Ultimate Big Data Analytics, The Engine of Automated Credit Scoring Decision.

Jean de Dieu Murera Gisa

August 22, 2019

Abstract

The financial institutions play a significant role in people welfare and in country economy at large by offering the affordable financial products and services mostly loans and credit to both retail and non-retail (i.e. SMEs and Corporate companies) customers. Traditionally, the credit experts do a pretty good job of discriminating between the customer who is likely to default the issued credit or not using their business expertise, experience and common sense. They are referring on the customer's historical information such as age, gender, marital status, character and integrity, income, employment status, collateral size and so on to decide about the credit risk. During the qualitative analysis of such client information, a subjective evaluation of the credit risk is made and put in place by business experts. Although this classical credit scoring approach might miss-lead in credit worthiness scoring, it is still quite commonly used in almost all banks in Rwanda for every specific credit portfolios. For example in business project finance and individual loan. The intention of this work is to present the superiority of statistical machine learning approach in automatic credit worthiness scoring based on the huge customer information captured everyday and being stored in bank core system. In doing so, we employed several supervised thirteen machine learning algorithms on acquired loan data set from bank of Kigali and select the optimal model basing on the individual model predictive performance. Besides, the Mathews's Correlation Coefficient has been used as the model performance metric to measure the goodness of predictive performance of each learning machines algorithm employed. The adaboost model has been selected and presented as the superior optimal model for the credit worthiness scoring task with maximum correlation (agreement) between the actual and predicted credit customer categories.

1 Introduction

Granting credit and loan to the retail and corporate customers is the backbone of business performance in financial institutions and in other lender organizations such as mobile money, insurence, digital financial companies, landlords and government departments employ same technique. In doing so, those credit cards companies need to have the efficient systems to decide whom to grant credit or not. A credit scoring is data-driven analysis of a customer's credit files, to represent his/her credit worthiness. Credit scoring is a key risk assessment approach to evaluate and quantify which customers are likely to make the required payments on their debt obligations before reaching maturity. The credit scoring is used not only to determine whether credit should be approved or denied to an applicant, but also for setting the debt limit issued to customers.

Nowadays, the banks are gathered and collected plenty of customers information describing their defaulting behaviour together with a tons of business experience about a bank's credit products. All this data are nicely stored into large relational database or into electronic data warehouses of many of the banks. It is now good time to analyze efficiently and sufficiently those huge amount of customers's data from both sources and come up with a statistical machine learning based decision model that allows scoring future credit applications at Bank of Kigali, Rwanda. This credit scoring model will optimal deciding whether the borrower customer is worthy credit granted or not basing on his/her default probability.

To summarize, this big data-driven credit scoring solution is a sensitive key in credit risk management and will enable the bank to optimally manage, understand, and to accurately predict loaner(borrower) payment behaviour. This leads to sufficiently making a financial company more profitable and keep issuing loan and credit to continuously add value in real time.

2 Methodology

This section begins by first describing data set used and presenting its source. Then, explain the systematic approaches carried out to classify, and predict the credit worthiness category using different thirteen statistical learning machine algorithms. And later, the model performance metric has been presented.

2.1 Source and the Description of Data set

The loan data set is real-valued continuous multivariate set of data which has been sourced from the business department at Bank of Kigali PLC. The original data set presents $58,097$ observations (customers) and 28 attributes with the intention of evaluating the customer ability to pay back the issued loan or defaulting the loan offered.

Table 1: BK Loan data set Description.

Data Set Characteristics	Multivariate	Observational Instances	58,097
Attributes Characteristics	mixed	Number of Attributes	28
Date Donated	15/08/2019	Missing Values	1372

2.2 Experimental Systematic Approach Undertaken

2.2.1 Statistical Tool, Libraries and Packages

The R programming language were used as statistical tool to implement the thirteen machine learning algorithms to classify and predict the credit worthiness categories. R language is an open-source software for statistical analysis, computing, graphics representations and reporting (Fox and Andersen, 2005). R is freely available under the GNU, general public and pre-compiled binary versions are provided for various operating systems like Unix, Windows and Mac.

The statistical R libraries and packages relevant for this paper are differ from their assigned specific tasks and functions. Here below are those R libraries utilized to implement the thirteen statistical learning machines to detect the credit worthiness classes.

- **The package caret:**[1] It stands for Classification and Regression Training (Kuhn *et al.*, 2008). It accommodates and integrates more than two hundreds and nearly every major machine learning algorithms available in R language. This library has been widely used in this paper since it loads all thirteen algorithms used for the experiment.

- **The tidyverse package:**[2] The tidyverse is an opinionated collection of R packages designed for data manipulation, exploration and visualization. All hosted packages share an underlying design philosophy, grammar, and data structures (Wickham).

To implement each and every supervised statistical machine learning algorithm considered to detect, classify and predict the prevalence classes of issued credit worthiness, we have used the function *train()* supported by library *caret* and mentioned the recognized name of the algorithm in argument *method()*.

