10462-011-9272-4

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets : A review. *Science, 30*(1), 25–36. https://doi.org/10.1007/978-0-387-09823-4_45

Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-Validation Approaches for Replicability in Psychology. *Frontiers in Psychology, 9*, 1117. https://doi.org/10.3389/fpsyg.2018.01117

Kowalski, R. M. (1996). Complaints and complaining: Functions, antecedents, and consequences. *Psychological Bulletin, 119*(2), 179–196. https://doi.org/10.1037/0033-2909.119.2.179

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kreuter, F., & Peng, R. (2013). Extracting information from big data: Issues of measurement, inference and linkage. In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 257–275. https://doi.org/10.1017/CBO9781107590205.016

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1–27. https://doi.org/10.1007/BF02289565

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE, 9*(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York, USA: Springer Nature.

Kuhn, M., & Johnson, K. (2019). Feature Engineering and Selection : a Practical Approach for Predictive Models. CRC Press LLC.

Kuhn, M., Weston, S., & Coulter, N. (2015). *C50: C5.0 Decision Trees and Rule-Based Models*. Retrieved from https://cran.r-project.org/package=C50

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., … Hunt, T. (2017). *caret: Classification and Regression Training*. Retrieved from https://cran.r-project.org/package=caret

Kumar, V., & Reinartz, W. J. (2012). Customer relationship management : concept, strategy, and tools. Springer.

Kumar, V., Bhagwat, Y., & Zhang, X. (2015). Regaining "lost" customers: The predictive power of first-lifetime behavior, the reason for defection, and the nature of the win-back offer. *Journal of Marketing, 79*(4), 34-55. https://doi.org/10.1509/jm.14.0107

Kumar, V., Leszkiewicz, A., & Herbst, A. (2018). Are you back for good or still shopping around? Investigating customers' repeat churn behavior. *Journal of marketing research, 55*(2), 208-225. https://doi.org/10.1509/jmr.16.0623

Kuncheva, L. I. (2004). *Combining Pattern Classifiers*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471660264

Lai, L. (2011). Maximizing and customer loyalty: Are maximizers less loyal? *Judgment and Decision Making, 6*(4), 307–313.

Lam, S. Y., Shankar, V., Erramilli, M. K., & Murthy, B. (2004, June). Customer value, satisfaction, loyalty, and switching costs: An illustration from a business-to-business service context. Journal of the Academy of Marketing Science. https://doi.org/10.1177/0092070304263330

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies. https://doi.org/10.1016/j.infsof.2008.09.005

Lathia, N., Sandstrom, G. M., Mascolo, C., & Rentfrow, P. J. (2017). Happier People Live More Active Lives: Using Smartphones to Link Happiness and Physical Activity. PLOS ONE, 12(1), e0160589. https://doi.org/10.1371/journal.pone.0160589

Lave, J. (1988). *Cognition in Practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511609268

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science, 343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., … Van Alstyne, M. (2009). Social science. Computational social science. *Science, 323*(5915), 721–723. https://doi.org/10.1126/science.1167742

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, E.-B., Kim, J., & Lee, S.-G. (2017). Predicting customer churn in mobile industry using data mining technology. *Industrial Management & Data Systems, 117*(1), 90–109. https://doi.org/10.1108/IMDS-12-2015-0509

Lee, K. C., & Jo, N. Y. (2010). Bayesian network approach to predict mobile churn motivations: Emphasis on general Bayesian network, Markov Blanket, and what-if simulation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6485 LNCS*, 304–313. https://doi.org/10.1007/978-3-642-17569-5_30

Lee, R. Y. M., & Murphy, J. (2005). From Loyalty to Switching: Exploring the Determinants in the Transition (pp. 196–203). The University of Western Australia.

Leonelli, S. (2012). Introduction: Making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*. https://doi.org/10.1016/j.shpsc.2011.10.001

Leung, K., Li, W. K., & Au, Y. F. (1998). The Impact of Customer Service and Product Value on Customer Loyalty and Purchase Behavior. Journal of Applied Social Psychology, 28(18), 1731–1741. https://doi.org/10.1111/j.1559-1816.1998.tb01343.x

Levin, N., & Zahavi, J. (1998). Continuous predictive modeling—a comparative analysis. *Journal of Interactive Marketing, 12*(2), 5-22. https://doi.org/10.1002/(SICI)1520-6653(199821)12:2<5::AID-DIR2>3.0.CO;2-D

Li, R., Wang, P., & Chen, Z. (2016). A Feature Extraction Method Based on Stacked Auto-Encoder for Telecom Churn Prediction. In *Asian Simulation Conference* (Vol. 643, pp. 568–576). https://doi.org/10.1007/978-981-10-2663-8

Li, S., Sun, B., & Wilcox, R. T. (2005). Cross-selling sequentially ordered products: An application to consumer banking services. *Journal of Marketing Research, 42*(2), 233-239. https://doi.org/10.1509/jmkr.42.2.233.62288

Li, Y. (2019). Innovation Research on Psychological Health Education of Contemporary College Students under the Background of Big Data. In *2018 International Workshop On Advances In Social Sciences* (Iwass 2018). Retrieved from https://webofproceedings.org/proceedings_series/article/artId/4545.html

Liao, H. Y., Chen, K. Y., Liu, D. R., & Chiu, Y. L. (2015). Customer Churn Prediction in Virtual Worlds. In *IIAI 4th International Congress on Advanced Applied Informatics* (pp. 115–120). https://doi.org/10.1109/IIAI-AAI.2015.265

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18–22. Retrieved from https://cran.r-project.org/doc/Rnews/

Lima, E., Mues, C., & Baesens, B. (2009). Domain Knowledge Integration in Data Mining Using Decision Tables: Case Studies in Churn Prediction. *The Journal of the Operational Research Society Journal of the Operational Research Society, 60*(8), 1096–1106. https://doi.org/10.1057/jors.2008.161

Ling, R., & Yen, D. C. (2001). Customer relashioship management: an analysis framework and implementation strategies. *Journal of Computer Information Systems*. https://doi.org/10.1080/08874417.2001.11647013

Linoff, G., & Berry, M. J. A. (2011). Data mining techniques : for marketing, sales, and customer relationship management (3rd ed.). Wiley.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York., 381. https://doi.org/10.1002/9781119013563

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Second edition. https://doi.org/10.2307/1533221

Liu, M., Qiao, X. Q., & Xu, W. L. (2011). Three categories customer churn prediction based on the adjusted real adaboost. *Communications in Statistics: Simulation and Computation, 40*(10), 1548–1562. https://doi.org/10.1080/03610918.2011.589732

Lix, T. S., Berger, P. D., & Magliozzi, T. L. (1995). New customer acquisition: prospecting models and the use of commercially available external data. *Journal of Direct Marketing, 9*(4), 8–18. https://doi.org/10.1002/dir.4000090403

Lohr, S. (2014). For Data Scientists, “Janitor Work” is Hurdle to Insights. The New York Times.

