Paper 093-2013

Using the Boosting Technique to Improve the Predictive Power of a Credit Risk Model

Alejandro Correa, Luxembourg University, Luxembourg.

Andrés González, Banco Colpatria, Bogotá, Colombia

Darwin Amézquita, Banco Colpatria, Bogotá, Colombia

ABSTRACT

In developing a predictive model, the complexity of the population used to build the model can lead to very weak scorecards when a traditional technique such as logistic regression or an MLP neural network is used. For these cases some nontraditional methodologies like boosting could help improve the predictive power of any learning algorithm. The idea behind this technique is to combine several weak classifiers to produce a much more powerful model. In this paper, boosting methodology is used to enhance the development of a credit risk scorecard in combination with several different techniques, such as logistic regression, MLP neural networks, and others, in order to compare the results of all methodologies and determine in which cases the boosting algorithm increases model performance.

INTRODUCTION

It can be said that the predictive power of a credit risk scorecard increases or decreases to the extent that it is able to identify and efficiently separate good clients from bad clients throughout the entire development distribution. Unfortunately, it is very common to find that the development distribution is extremely complex and just as the blind men and the elephant poem[1] by John Godfrey Saxe, where six blind men described an elephant and each of them described it in a different way based on what part of the body they where touching (Figure 1). A traditional predictive model can be very powerful regarding the task of separating good clients from bad clients of a specific part of the development distribution but very weak regarding the global distribution. Thus, if two predictive models were built on the same complex population using different variables and techniques, the result would show that both models are powerful in explaining different portions of the total distribution.

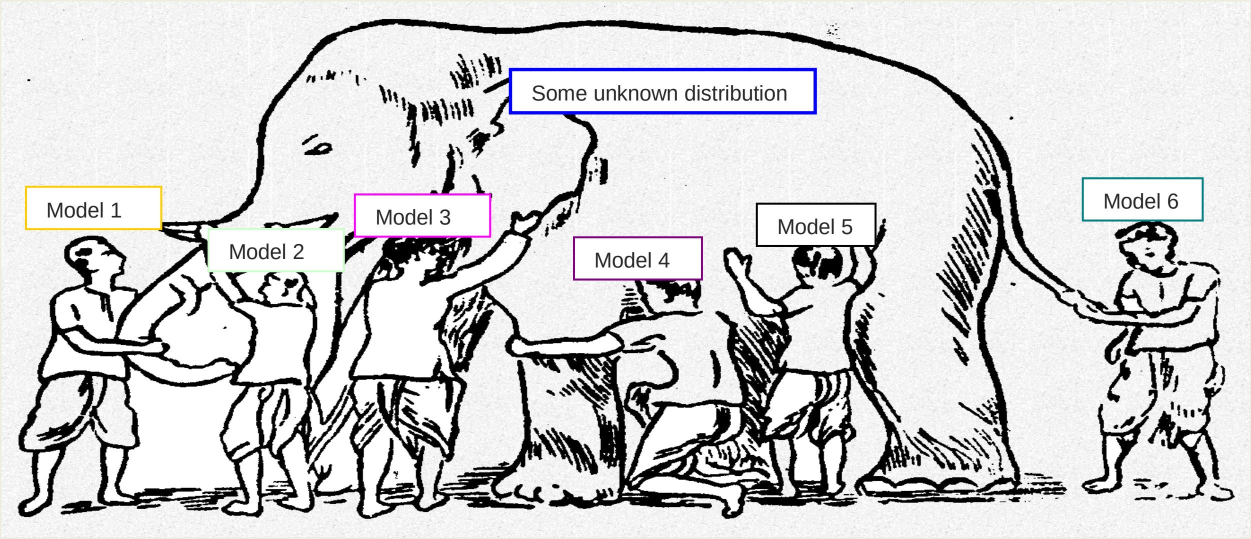

Figure 1.The Blind Men and the Elephant

[1] http://courses.cs.vt.edu/~cs1104/Introduction/6.blind.men.html

Given this, experts realized that for complex distributions a single traditional predictive model would not work. If the distribution is broken down into multiple sub-distributions and a different model for each one is develop, this would do the job. From this divide and conquer approach, the Ensemble and Boosting techniques were born. Both methods combine multiple weak classifiers in order to produce a more powerful one. The major difference between these two techniques depends on the way that classifiers are created.

It's been over three years since these methods became famous in the Netflix Prize competition[2] given their supremacy when describing the "Elephant" of the total distribution over the traditional single models. Yet, the financial institutions continue to develop credit risk scorecards using the "traditional" technique.