[1] http://CRAN.R-project.org/package=caret

[2] https://www.tidyverse.org/

2.2.2 Experiment Set Up and Data Pre-Processing

The employed BK Loan data set has been well prepared, pre-processed and splitted into training and independent test data sets before heading to the models estimation. The predictive performance of each learning machine model has been estimated by using the algorithmic syntaxes supported by *caret* library. It has several functions that attempt to streamline the predictive model building and evaluation processes, as well as feature selection, positioning and scaling, data splitting, pre-processing, and variable importance estimation. The fundamental and principal tool in such library is the ***train()*** function which has been used to train each model performance on splitted train set of fixed sample dimension of $(46478, 12)$. And it has been used to select the optimal model across the hyper-parameter tuned parameters through utilization of 5-fold cross validation repeatedly three times as re-sampling method. Therefore, the models have been validated on test data set of sample dimension of $(11618, 12)$ to predict whether customer is likely to pay back the issued loan or not. The author can't leave behind the fact of setting pseudo-random number generator for the avoidance of result variability in order to obtain the nearly trusted experimental results.

More formally, the *train()* Pseudocode is used for simplicity and easy understanding of each model algorithm infrastructure and program skeleton to implement all considered statistical machine learning models which were compiled without errors.

Algorithm 1 :Model Training and Hyper-parameter Tuning

```
Define sets of model parameter to evaluate
for each parameter set do
    for each re-sampling iteration do
        Hold-out specific samples
        [Optional] Pre-process the data
        Fit the model on the remainder
        Predict the hold-out samples
    end for
    Calculate the average performance across hold-out predictions
end for
Determine the optimal parameter set
Fit the final model to all the training data using the optimal parameter set
```

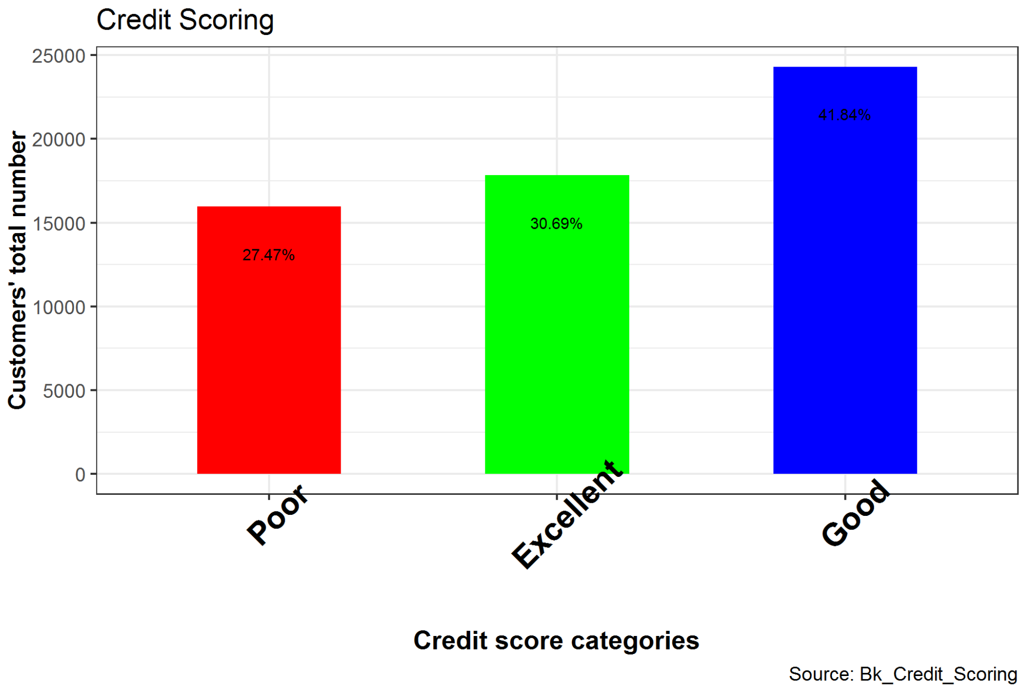
2.3 Optimal Model Selection and Performance Evaluation

The predictive built models have been evaluated to quite measure the model's quality and efficacy performance by using the Mathew's correlation coefficient (mcc). It indicates the agreement between the actual(observed) and predicted prevalence classes in pattern recognition and classification tasks. The predictive model performance were recorded and presented in table format. Then a model with maximum correlation coefficient has been selected and chosen as an optimal model due to the fact that it represents the perfect prediction relative to others.

3 Result and Discussion

This section represents the graphical analysis of acquired data and experimental results obtained from training thirteen classification supervised learning machine algorithms on such data set in order to significantly detect and predict the credit worthiness at Bank of Kigali. The quality of the model has been presented in terms of its Mathew's correlation coefficient.

Firstly, let us understand the loan data used through the different graphs and plots.