Long, X., Yin, W., An, L., Haiying, N., Huang, L., Luo, Q., & Yan, C. (2012). Churn Analysis of Online Social Network Users Using Data Mining Techniques. In *Proceeding of international multi conference of engineers and computer scientists* (Vol. I, pp. 14–16).

Lopes, L., Brito, C., & Alves, H. (2013). Customer relationship reactivation in the telecommunications sector. In *3rd International Network of Business & Management Conference*. Lisboa, Portugal.

Retrieved from https://bibliotecadigital.ipb.pt/bitstream/10198/9563/1/Full paper SIJ_INBAM2013_final.pdf

Lu, H., Pan, W., Lane, N. D., Choudhury, T., & Campbell, A. T. (2009). SoundSense: Scalable sound sensing for people-centric applications on mobile phones. In MobiSys'09 - Proceedings of the 7th ACM International Conference on Mobile Systems, Applications, and Services (pp. 165–178). https://doi.org/10.1145/1555816.1555834

Mahajan, V., Misra, R., & Mahajan, R. (2015). Review of data mining techniques for churn prediction in telecom. *Journal of Information and Organizational Sciences, 39*(2), 183–197. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84954065213&partnerID=40&md5=f907e68d2cf9b7ff0e558e5f6f8fd366

Maimon, O., & Rokach, L. (2010). *Data mining and knowledge discovery handbook*. Springer.

Malhotra, A., & Malhotra, C. K. (2013). Exploring switching behavior of US mobile service customers. Journal of Services Marketing, 27(1), 13–24. https://doi.org/10.1108/08876041311296347

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Markou, M., & Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal Processing, 83*(12), 2481–2497. https://doi.org/10.1016/j.sigpro.2003.07.018

Marr, B. (2019). *Big Data in Practice*. Retrieved August 20, 2019, from https://www.bernardmarr.com/default.asp?contentID=1076

Martensen, A., & Grønholdt, L. (2010). Measuring and managing brand equity. International Journal of Quality and Service Sciences, 2(3), 300–316. https://doi.org/10.1108/17566691011090044

Martensen, A., Gronholdt, L., & Kristensen, K. (2000). The drivers of customer satisfaction and loyalty: Cross-industry findings from Denmark. Total Quality Management, 11(4–6), 544–553. https://doi.org/10.1080/09544120050007878

Mas, A., & Moretti, E. (2009). Peers at Work. *American Economic Review, 99*(1), 112–145. https://doi.org/10.1257/aer.99.1.112

Mashey, R. J. (1998). *Big Data and the Next Big Wave of InfraStress*. Retrieved from http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

Matz, S. C., & Netzer, O. (2017). Using Big Data as a window into consumers' psychology. *Current Opinion in Behavioral Sciences, 18*, 7–12. https://doi.org/10.1016/J.COBEHA.2017.05.009

Mau, S., Pletikosa, I., & Wagner, J. (2018). Forecasting the next likely purchase events of insurance customers: A case study on the value of data-rich multichannel environments. *International Journal of Bank Marketing, 36*(6), 1125–1144. https://doi.org/10.1108/IJBM-11-2016-0180

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods, 9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

McDaniel, T. M., & Groothuis, P. A. (2012). Retail competition in electricity supply-Survey results in North Carolina. Energy Policy, 48, 315–321. https://doi.org/10.1016/j.enpol.2012.05.028

McFall, J. (1969). Priority Patterns and Consumer Behavior. *Journal of Marketing, 33*(4), 50. https://doi.org/10.2307/1248673

McFarland, D. A., & McFarland, H. R. (2015). Big Data and the danger of being precisely inaccurate. *Big Data & Society, 2*(2), 2053951715602495. https://doi.org/10.1177/2053951715602495

McKinney, W. (2017). Python for data analysis : data wrangling with pandas, NumPy, and IPython. Sebastopol, CA, USA: Oreilly & Associates Inc.

Meire, M., Ballings, M., & Van den Poel, D. (2017). The added value of social media data in B2B customer acquisition systems: A real-life experiment. *Decision Support Systems, 104*, 26–37. https://doi.org/10.1016/j.dss.2017.09.010

Methlie, L. B., & Nysveen, H. (1999). Loyalty of on-line bank customers. Journal of Information Technology, 14(4), 375–386. https://doi.org/10.1080/026839699344485

Meyer-Waarden, L. (2007). The effects of loyalty programs on customer lifetime duration and share of wallet. *Journal of Retailing, 83*(2), 223–236. https://doi.org/https://doi.org/10.1016/j.jretai.2007.01.002

Michel, S. (2001). Analyzing service failures and recoveries: A process approach. International Journal of Service Industry Management, 12(1), 20–33. https://doi.org/10.1108/09564230110382754

Microsoft Corporation, & Weston, S. (2015). *foreach: Foreach looping construct for R*. Retrieved from https://cran.r-project.org/package=foreach

Miguéis, V. L., Van den Poel, D., Camanho, A. S., & Falcão e Cunha, J. (2012). Modeling partial customer churn : On the value of first product-category purchase sequences. *Expert System with Applications, 39*, 11250–11256. https://doi.org/10.1016/j.eswa.2012.03.073

Mikelsons, G., Smith, M., Mehrotra, A., & Musolesi, M. (2017). Towards Deep Learning Models for Psychological State Prediction using Smartphone Data: Challenges and Opportunities. *ArXiv*. Retrieved from http://arxiv.org/abs/1711.06350

