GAMLSS FOR BIG DATA: ROC CURVE PREDICTION USING TWITTER DATA

Paolo Mariani [1], Andrea Marletta [1] and Mariangela Sciandra [2]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, (e-mail: paolo.mariani@unimib.it, andrea.marletta@unimib.it)

[2] Dipartimento Scienze Economiche Aziendali e Statistiche, University of Palermo, (e-mail: mariangela.sciandra@unipa.it)

ABSTRACT: During last years, Big Data appears as one of the most innovative and growing scientific area of interest. In this field, finding reliable methods to make accurate predictions represents one of the most inspirational challenges. The way to make prediction in the following paper is the use of ROC (Receiver Operating Characteristic) Curve, a binary prediction tool, often used for medical tests. The attention is focused in particular on the implementation of ROC Curve in GAMLSS (Generalized Additive Models for Location Scale and Shape), semi-parametric models suitable for huge and flexible dataset. An application will be shown where the class of GAMLSS is applied to Twitter data in order to predict number of interactions for a tweet given a set of explanatory variables.

KEYWORDS: GAMLSS, ROC curve, Twitter, Big Data

1 Introduction

Big Data analysis represent the new challenge to face for mining information from data. The term 'Big' seems to pertain to quantity of the data, but actually there is also a new way to look for data focusing the attention on their quality. On account of this, it is not sufficient to have a dataset with a billion of observations to classify it as 'Big Data'.

This type of data have to be endowed with some particular features as Volume, Velocity, Variety, Value, Veridicality and Validity (Liberati and Mariani, 2016). These characteristics listed above shows that data has to be a huge quantity but they have also to be susceptible to changes and continually updated without losing property of truth and effectiveness. Social media seems to be the typical area where this happens, this is why a lot of times Big Data is associated with the term 'Internet of Things'. Moreover, it is a field where measuring the possibility of interactions between profiles is one of the most

interesting purposes to chase. In this paper the attention is focused on social media data, in particular on Twitter data. Twitter was created in 2006 and nowadays represents one the biggest used social network in the world with more than 319 million monthly active users. Peculiarity of Twitter stands in the possibility to post and interact with messages, "tweets", restricted to 140 characters. Main objective of this paper stands in using the ROC Curve in GAMLSS to measure the interactions in the tweet history for a Twitter user.

The paper is organized as follows. Section 2 is devoted to the presentation of GAMLSS. The proposal of implementing ROC Curve in GAMLSS is described in section 3. An application of this method about Twitter data is showed in section 4. Finally, conclusions and main remarks are discussed in the last part of the paper.

2 The GAMLSS models

General Additive Models for Location Scale and Shape were introduced firstly by Rigby and Stasinopoulos (2001) as a way of overcoming some of the limitations associated with GLM and GAM (Nelder & Wedderburn, 1972 and Hastie & Tibshirani, 1990). They represent a class of semi-parametric models where all the parameters of the assumed distribution for the response can be modelled as additive functions of the explanatory variables. Since GAMLSS are very flexible, they appears to be particularly suited for analysing Big Data. The basic hypothesis is that since Big Data are usually very complex to inspect, then working with multiple equation models could be a possible solution. So, assuming the response variable Y to follow a four parameters distribution $Y \sim D(\theta)$ with $\theta = (\mu, \sigma, \nu, \tau)$, where μ and σ are location and scale parameters while ν and τ shape parameters. Equation (1) represents the formulation of GAMLSS given by Rigby and Stasinopoulos (2005):

$$g_k(\theta_k) = \eta_k = \mathbf{X}_k \beta_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \qquad k = 1,2,3,4 \tag{1}$$

where $g_k(.)$ are known monotonic link functions relating in a parametric way the distribution parameters to the explanatory variables $\mathbf{X}_k$ and h_{jk} represent the non-parametric additive terms. The vector of parameters β_k and the non parametric terms are estimated by maximizing a penalized likelihood function l_p given by

$$l_p = l - \frac{1}{2} \sum_{k=1}^{p} \sum_{j=1}^{J_k} \lambda_k \gamma'_{jk} G_{jk} \gamma_{jk} \tag{2}$$