This paper seeks to compare the "traditional" single credit risk models developed using different techniques such as logistic regression, decision trees and neural networksversus the Boosting and Ensemble models. Results showed that Ensemble models improve accuracy and predictive power over single model methods and therefore are more efficient at the task of separating good from the bad clients.

GENERAL CONCEPTS

In this section the different algorithms used for this project are described.

LOGISTIC REGRESION

A logistic regression is a type of regression analysis that is used to find the relationship that exists between a dichotomous dependent variable and the independent variables which can be continuous or categorical. It is used to predict the probability of occurrence of an outcome given that the output values are adjusted between 0 and 1 using the logistic function.

Logistic regression is the traditional method used by financial institutions to develop the credit risk scorecards which attempt to attribute a rating (score) to a client indicating the predicted probability that the customer reflects a certain behavior. It is still one of the most widely used techniques because it has two strong features in its favor: i) Simple model development and ii) ease of interpretability.

ARTIFICIAL NEURAL NETWOR

An artificial neural network is a mathematical model created to replicate the structure and functionality of the real nervous system (Rosenblatt, 1962) that consists of a set of units called neurons that are highly interconnected. In an artificial neural network, there is an input layer that represents the input variables to be used in the model. It also has hidden layers and each layer contains hidden units. The hidden units receive a weighted sum of the inputs and an activation function is applied. Finally, the neural network has an output layer that computes for the result of the whole process, by receiving a weighted sum of the hidden units output and applying an activation function to this sum. Information passes from one layer to the next by unidirectional connections. The hidden units and the output unit can have a variety of activation functions. The neural network finds the weights by an iterative process through different types of algorithms. For further information regarding the usage of an MLP neural network on credit scoring, please refer to Correa, Gonzalez and Ladino (2011).

DECISION TREE MODEL

A Decision tree is a commonly used methodology to create models for predicting a target variable based on the input features values. This supervised algorithm attempts to split the entire population into sub-population groups according to their characteristics. In classification trees, like in the case of this paper, the goal is find groups that help to predict the target variable.

[2] http://bloggasm.com/how-the-ensemble-squeaked-by-to-win-the-netflix-prize

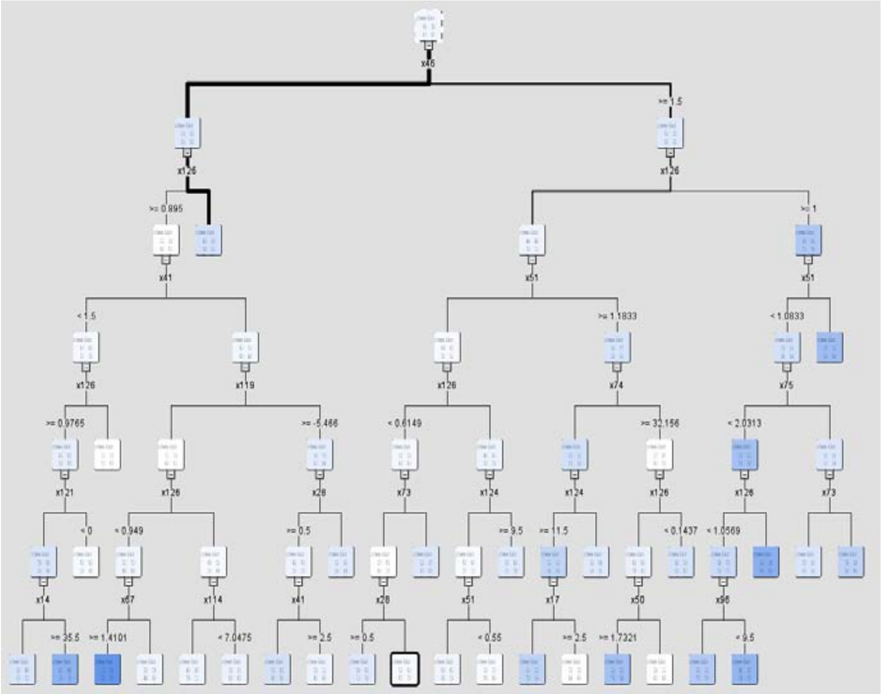

Figure 2. Decision tree model

As can be seen in Fig2., each node corresponds to one split of the population, where nodes are found using machine learning. In this paper the process of splitting the whole population in subsets is performed using a top-down methodology, where the entire population is split into two subsets at once, and this process is repeated until the desired level of division is reached.