Miller, H. J. (2010). The Data Avalanche Is Here. Shouldn't We Be Digging? Journal of Regional Science, 50(1), 181–201. https://doi.org/10.1111/j.1467-9787.2009.00641.x

Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Retrieved from http://arxiv.org/abs/1706.07269

Milošević, M., Živić, N., & Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications, 83*, 326–332. https://doi.org/10.1016/j.eswa.2017.04.056

Mishra, H. G., Sinha, P. K., & Koul, S. (2017). Customer dependence and customer loyalty in traditional and modern format stores. *Journal of Indian Business Research, 9*(1), 59–78. https://doi.org/10.1108/JIBR-12-2015-0126

Mitchell, T. R., & James, L. R. (2001). Building Better Theory: Time and The Specification of When Things Happen. *Academy of Management Review, 26*(4), 530–547. https://doi.org/10.5465/amr.2001.5393889

Mitrović, S., Baesens, B., Lemahieu, W., & De Weerdt, J. (2018). On the operational efficiency of different feature types for telco Churn prediction. *European Journal of Operational Research, 267*(3), 1141–1155. https://doi.org/10.1016/j.ejor.2017.12.015

Mitrović, S., Singh, G., Baesens, B., Lemahieu, W., & De Weerdt, J. (2017). Scalable RFM-Enriched representation learning for churn prediction. *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017*, 79–88. https://doi.org/10.1109/DSAA.2017.42

Modani, N., Dey, K., Gupta, R., & Godbole, S. (2013). CDR analysis based Telco churn prediction and customer behavior insights: A case study. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8181 LNCS(PART 2)*, 256–269. https://doi.org/10.1007/978-3-642-41154-0_19

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems, 72*, 72–81. https://doi.org/10.1016/j.dss.2015.02.007

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., & Kumar, S. (2011). Google Correlate Whitepaper.

Molnar, C. (2019). *Interpretable Machine. A guide for making black box models explainable*. Retrieved from https://christophm.github.io/interpretable-ml-book/

Mols, N. P. (1998). The Internet and the banks' strategic distribution channel decisions. *Internet Research, 8*(4), 331–337. https://doi.org/10.1108/10662249810231087

Moon, S., & Russell, G. J. (2008). Predicting Product Purchase from Inferred Customer Similarity: An Autologistic Model Approach. *Management Science, 54*(1), 71–82. https://doi.org/10.1287/mnsc.1070.0760

Morita, J. G., Lee, T. W., & Mowday, R. T. (1993). The regression-analog to survival analysis: A selected application to turnover research. *Academy of Management Journal, 36*(6), 1430–1464. https://doi.org/10.2307/256818

Moustafa, A. A., Diallo, T. M. O., Amoroso, N., Zaki, N., Hassan, M., & Alashwal, H. (2018). Applying Big Data Methods to Understanding Human Behavior and Health. *Frontiers in Computational Neuroscience, 12*, 84. https://doi.org/10.3389/fncom.2018.00084

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Churn Reduction in the Wireless Industry. *Advances in Neural Information Processing Systems, 12*, 935–941.

Muchnik, L., Aral, S., & Taylor, S. (2013). Social Influence Bias: A Randomized Experiment. *Science, 341*, 647–651. https://doi.org/10.1126/science.1240466

Murnane, E. L., Abdullah, S., Matthews, M., Kay, M., Kientz, J. A., Choudhury, T., … Cosley, D. (2016). Mobile manifestations of alertness: Connecting biological rhythms with patterns of smartphone app use. In Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2016 (pp. 465–477). Association for Computing Machinery, Inc. https://doi.org/10.1145/2935334.2935383

Murphy, K. R., & Russell, C. J. (2017). Mend It or End It: Redirecting the Search for Interactions in the Organizational Sciences. *Organizational Research Methods, 20*(4), 549–573. https://doi.org/10.1177/1094428115625322

Nabareseh, S. (2014). Exploring Roles of Females in Contemporary Socio-Politico-Economic Governance: An Association Rule Approach. *Mediterranean Journal of Social Sciences*. https://doi.org/10.5901/mjss.2014.v5n23p2178

Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D., & Rentfrow, J. (2018). Musical Preferences Predict Personality: Evidence From Active Listening and Facebook Likes. *Psychological Science, 29*(7), 1145–1158. https://doi.org/10.1177/0956797618761659

Neslin, Scott A., Gupta, S., Kamakura, W., Lu, J. X., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204–211. https://doi.org/10.1509/jmkr.43.2.204

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. Marketing Science, 31(3), 521–543. https://doi.org/10.1287/mksc.1120.0713

Nevin, J. R. (1995). Relationship Marketing and Distribution Channels: Exploring Fundamental Issues. *Journal of the Academy of Marketing Science, 23*(4), 327–334. https://doi.org/10.1177/009207039502300413

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management : A literature review and classification. *Expert Systems With Applications, 36*(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

Ngonmang, B., Viennet, E., & Tchuente, M. (2012). Churn prediction in a real online social network using local community analysis. In *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012* (pp. 282–288). https://doi.org/10.1109/ASONAM.2012.55

Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications, 38*(12), 15273–15285. https://doi.org/10.1016/j.eswa.2011.06.028

Nitzan, I., & Libai, B. (2011). Social effects on customer retention. *Journal of Marketing, 75*(6), 24-38. https://doi.org/10.1509/jmkg.75.6.24

Olle, G. D., & Cai, S. (2014). A Hybrid Churn Prediction Model in Mobile Telecommunication Industry. *International Journal of E-Education, e-Business, e-Management and e-Learning, 4*(1), 55–62. https://doi.org/10.7763/ijeeee.2014.v4.302

Olson, D. L., & Chae, B. K. (2012). Direct marketing decision support through predictive customer response modeling. *Decision Support Systems, 54*(1), 443-451. https://doi.org/10.1016/j.dss.2012.06.005

Oommen, T., Baise, L. G., & Vogel, R. M. (2011). Sampling Bias and Class Imbalance in Maximum-likelihood Logistic Regression. *Mathematical Geosciences, 43*(1), 99–120. https://doi.org/10.1007/s11004-010-9311-8