3 ROC Curve in GAMLSS

Receiver Operating Characteristic (ROC) curve is one of the most used tool to measure the accuracy of a binary test. In presence of a dichotomous outcome and a binary prediction, four different situations can appear: True Positive (TP) when outcome and prediction test are both positive; True Negative (TN) when outcome and prediction test are both negative; False Positive (FP) when outcome is negative and prediction test is positive; False Negative (FN) when outcome is positive and prediction test is negative (Pepe, 2003). A test is characterized by level of accuracy, sensitivity and specificity. Accuracy of a test could be computed as the sum of the main diagonal ($TP + TN$) over the n total number of observations. Sensitivity and specificity measure the proportion of units that are correctly predicted when outcome is positive or negative.

For a binary test, ROC curve is a graphical plot of sensitivity vs (1 - specificity), where each point of the curve represents a different value for the cutoff to classify a statistical unit. Another way to check if a prediction test is informative is to compute the Area Under a ROC Curve (AUC). This index is the most commonly used method for summarizing a diagnostic test's overall accuracy. It ranges from 0 to 1 (perfect classification) and takes value 0.5 for a random test.

ROC curves are suitable to binary data because in logistic regression sensitivity and specificity are computed starting by fitted values of $\hat{p} = P(Y = 1)$ ranged in (0,1). Using GAMLSS, fitted values are not necessary ranged in (0,1), so it is necessary to calculate probabilities $\hat{p}$ starting from GAMLSS fitted values: in the proposed approach, $\hat{p}$ are obtained using the difference between 1 and the density function of the selected distribution in GAMLSS at an established cut-off α, where parameters are substituted with fitted values computed for GAMLSS model. Using this approach there exists a direct correspondence between each observation y and a probability $\hat{p}$ that lies in (0,1). Given the probabilities it is possible to derive the ROC curve.

The use of this approach needs to be validated comparing it with other statistical models. When ROC curves are used, two possible ways of comparing different statistical models are possible. The first one is a graphical comparison, where different ROC curves are drawn in order to identify the higher curve. The higher the curve, the better the prediction. Secondly, the AUC index can be computed for all models; the model with a higher AUC index will be the best.

4 Application and results

An application based on real data is here described to compare prediction from selected GAMLSS with discrete GLM. Statistical units are represented by tweets extracted in 2016 from the official account of F1 Italy Circuit in Monza (@Autodromo_Monza). Total number of observations is 737. The selected GAMLSS shows relationships between count of "likes" for a tweet and three explanatory variables: count of hashtags (#), count of tags (@) and count of links. Selected distribution for count of likes among discrete in GAMLSS is the Sichel distribution (Sichel, 1973). In order to obtain the ROC curve as a prediction tool, it is necessary to split up the dataset in two subsets: the training (75%) and the validation set (25%).

The selected GAMLSS represents the starting point for the estimate of the ROC curve. This model is fitted on the complete dataset with different weights for training ($w = 1$) and validation ($w = 0$) sets. Predicted values $\hat{\mu}, \hat{\sigma}, \hat{\nu}$ are extracted for this weighted model and included in $1 - F(\alpha)$ where F is the density function for Sichel distribution and α is the selected cutoff corresponding number of "likes" from 1 over to 5. The use of the cutoff allows to dichotomize a tweet as likeable or not likeable. If for example $\alpha = 1$, a tweet is likeable if it received at least 1 "like", and so on.

For comparative purposes, Poisson and Negative Binomial response GLM (Generalized Linear Model) are presented since the response variable distribution is discrete.

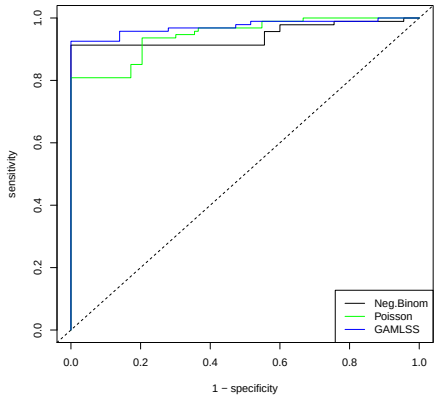

Figure 1. *ROC Curves for different statistical models*

In Figure 1, for $\alpha = 1$ and a single training and validation sample, ROC curves are shown for different statistical models. As it is possible to observe, the blue one representing GAMLSS with Sichel distribution is slightly above all the other curves. Graphical comparison is not enough because it could depend on the chosen sample. Using the AUC index it is possible to compute an average measure for each model for all samples (see Table 1).

