Analysis and comparison of machine learning classification models applied to credit approval

Jorge Alarcón Flores, Jiam Lopez Malca, Luiz Ruiz Saldarriaga and Christian Sarmiento Román
Maestría en Informática, Pontificia Universidad Católica del Perú, Lima, Perú
brian.alarcon@pucp.edu.pe, jiam.lopez@pucp.edu.pe, lruizs@pucp.pe,
cwsarmiento@pucp.edu.pe

Abstract

The credit granting decision is one of the most important process of all, and in whose accuracy, rests the good performance of several critical business KPI's such as loans level, credit recoveries level and non-performing loan ratios. In the last decade, the developing of certain technology such AI and machine learning has allowed this process automation. The present paper has its main goal, the analysis of credit granting predictions to collaborate with current knowledge in this issue, giving an objective explanation of the results and suggesting following researches to be developed in order to get better results in existing mathematical algorithms. As results of the experimentation determined that the best model was Gradient Boosting, with an accuracy of 83.71%.

1 Introduction

Credit institutions must establish efficient schemes for management and control of credit risk to which they are exposed in their business development,in accordance with their own risk profile and market segmentation, according to the markets characteristics in which they operate and the products they offer. To do this, it is necessary to adopt a research and analysis procedure, which can be seen reflected in a credit scoring or credit risk assessment, being considered as one of the most important processes of any financial institution, that can allow to have a the customer's first step in the credit admission process, which can give the financial institution an element of competitive advantage within the industrial sector. The problem therefore seeks to work a mathematical model based on machine learning that allows accurately predict the qualification of a client (good or bad), which provides us with an additional element for decision making, which will ultimately be the granting or non-granting of a credit.

2 Experimentation Design

2.1 Dataset Description

The original dataset consists of 1000 observations and 20 attributes, in addition to the classifier. The number of observations in the training sample is 750, and in the test, 250. The target has two values; namely 1 if the credit is approved or 2 if the credit is rejected. The dataset is clearly unbalanced, because there are 700 instances with classification value 1 (accepted credit) and 300 with classification 2 (credit rejected).

2.2 Data preparation strategy

One of the first problems to be faced is preparing the data for training and validation of the different learning models to use. The following points must be resolved:

Dataset balancing strategy: After the first experimentations it was found that there is a more efficient oversampling technique called SMOTE. The advantage of this technique is that it allows to generate new observations according to the distribution of the characteristics of the dataset. In this way the data set was balanced with 1400 instances (700 of each class).

Analysis and categorization of existing characteristics: Regarding to the numerical variables, we have worked on the conversion of only two variables. For the case of categorical variables, the one-hot encoding technique was applied to transform non-ordinal data into binary numerical data. This allows the multiplication of the characteristics, from 20 of the original dataset to 62 under this technique.

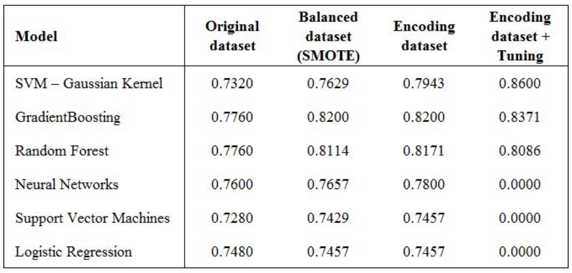
Algorithms to be used: The algorithms or models selected for the tests are as follows: Logistic Regression, Neural Networks, Support Vector Machine, Random Forest and Gradient Boosting.

Selection and justification of the quality measure: Two quality measures were selected in order to compare the validity and effectiveness of the different models to be used. The first quality indicator is the accuracy of each model. The second quality indicator chosen is the ROC curve.

Selection of the most relevant characteristics for learning the model: Once the best classification model was selected, the analysis of the importance of variables was performed, in order to find those characteristics of greater contribution in the prediction of credit risk.

3 Experimentation and Results

This research will serve as a baseline and reference point for the presentation of the results of the present study. The results and comparison of the mathematical models selected is showed in the table I.



Model	Original dataset	Balanced dataset (SMOTE)	Encoding dataset	Encoding dataset + Tuning
SVM – Gaussian Kernel	0.7320	0.7629	0.7943	0.8600
GradientBoosting	0.7760	0.8200	0.8200	0.8371
Random Forest	0.7760	0.8114	0.8171	0.8086
Neural Networks	0.7600	0.7657	0.7800	0.0000
Support Vector Machines	0.7280	0.7429	0.7457	0.0000
Logistic Regression	0.7480	0.7457	0.7457	0.0000

Table 1: Results comparisson of models applied.

4 Results Discusion

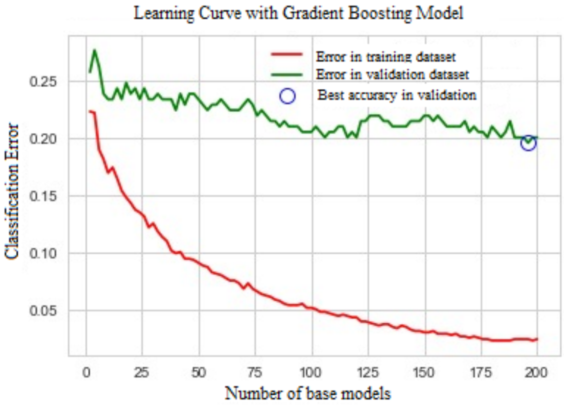


Figure 1: Learning curve with Gradient Boosting

In the experimentation, the model with more accuracy was the one of SVM with Gaussian Kernel. However, there are doubts to take the results of this algorithm because there is a limited number of observations compared to the number of characteristics of the training dataset. The reason for this decision is because it is believed that there is a risk of overfitting. The Gradient Boosting model is the one that had the best regularity during the experiments. In Figure 1 we observe the learning curve obtaining its smallest error (200 nodes).

5 Conclusions

The model that behaved in a more stable way with the different experimental scenarios was that of Gradient Boosting with an accuracy of 83.71%, and value ROC=0.834. The most important characteristics for the Gradient Boosting model were the balance account, duration of credit, credit amount, length of current employment and age. All these variables are coincidentally the ones that the experts of the financial sector.

6 References

Cheng-Lung Huang, Mu-Chen Chen and Chieh-Jen Wang, 2008. Credit scoring with a data mining approach based on support vector machines, *Expert Systems with Applications*.

Joao Bastos, 2008. Credit scoring with boosted decision trees, *MPRA Paper,* 27(1), pp. 262-273.

Josphat Kipchumba, 2012. Credit evaluation model using Naive Bayes classifier: A Case of a Kenyan Commercial Bank, *University of Nairobi*.

Nazeeh Ghatasheh, 2014. Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study, *International Journal of Advanced Science and Technology*.

Shashi Dahiya, S.S Handa y N.P. Singh, 2015. Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set, *Industrija* Vol.43, No.4, pp. 163-174.

Van Sang Ha, Ha Nam Nguyen and Duc Nhan Nguyen, 2016. A novel credit scoring prediction model based on Feature Selection approach and parallel random forest, *Indian Journal of Science* Vol. 9(20).