Open Science Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Óskarsdóttir, M., Baesens, B., & Vanthienen, J. (2018). Profit-Based Model Selection for Customer Retention Using Individual Customer Lifetime Values. *Big Data, 6*(1), 53–65. https://doi.org/10.1089/big.2018.0015

Paas, L. J. (1998). Mokken scaling characteristic sets and acquisition patterns of durable- and financial products. *Journal of Economic Psychology*. https://doi.org/10.1016/S0167-4870(98)00011-7

Padmanabhan, B., Hevner, A., Cuenco, M., & Shi, C. (2011). From information to operations: Service Quality and Customer Retention. In ACM Transactions on Management Information Systems (Vol. 2, pp. 1–21). https://doi.org/10.1145/2070710.2070712

Pan, L., Zhou, H., Liu, Y., & Wang, M. (2019). Global event influence model: integrating crowd motion and social psychology for global anomaly detection in dense crowds. *Journal of Electronic Imaging, 28*(2), 1–18. Retrieved from https://doi.org/10.1117/1.JEI.28.2.023033

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411–419. https://doi.org/10.1037/t69659-000

Paolacci, Gabriele, & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science, 23*(3), 184–188. https://doi.org/10.1177/0963721414531598

Patil, P. N., Lathi, R., & Chitre, V. (2012). Comparison of C5 . 0 & CART Classification algorithms using pruning technique. *International Journal of Engineering Research & Technology*, 1(4), 1–5. Patterson, P. G. (2004). A contingency model of behavioural intentions in a services context.

European Journal of Marketing, 38(9/10), 1304–1315. https://doi.org/10.1108/03090560410548997

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In *Handbook of research methods in personality psychology*. (pp. 224–239). New York, NY, US: The Guilford Press.

Paxton, A., & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods, 49*(5), 1630–1638. https://doi.org/10.3758/s13428-017-0874-x

Pearl, J., & Mackenzie, D. (2019). The book of why : the new science of cause and effect.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Peng, J., Zhang, J., Zhang, Y., Gong, P., Han, B., Sun, H., … Miao, D. (2018). A New Look at the Impact of Maximizing on Unhappiness: Two Competing Mediating Effects. *Frontiers in Psychology, 9*, 66. https://doi.org/10.3389/fpsyg.2018.00066

Pentland, A., & Heibeck, T. (2008). Honest signals : how they shape our world.

Perlich, C., & Provost, F. (2006). Distribution-based aggregation for relational learning with identifier attributes. *Machine Learning, 62*(1–2), 65–105. https://doi.org/10.1007/s10994-006-6064-1

Perugini, M., Richetin, J., & Zogmaister, C. (2010). Prediction of Behavior. In B. Gawronski & K. B. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 242–277). New York: Guilford Press.

Petrison, L. A., Blattberg, R. C., & Wang, P. (1993). Database marketing.Past, present, and future. *Journal of Direct Marketing, 7*(3), 27–43. https://doi.org/10.1002/dir.4000070306

Petrison, L. A., Blattberg, R. C., & Wang, P. (1997). Database marketing: Past, present, and future. *Journal of Direct Marketing, 11*(4), 109-125. https://doi.org/10.1002/(SICI)1522-7138(199723)11:43.0.CO;2-G

Petrozziello, A., Jordanov, I., & Sommeregger, C. (2018). Distributed Neural Networks for Missing Big Data Imputation. In 2018 *International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN.2018.8489488

Phadke, C., Uzunalioglu, H., Mendiratta, V. B., Kushnir, D., & Doran, D. (2013). Prediction of subscriber churn using social network analysis. *Bell Labs Technical Journal, 17*(4), 63–75. https://doi.org/10.1002/bltj.21575

Phan, T. Q., & Airoldi, E. M. (2015). A natural experiment of social network formation and dynamics. Proceedings of the National Academy of Sciences of the United States of America, 112(21), 6595–6600. https://doi.org/10.1073/pnas.1404770112

Prasad, U., & Madhavi, S. (2012). Prediction of Churn Behavior of Bank Customers. *Business Intelligence Journal, 5*(1), 96–101.

Prinzie, A., & Van den Poel, D. (2006a). Incorporating sequential information into traditional classification models by using an element / position-sensitive SAM. Decision Support Systems, 42, 508–526. https://doi.org/10.1016/j.dss.2005.02.004

Prinzie, A., & Van den Poel, D. (2006b). Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models. *European Journal of Operational Research, 170*(3), 710–734. https://doi.org/10.1016/j.ejor.2004.05.004

Pudipeddi, J. S., Akoglu, L., & Tong, H. (2014). User churn in focused question answering sites. In *International World Wide Web Conference Committee* (pp. 469–474). https://doi.org/10.1145/2567948.2576965

Pushpa, & Shobha, G. (2014). Social network classifier for churn prediction in telecom data. In ICACCS 2013 - Proceedings of the 2013 International Conference on Advanced Computing and Communication Systems: Bringing to the Table, Futuristic Technologies from Around the Globe. https://doi.org/10.1109/ICACCS.2013.6938744

Qin, X., Cunningham, P., & Salter-Townshend, M. (2016). Online Trans-dimensional von Mises-Fisher Mixture Models for User Profiles. *Journal of Machine Learning Research, 17*(200), 1–51. Retrieved from http://jmlr.org/papers/v17/15-454.html

Quercia, D., Kosinski, M., Stillwell, D., & Crowcroft, J. (2011). Our Twitter Profiles, Our Selves: Predicting Personality with Twitter. In *2011 IEEE Third Int'l Conference on Privacy, Security,*

Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing (pp. 180–185). IEEE. https://doi.org/10.1109/PASSAT/SocialCom.2011.26

Quinlan, J. R. (2004). *Data Mining Tools See5 and C5.0*. Retrieved from citeulike-article-id:8970822

R Development Core Team. (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Rachuri, K. K., Musolesi, M., Mascolo, C., Rentfrow, P. J., Longworth, C., & Aucinas, A. (2010).

EmotionSense: A mobile phones based adaptive platform for experimental social psychology research.