BOOSTING

Boosting is a machine learning algorithm that attempts to create a single strong classifiers starting from a set of weak classifiers (Schapire, 1990). Weak classifiers are those whose error rate is only slightly better than random assignation while a strong classifier is highly accurate and close to true classification.

Boosting algorithms are iterative learning processes that combine weak classifiers in order to create a final strong classifier. At each step of the iteration a new weak classifier is added and weighted according the learner’s accuracy and/or the step in the iterative process. Meanwhile, the data is reweighted by assigning more importance to the still misclassified observations; as a consequence newest classifiers focus more on the population that was not correctly classified in previous steps.

ENSEMBLE

Ensemble, as its name suggest, is the combination of techniques in order to create a powerful classifier. The idea behind the ensemble methodology is that using different techniques trained with different sets of features or databases will jointly perform better than a single technique. An ensemble model may incorporate any type of models, even other ensembles. The intuition behind why the ensemble modeling works, is that, by combining different sources of knowledge a better understanding of the problem is attached and by doing that better predictions can be made.

The crucial part of the ensemble is how to combine the different models. There are several types of combination: Simple averaging; weighted averaging; majority voting; regression; and optimization (Gao, 2010).

There is a clear relationship between the Boosting and Ensemble models given that both combine models in order to produce better classifiers. The main difference between these techniques is that weak Ensemble classifiers are built separately using different uncorrelated techniques while Boosting creates weak classifiers using a particular optimization algorithm that focuses on the misclassified observations from previous steps.

DATABASE DESCRIPTION

For this project a dataset containing examples of bad and good credit card customers was used. The dataset contains a total of 175,093 observations, where 95,058 performed as good clients, which is equivalent to 54.29% of the entire database. The remaining 80,035 observations performed as bad clients, representing 45.71%. The original dataset was randomlydivided into three different datasets: 40% for development, 30% for validation and the remaining 30% for testing.

For each one of the developed models a particular methodology was used to select the final variables, such as variable selection, stepwise, principal components and principal canonical components, among others. These selection methodologies are out of the scope of the paper thus they will not be explained.

METHODOLOGY

In order to control over-fitting and have comparable datasets, the first step was to split the entire population into three groups using the data partition tool; development, validation and testing.

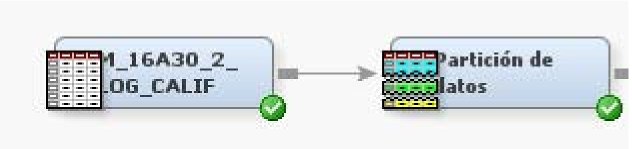

Figure 3. Data partition

Afterwards, four different credit risk models where developed using SAS Enterprise Miner: The Decision tree model, the Boosting tree model, the Logistic Regression model and the Neural Network model. As shown on figure 4.

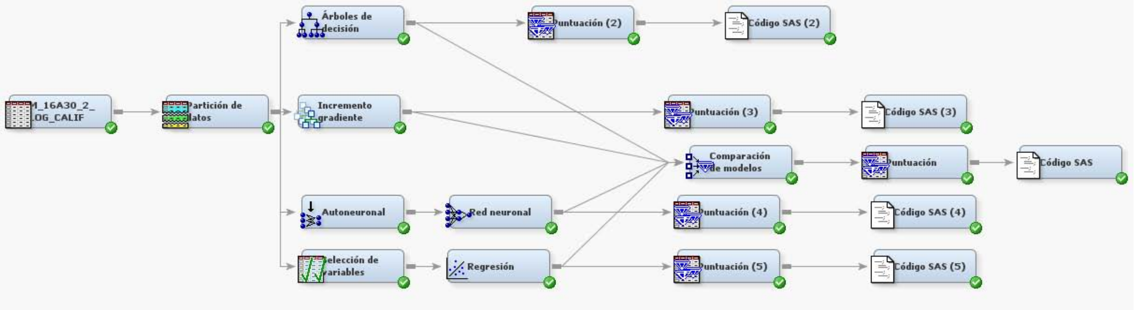

Figure 4.Enterprise miner project flow

Additionally, a MPL Neural Network model was developed using genetic algorithm optimization to select the best architecture (Correa Bahnsen & Gonzalez, 2011). This model was introduced in order to compare the boosting and ensemble models to a more powerful methodology than traditionally used.

Finally, the Ensemble model was developed combining 11 different Neural Network models, each one built using different features and structures in order to have many weak classifiers. All the models where blended into one powerful classifier using a Ridge Linear Regression model. (Bell, Koren, & Volinsky, 2007).