Model	GLM NBin	GLM Pois	**GAMLSS**
$\overline{AUC}$	0.951	0.949	**0.967**

Table 1. *Mean comparison of AUC for selected models*

In Figure 2, in a scenario with $\alpha = 1$, on the left the AUC indices for 50 training and validation samples are shown on the same graph. On the x-axis we have the sampling index and each point represents a resulting value of AUC. It is possible to note that blue points denoting GAMLSS are the highest point. On the right, boxplot for 50 AUC indexes are displayed. The blue box-plot related to GAMLSS is the highest as it is reasonable to expect from previous considerations. Furthermore, AUC indices for GAMLSS present a smaller variability denoting less dependence on the choice of sample for the splitting.

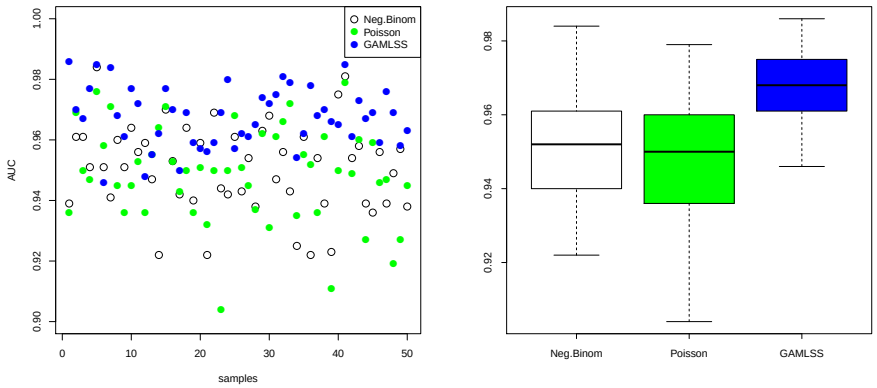

Figure 2. *AUC indices for 50 test and validation samplings*

5 Conclusions

The article proposes the implementation of ROC Curve in GAMLSS as a prediction tool in a Big Data context. Moreover, an application of the proposed approach is presented for Twitter data. Count of "like" is modelled using GAMLSS with a Sichel probability distribution.

GAMLSS seems to be a possible alternative choice in modelling Big Data thanks to their flexibility. Particularly, the implementation of ROC Curve in GAMLSS proves to be a good prediction tool performing better than usual statistical models. In order to make a comparison with other classes of statistical models, AUC index has been computed for Poisson and Negative Binomial regression. The output of possible choice has been measured in terms of AUC and when the Twitter dataset has been used, AUC values derived from GAMLSS are higher than Poisson and Negative Binomial regression.

Future research will focus the attention on results extracting from prediction for different values of α or adding other explanatory variables or non-linear terms in the Twitter selected model.

References

HASTIE, T., & TIBSHIRANI, R. 1990. *Generalized Additive Models*. Chapman and Hall.

LIBERATI, C., & MARIANI, P. 2015. *Book of Abstracts, 10th Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society*. CUEC Editrice.

NELDER, J. A., & WEDDERBURN, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A*, **135**(3), 370–382.

PEPE, MARGARET SULLIVAN. 2003. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

RIGBY, R.A., & STASINOPOULOS, D. M. 2001. The GAMLSS project: a Flexible Approach to Statistical Modelling. *Proceedings of the 16th International Workshop on Statistical Modelling*, 249–256.

RIGBY, R.A., & STASINOPOULOS, D. M. 2005. Generalized additive models for location, scale and shape). *Applied Statistics*, **54**(3), 507–544.

SICHEL, HS. 1973. Statistical valuation of diamondiferous deposits. *Journal of the Southern African Institute of Mining and Metallurgy*, **73**(7), 235–243.