In UbiComp'10 - Proceedings of the 2010 ACM Conference on Ubiquitous Computing (pp. 281–290).

https://doi.org/10.1145/1864349.1864393

Radosavljevik, D., Van Der Putten, P., & Larsen, K. K. (2010). The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications: What to Predict, for Whom and Does the Customer Experience Matter? *Transactions on Machine Learning and Data Mining Journal, 3*(2), 80–99.

Ramli, S., & Nordin, S. (2018). Personality Prediction Based on Iris Position Classification Using Support Vector Machines. *Indonesian Journal of Electrical Engineering and Computer Science, 9*(3), 667. https://doi.org/10.11591/ijeecs.v9.i3.pp667-672

Ranganathan, C., Seo, D. B., & Babad, Y. (2006). Switching behavior of mobile users: Do users'

relational investments and demographics matter? European Journal of Information Systems, 15(3),

269–276. https://doi.org/10.1057/palgrave.ejis.3000616

Rehman, A., & Raza Ali, A. (2014). Customer Churn Prediction, Segmentation and Fraud Detection in Telecommunication Industry. In *ASE BigData/SocialInformatics/PASSAT/BioMedCom 2014 Conference* (pp. 1–9). Retrieved from https://www.researchgate.net/publication/304822719

Reinartz, W. J., & Kumar, V. (2003). The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing, 67*(1), 77–99. https://doi.org/10.1509/jmkg.67.1.77.18589

Rezzani, A. (2013). Big data. Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati. Apogeo Education - Maggioli Editore.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology, 7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Richards, N. M., & King, J. H. (2013). Three Paradoxes of Big Data. *Stanford Law Review Online, 66*. Retrieved from https://heinonline.org/HOL/Page?handle=hein.journals/slro66&id=41&div=7&collection=journals

Rizkallah, J. (2017). *The Big (Unstructured) Data Problem*. Retrieved September 28, 2019, from https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#35a610d9493a

Roe, R. A. (2008). Time in Applied Psychology. *European Psychologist, 13*(1), 37–52. https://doi.org/10.1027/1016-9040.13.1.37

Rokach, L., & Maimon, O. (2015). *Data mining with decision tree. Theory and applications* (2nd ed.). Singapore: World Scientific Publishing Co.

Ronin, A. (2019). How many programming languages are there today? Retrieved July 26, 2019, from https://frontnet.eu/how-many-programming-languages-are-there-today/

Roth, G., Assor, A., Niemiec, C. P., Ryan, R. M., & Deci, E. L. (2009). The emotional and academic consequences of parental conditional regard: Comparing conditional positive regard, conditional negative regard, and autonomy support as parenting practices. *Developmental psychology, 45*(4), 1119. https://doi.org/10.1037/a0015272

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471725382

Rowe, M. (2016). Mining User Development Signals for Online Community Churner Detection. In *ACM Transactions on Knowledge Discovery from Data* (Vol. 10, pp. 1–28). https://doi.org/10.1145/2798730

Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science, 290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*(3), 581. https://doi.org/10.2307/2335739

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. (D. B. Rubin, Ed.). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

Ryan, R. M. (1995). Psychological Needs and the Facilitation of Integrative Processes. *Journal of Personality, 63*(3), 397–427. https://doi.org/10.1111/j.1467-6494.1995.tb00501.x

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*. US: American Psychological Association. https://doi.org/10.1037/0003-066X.55.1.68

Ryan, R. M., Rigby, S., & King, K. (1993). Two types of religious internalization and their relations to religious orientations and mental health. *Journal of personality and social psychology*, 65(3), 586. https://doi.org/10.1037/0022-3514.65.3.586

Rygielski, C., Wang, J.-C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society, 24*(4), 483–502. https://doi.org/10.1016/S0160-791X(02)00038-6

Sammut, G., & Sartawi, M. (2012). Perspective-Taking and the Attribution of Ignorance. *Journal for the Theory of Social Behaviour, 42*(2), 181–200. https://doi.org/10.1111/j.1468-5914.2011.00485.x

Sandstrom, G. M., Lathia, N., Mascolo, C., & Rentfrow, P. J. (2017). Putting mood in context:

Using smartphones to examine how people feel in different locations. Journal of Research in

Personality, 69, 96–101. https://doi.org/10.1016/j.jrp.2016.06.004

Sato, T., Huang, B., Lefait, G., Kechadi, M. T., & Buckley, B. (2009). Kernel-based principal components analysis on large telecommunication data. *Conferences in Research and Practice in Information Technology Series*, 101(AusDM), 109–115.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. Organizational Research Methods, 12(2), 347-367. https://doi.org/10.1177/1094428107308906

Schmid Mast, M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social

Sensing for Psychology. Current Directions in Psychological Science, 24(2), 154–160.

https://doi.org/10.1177/0963721414560811

Schwartz, B., Ward, A., Lyubomirsky, S., Monterosso, J., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology, 83*(5), 1178–1197. https://doi.org/10.1037//0022-3514.83.5.1178

Schwartz, H., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., … Ungar, L. (2014). Towards Assessing Changes in Degree of Depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. https://doi.org/10.3115/v1/W14-3214

Semrl, J., & Matei, A. (2017). Churn prediction model for effective gym customer retention. *Proceedings of 4th International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2017*, 1–3. https://doi.org/10.1109/BESC.2017.8256379

Shao, J., Li, X., & Liu, W. (2007). The Application of AdaBoost in Customer Churn Prediction. In *2007 International Conference on Service Systems and Service Management*.

Shapley, L. S. (1988). A value for n-person games. In A. E. Roth (Ed.), *The Shapley value* (pp. 31–40). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511528446.003

Shaw, Z. (2013). Learn Python the hard way: A very simple introduction to the terrifyingly beautiful world of computers and code (3rd ed.). Addison-Wesley Professional.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1–32. Retrieved from https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science, 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Shukla, P. (2004). Effect of product usage, satisfaction and involvement on brand switching behaviour. Asia Pacific Journal of Marketing and Logistics, 16(4), 82–104. https://doi.org/10.1108/13555850410765285

Sidhu, A. (2005). Canadian Cellular Industry: Consumer Switching Behaviour. Fraser University.