RESULTS

In order to compare the different models, the receiving operator characteristic Statistic (ROC) was used. The ROC curve shows the relationship between true positives and false positives through all scores ranges. This measure allows an accurate and unbiased model comparison taking into account the entire score range.

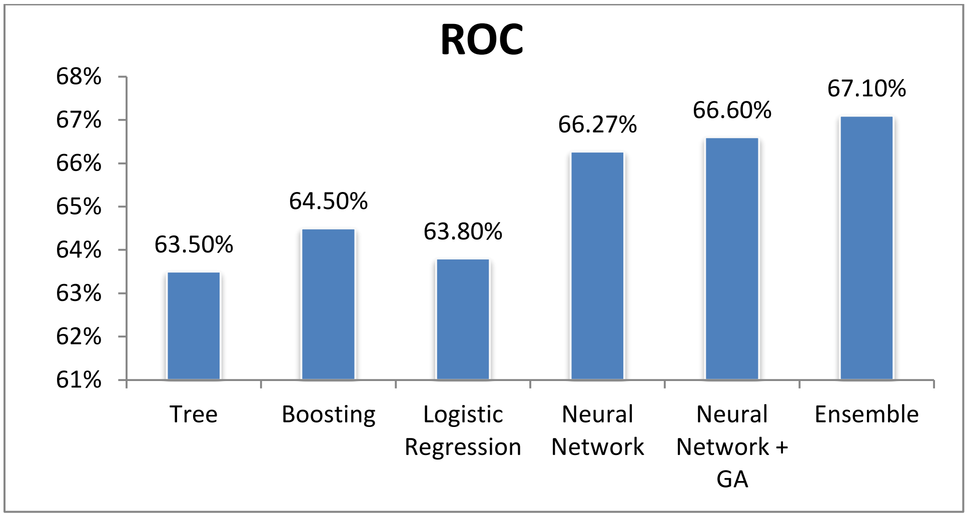

Figure 5. ROC statistic measure of the different methodologies

As can be seen in Fig 5., the best model based on the ROC statistic is the Ensemble model, the second best is the MLP Neural Network optimized with genetic algorithms, followed by the default Artificial Neural Network developed on SAS Enterprise Miner, and finally the Boosting Tree, the Logistic Regression and the Predictive Tree subsequently.

The Boosting Tree model which is a combination of different trees, improves the ROC of the single Predictive Tree model and exceeds the default Logistic Regression model built using the SAS Enterprise Miner. Nevertheless, the improvement is not enough to overcome the performance of the Neural Networks models and the Ensemble Model.

It is important to clarify that although methodologies like Boosting and Ensembles overcome the performance of traditional models, these methods require the development of several models and therefore they consume a greater amount of computational time and effort.

CONCLUSIONS

The main conclusion of this paper is that the model developed using the Ensemble methodology presents a greater predictive power and therefore is better than the traditional classifiers or even the newest methodologies such as the Artificial Neural Networks.

Given these results, it can be said that the methods of combining several weak classifiers generate more powerful models when developing credit risk scorecards, but they are more time consuming. Also given their complexity the relationship between variables is very hard to explain and understand, unlike the more simple traditional models.

REFERENCES

A. Correa, A. Gonzalez, C. Ladino. 2011. Genetic Algorithm Optimization for Selecting the Best Architecture of a Multi-Layer Perceptron Neural Network: A Credit Scoring Case. SAS Global Forum.

Bell, R. M., Koren, Y., & Volinsky, C. (2007). The BellKor solution to the Netflix Prize A factorization model. AT&T Labs – Research.

Correa Bahnsen, A., & Gonzalez, A. F. (2011). Evolutionary Algorithms for Selecting the Architecture of a MLP Neural Network: A Credit Scoring Case. 2011 IEEE 11th International Conference on Data Mining Workshops (págs. 725-732). Vancouver: IEEE.

Gao, F. H. (2010). On the Power of Ensemble: Supervised and Unsupervised Methods Reconciled. SIAM Inernational Conference in Data Mining.

Schapire, R. E. (1990). The Strength of Weak Learnability. Machine Learning, 97-227.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Alejandro Correa Bahnsen
Enterprise: Luxembourg University
City: Luxembourg, Luxembourg.
E-mail: al.bahnsen@gmail.com

Name: Andrés González
Enterprise: Banco Colpatria
City: Bogotá, Colombia
Phone: (+57)3103595239
E-mail: andrezfg@gmail.com

Name: Darwin Amezquita
Enterprise: Banco Colpatria
City: Bogotá, Colombia
Phone: (+57) 3013372763
E-mail: amezqud@colpatria.com