Sijbrandij, S. (2017). *Coding Careers: Developers As The Next Mass Profession*. Retrieved September 27, 2019, from https://www.forbes.com/sites/forbestechcouncil/2017/12/12/coding-careers-developers-as-the-next-mass-profession/#3e7ca02febd9

Sivasankar, E., & Vijaya, J. (2017). Customer segmentation by various clustering approaches and building an effective hybrid learning system on churn prediction dataset. In *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-981-10-3874-7_18

Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry, 27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044

Spreng, R. A., Harrell, G. D., & Mackoy, R. D. (1995). Service recovery: Impact on satisfaction and intentions. Journal of Services Marketing, 9(1), 15–23. https://doi.org/10.1108/08876049510079853

Srinivasan, S. S., Anderson, R., & Ponnavolu, K. (2002). Customer loyalty in e-commerce: an exploration of its antecedents and consequences. *Journal of Retailing, 78*(1), 41–50. https://doi.org/10.1016/S0022-4359(01)00065-3

Srivastava, K., & Sharma, N. K. (2013). Service Quality, Corporate Brand Image, and Switching Behavior: The Mediating Role of Customer Satisfaction and Repurchase Intention. Services Marketing Quarterly, 34(4), 274–291. https://doi.org/10.1080/15332969.2013.827020

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., … Bühner, M. (2019). Personality Research and Assessment in the Era of Machine Learning. *PsyArXiv*. https://doi.org/10.31234/OSF.IO/EFNJ8

Stauss, B., & Friege, C. (1999). Regaining service customers: Costs and benefits of regain management. *Journal of Service Research, 1*(4), 347–361. https://doi.org/10.1177/109467059914006

Steinberg, D. (2009). CART: Classification and Regression Trees. In X. Wu & V. Kumar (Eds.), *The Top Ten Algorithms in Data Mining* (pp. 193–216). Chapman and Hall/CRC. https://doi.org/10.1201/9781420089653-17

Stephens-Davidowitz, S., & Pinker, S. (2017). *Everybody lies: big data, new data, and what the Internet can tell us about who we really are* (First edition.). New York NY: Dey St. an imprint of William Morrow. Retrieved from https://www.worldcat.org/title/everybody-lies-big-data-new-data-and-what-the-internet-can-tell-us-about-who-we-really-are/oclc/985108386

Stevens, J. R., & Soh, L.-K. (2018). Predicting similarity judgments in intertemporal choice with machine learning. *Psychonomic Bulletin & Review, 25*(2), 627–635. https://doi.org/10.3758/s13423-017-1398-1

Subramanya, K. B., & Somani, A. K. (2017). Enhanced feature mining and classifier models to predict customer churn for an e-retailer. *Big Data Analytics: Tools and Technology for Effective Planning*, 293–309. https://doi.org/10.1201/b21822

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to data mining and its applications*. Springer.

Sundarkumar, G. G., & Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. *Engineering Applications of Artificial Intelligence, 37*, 368–377. https://doi.org/10.1016/j.engappai.2014.09.019

Sundarkumar, G. G., Ravi, V., & Siddeshwar, V. (2016). One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. In *2015 IEEE International Conference on Computational Intelligence and Computing Research*, ICCIC 2015. https://doi.org/10.1109/ICCIC.2015.7435726

Tamassia, M., Raffe, W., Sifa, R., Drachen, A., Zambetta, F., & Hitchens, M. (2016). Predicting player churn in destiny: A Hidden Markov models approach to predicting player departure in a major online game. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)* (pp. 1–8). IEEE. https://doi.org/10.1109/CIG.2016.7860431

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Association Analysis: Basic Concepts and Algorithms. In P.-N. Tan, M. Steinbach, & V. Kumar (Eds.), *Introduction to Data mining*. Pearson Addison Wesley. https://doi.org/10.1111/j.1600-0765.2011.01426.x

Tang, L., Thomas, L., Fletcher, M., Pan, J., & Marshall, A. (2014). Assessing the impact of derived behavior information on customer attrition in the financial service industry. *European Journal of Operational Research, 236*(2), 624–633. https://doi.org/10.1016/j.ejor.2014.01.004

Tax, S. S., Brown, S. W., & Chandrashekaran, M. (1998). Customer evaluations of service complaint experiences: Implications for relationship marketing. Journal of Marketing, 62(2), 60–76. https://doi.org/10.2307/1252161

Taylor, S. A., Jaques, N., Nosakhare, E., Sano, A., & Picard, R. (2017). Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing*, 1–1. https://doi.org/10.1109/TAFFC.2017.2784832

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science, 290*(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

Teng, H. S., Chen, K., & Lu, S. C. (1990). Adaptive real-time anomaly detection using inductively generated sequential patterns. In *Proceedings. 1990 IEEE Computer Society Symposium on Research in Security and Privacy* (pp. 278–284). IEEE. https://doi.org/10.1109/RISP.1990.63857

Therneau, T., Atkinson, B., & Ripley, B. (2017). *rpart: Recursive Partitioning and Regression Trees*. Retrieved from https://cran.r-project.org/package=rpart

Thomas, J. S. (2001). A methodology for linking customer acquisition to customer retention. *Journal of marketing research, 38*(2), 262-268. https://doi.org/10.1509/jmkr.38.2.262.18848

Thomas, J. S., Blattberg, R. C., & Fox, E. J. (2004). Recapturing Lost Customers. *Journal of Marketing Research, 41*(1), 31–45. Retrieved from https://doi.org/10.1142/9789814287067_0015

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., … Li, L.-J. (2015). YFCC100M: *The New Data in Multimedia Research*. https://doi.org/10.1145/2812802

Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert Systems with Applications, 39*(3), 2597–2605. https://doi.org/10.1016/j.eswa.2011.08.115

Tonidandel, S., King, E. B., & Cortina, J. M. (2016). Big Data Methods: Leveraging Modern Data Analytic Techniques to Build Organizational Science. *Organizational Research Methods, 21*(3), 525–547. https://doi.org/10.1177/1094428116677299

Torgo, L. (2010). *Data Mining with R, learning with case studies*. New York: Chapman and Hall/CRC data mining and knowledge discovery series. Retrieved from http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR

Troester, M. (2012). Big Data meets Big Data analytics: Three Key Technologies for Extracting Real-Time Business Value from the Big Data That Threatens to Overwhelm Traditional Computing Architectures. SAS. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=

2ahUKEwjgvKfxuZTlAhXOEVAKHU23C9YQFjAAegQIAhAC&url=http%3A%2F%2Feric.univ-lyon2.fr%2F~ricco%2Fcours%2Fslides%2Fsources%2Fbig-data-meets-big-data-analytics-105777.pdf&usg=AOv

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management, 24*(1), 53–62. https://doi.org/10.1016/j.jom.2005.03.001

Turban, E., Aronson, J. E., Liang, T.-P., & Sharda, R. (2007). *Decision Support Systems and Business Intelligence* (7th ed.). Pearson. https://doi.org/10.1017/CBO9781107415324.004

Turner, R. M., Bird, S. M., & Higgins, J. P. T. (2013). The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews. *PLoS ONE, 8*(3), e59202. https://doi.org/10.1371/journal.pone.0059202

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). *The Anatomy of the Facebook Social Graph*. Retrieved from http://arxiv.org/abs/1111.4503

Uncles, M. D., East, R., & Lomax, W. (2013). *Good customers: The value of customers by mode of acquisition*. Australasian Marketing Journal. https://doi.org/10.1016/j.ausmj.2013.02.003

Ushey, K., Allaire, J. J., Tang, Y., Lewis, B., & Geelnard, M. (2019). *reticulate: Interface to "Python."*

van Capelleveen, G., Poel, M., Mueller, R. M., Thornton, D., & van Hillegersberg, J. (2016). Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. *International Journal of Accounting Information Systems, 21*, 18–31. https://doi.org/10.1016/j.accinf.2016.04.001

Van den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*, 196–217. https://doi.org/10.1016/S0377-2217(03)00069-9

Van den Poel, D., & Larivière, B. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert System with Applications, 29*, 472–484. https://doi.org/10.1016/j.eswa.2005.04.043

van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2007). Multiple Imputation of Item Scores in Test and Questionnaire Data, and Influence on Psychometric Results. *Multivariate Behavioral Research, 42*(2), 387–414. https://doi.org/10.1080/00273170701360803

Vandecruys, O., Martens, D., Baesens, B., Mues, C., De Backer, M., & Haesen, R. (2008). Mining software repositories for comprehensible software fault prediction models. *Journal of Systems and Software, 81*(5), 823–839. https://doi.org/10.1016/j.jss.2007.07.034

Varian, H. (2014). Big Data: New Tricks for Econometrics. *The Journal of Economic Perspectives, 28*. https://doi.org/10.1257/jep.28.2.3

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Applied Soft Computing Journal, 14*(PART C), 431–446. https://doi.org/10.1016/j.asoc.2013.09.017

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications, 38*(3), 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023

Vercellis, C. (2009). *Business Intelligence*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470753866

Verhoef, P. C., & Donkers, B. (2005). The effect of acquisition channels on customer loyalty and cross-buying. *Journal of Interactive Marketing, 19*(2), 31–43. https://doi.org/10.1002/dir.20033

Verhoef, P. C., Kooge, E., & Walk, N. (2016). Creating value with big data analytics: Making smarter marketing decisions. Routledge.

Verhoef, P. C., Spring, P. N., Hoekstra, J. C., & Leeflang, P. S. H. (2003). The commercial use of segmentation and predictive modeling techniques for database marketing in the Netherlands. *Decision Support Systems, 34*(4), 471–481. https://doi.org/10.1016/S0167-9236(02)00069-6

Vieira, S., Pinaya, W. H. L., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews, 74*, 58–75. https://doi.org/10.1016/J.NEUBIOREV.2017.01.002

Vo, N. N., Liu, S., Brownlow, J., Chu, C., Culbert, B., & Xu, G. (2018). Client Churn Prediction with Call Log Analysis. In *International Conference on Database Systems for Advanced Applications* (Vol. 1, pp. 752–763). Springer International Publishing. https://doi.org/10.1142/9789814537308

Waddams Price, C., Webster, C., & Zhu, M. (2013). Searching and Switching: Empirical estimates of consumer behaviour in regulated markets. Retrieved from http://ec.europa.eu/consumers/consumer_research/index_en.htm,

Wang, C. K. J., Pyun, D. Y., Kim, J. Y., & Chatzisarantis, N. L. D. (2009). Testing for multigroup invariance of the perceived locus of causality in sport. *Personality and Individual Differences, 47*(6), 590–594. https://doi.org/https://doi.org/10.1016/j.paid.2009.05.008

Wang, Q., Guo, B., Peng, G., Zhou, G., & Yu, Z. (2016). CrowdWatch: Pedestrian safety assistance with mobile crowd sensing. In UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 217–220). Association for Computing Machinery, Inc. https://doi.org/10.1145/2968219.2971433

Wang, R., Aung, M. S. H., Abdullah, S., Brian, R., Campbell, A. T., Choudhury, T., … Ben-Zeev, D. (2016). CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. In UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 886–897). Association for Computing Machinery, Inc. https://doi.org/10.1145/2971648.2971740

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., … Campbell, A. T. (2014). Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 3–14). Association for Computing Machinery, Inc. https://doi.org/10.1145/2632048.2632054

Wang, R., Harari, G., Hao, P., Zhou, X., & Campbell, A. T. (2015). SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 295–306). New York, NY, USA: ACM. https://doi.org/10.1145/2750858.2804251

Wang, X., Zhao, K., & Street, N. (2017). Analyzing and predicting user participations in online health communities: A social support perspective. *Journal of Medical Internet Research, 19*(4). https://doi.org/10.2196/jmir.6834

Wedel, M., & Kannan, P. K. (2016). Marketing Analytics for Data-Rich Environments. Journal of Marketing, 80(6), 97–121. https://doi.org/10.1509/jm.15.0413

Wei, C. P., & Chiu, I. T. (2002). Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications, 23*(2), 103–112. https://doi.org/10.1016/S0957-4174(02)00030-1

Weiss, G. M. (2004). Mining with Rarity: A Unifying Framework. *Sigkdd Explorations, 6*(1), 7–19. https://doi.org/10.1145/1007730.1007734

Weiss, G. M. (2010). Mining with Rare Cases. In O. Maimon & L. Rokach (Eds.), Data *Mining and Knowledge Discovery Handbook* (2nd ed., pp. 747–757). Springer. https://doi.org/10.1007/978-0-387-09823-4_38

Weiss, S. M., & Indurkhya, N. (1998). *Predictive data mining : a practical guide*. Morgan Kaufmann Publishers.

Welles, B. F., & Contractor, N. (2015). Individual Motivations and Network Effects: A Multilevel Analysis of the Structure of Online Social Relationships. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 180–190. https://doi.org/10.1177/0002716214565755

Wenzel, R., & Van Quaquebeke, N. (2018). The Double-Edged Sword of Big Data in Organizational and Management Research. *Organizational Research Methods, 21*(3), 548–591. https://doi.org/10.1177/1094428117718627

Whelan, T. J., & DuVernet, A. M. (2015). The big duplicity of big data. Industrial and Organizational Psychology, 8(4), 509–515. https://doi.org/10.1017/iop.2015.75

Wickham, H., & Grolemund, G. (2017). *R for data science : import, tidy, transform, visualize, and model data*. Sebastopol, CA, USA: O'Reilly Media.

Wind, J., & Rangaswamy, A. (2001). Customerization: The Next Revolution in Mass Customization. *Journal of Interactive Marketing, 15*(1), 13–32. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=4211345&site=ehost-live&scope=site

Wojcik, S. P., Hovasapian, A., Graham, J., Motyl, M., & Ditto, P. H. (2015). Conservatives report, but liberals display, greater happiness. *Science, 347*(6227), 1243–1246. https://doi.org/10.1126/science.1260817

Wrzus, C., & Mehl, M. R. (2015). Lab and/or Field? Measuring Personality Processes and Their Social Consequences. European Journal of Personality, 29(2), 250–271. https://doi.org/10.1002/per.1986

Wrzus, C., Brandmaier, A. M., von Oertzen, T., Müller, V., Wagner, G. G., & Riediger, M. (2012). A New Approach for Assessing Sleep Duration and Postures from Ambulatory Accelerometry. PLoS ONE, 7(10), e48089. https://doi.org/10.1371/journal.pone.0048089

Xie, X., Wang, C., Chen, S., Shi, G., & Zhao, Z. (2017). Real-Time Illegal Parking Detection System Based on Deep Learning. In *International Conference on Deep Learning Technologies* (pp. 23–27). https://doi.org/10.1145/3094243.3094261

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3 PART 1), 5445–5449. https://doi.org/10.1016/j.eswa.2008.06.121

Yabas, U., & Cankaya, H. C. (2013). Churn prediction in subscriber management for mobile and wireless communications services. *2013 IEEE Globecom Workshops*, GC Wkshps 2013, 991–995. https://doi.org/10.1109/GLOCOMW.2013.6825120

Yan, L. Y. L., Miller, D. J., Mozer, M. C., & Wolniewicz, R. (2001). Improving prediction of customer behavior in nonstationary environments. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings* (Cat. No.01CH37222) (Vol. 3). https://doi.org/10.1109/IJCNN.2001.938518

Yang, X. C., Wu, J., Zhang, X. H., & Lu, T. J. (2008). Using decision tree and association rules to predict cross selling opportunities. In *Proceedings of the 7th International Conference on Machine Learning and Cybernetics, ICMLC*. https://doi.org/10.1109/ICMLC.2008.4620698

Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yetton, B. D., Revord, J., Margolis, S., Lyubomirsky, S., & Seitz, A. R. (2019). Cognitive and Physiological Measures in *Well-Being Science: Limitations and Lessons. Frontiers in Psychology*, 10, 1630. https://doi.org/10.3389/fpsyg.2019.01630

Ylijoki, O., & Porras, J. (2016). Perspectives to Definition of Big Data: A Mapping Study and Discussion. *Journal of Innovation Management, 4*(1), 69–91. https://doi.org/10.24840/2183-0606_004.001_0006

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America, 112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). The Behavioral Consequences of Service Quality. *Journal of Marketing, 60*(2), 31–46. https://doi.org/10.2307/1251929

Zhang, X., Liu, Z., Yang, X., Shi, W., & Wang, Q. (2010). Predicting customer churn by integrating the effect of the customer contact network. In *Proceedings of 2010 IEEE International Conference on Service Operations and Logistics, and Informatics, SOLI 2010* (pp. 392–397). https://doi.org/10.1109/SOLI.2010.5551545

Zhang, X., Zhu, J., Xu, S., & Wan, Y. (2012). Predicting customer churn through interpersonal influence. Knowledge-Based Systems, 28, 97–104. https://doi.org/10.1016/j.knosys.2011.12.005

Zhang, Yingying, Qi, J., Shu, H., & Li, Y. (2006). Case Study on CRM : Detecting Likely Churners with Limited Information of Fixed-line Subscriber. In 2006 International conference on service systems and service management.

Zhang, Yongbin, Liang, R., Li, Y., Zheng, Y., & Berry, M. (2011). Behavior-based telecommunication churn prediction with neural network approach. In Proceedings - 2011 International Symposium on Computer Science and Society, ISCCS 2011 (pp. 307–310). https://doi.org/10.1109/ISCCS.2011.89

Zhang, Z.-Z., Chen, Q., Ke, S.-F., Wu, Y.-J., Qi, F., & Zhang, Y.-P. (2008). Ranking Potential Customers Based on Group-Ensemble. International Journal of Data Warehousing and Mining, 4(2), 79–89. https://doi.org/10.4018/jdwm.2008040109

Zhao, J., Zhang, M., He, C., & Zuo, K. (2019). Data-Driven Research on the Matching Degree of Eyes, Eyebrows and Face Shapes. *Frontiers in Psychology*, 10, 1466. https://doi.org/10.3389/fpsyg.2019.01466

Zhao, Q., & Hastie, T. (2019). Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, 1–10. https://doi.org/10.1080/07350015.2019.1624293

Zhu, B., Baesens, B., Backiel, A., & vanden Broucke, S. K. L. M. (2017). Benchmarking sampling techniques for imbalance learning in churn prediction. *Journal of the Operational Research Society*. https://doi.org/10.1057/s41274-016-0176-1