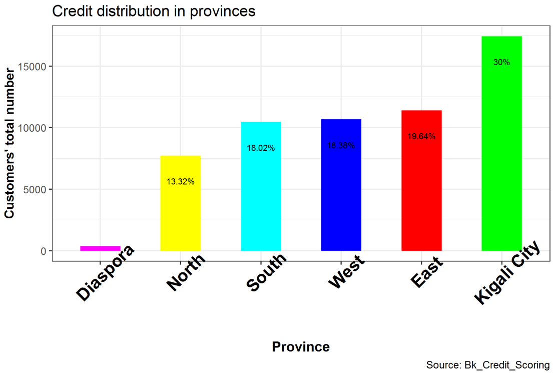


Credit Scoring

Source: Bk_Credit_Scoring

Figure 1: The prevalence customer credit worthiness. (wikipedia.org, 2015)

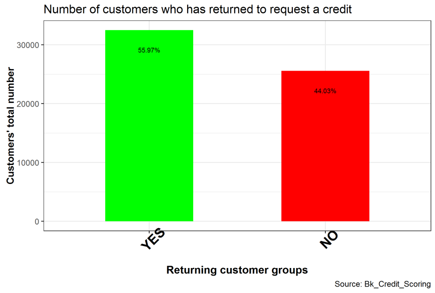
The above plot is representing the categories of customers at Bank of Kigali basing on their loan defaulting behavior.



Credit distribution in provinces

Source: Bk_Credit_Scoring

(a) The provincial loaners



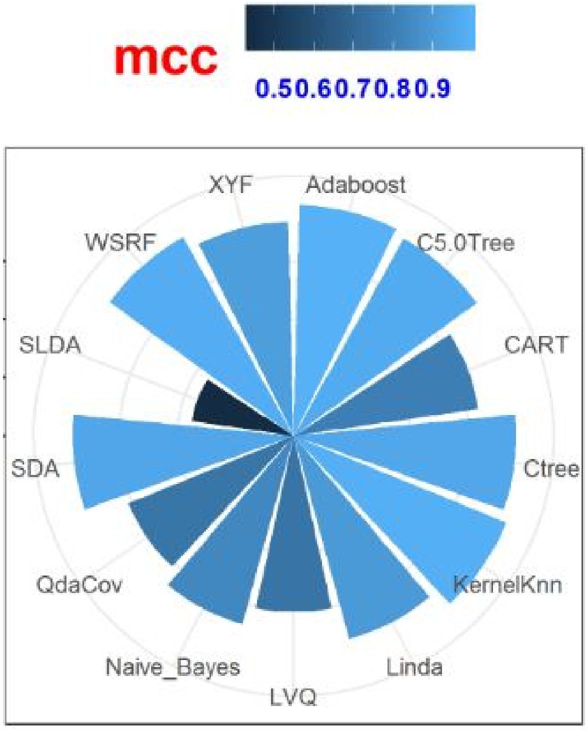
Number of customers who has returned to request a credit

Source: Bk_Credit_Scoring

(b) Old and New loaners at BK

The above Plot (2a) is showing the exact number of customers who were offered credit in different province and those staying abroad. At the other hand, the Plot (3b) indicates whether the customer issued credit is new or the old one.

Secondary, let us present the predictive model comparative results, later the optimal model basing on its maximum agreement between observable and predicted credit worthiness classes. The Table (3b) and Figure (3a) are presenting the comparative model performance with Mathews correlation coefficients.



Source: BankNotes_Authentication Data Set

(a) Model Performance Pie Chart.

(b) Model Performance by mcc metric

SN	ML Model	mcc
1	SLDA	0.4417760
2	LVQ	0.7562242
3	QdaCov	0.7668727
4	CART	0.8003889
5	Naive_Bayes	0.8375062
6	Linda	0.9044117
7	XYF	0.9204091
8	SDA	0.9498330
9	Ctree	0.9556730
10	C5.0Tree	0.9709653
11	WSRF	0.9778811
12	KernelKnn	0.9853554
13	Adaboost	0.9926452

Figure 3: Classifiers Comparative performance on BK Loan data.

The comparison of experimental measurement presented from Figure (3) gives the right decision to select the Adaptive Boosting meta algorithm (adaboost) as the superior optimal model relative to others, all trained on 80% training and 20% testing of Bk Loan data set. Since it is used in conjunction with other types of learning algorithms to improve performance, it won the detection, classification and prediction of credit worthiness classes at Bank of Kigali. In addition, it is often referred to as the best classifier due to its achievement of optimal performance of the 99.26% agreement between the actual and predicted credit worthiness categories.

4 Conclusion

This study considered 13 different statistical machine learning algorithms which have been carefully selected from the set of supervised parametric and/or non-parametric models. To select the best optimal model, the *caret* R package has been widely used to train and tune our models. The adaboost model takes the lead in accuracy relative to others with the maximum agreement (Mathews' correlation coefficient) between the actual and predicted banknotes prevailed classes. And finally, it is advised to use this statistical machine learning algorithm to automate the credit scoring decision at Bank of Kigali in all credit portfolios.

References

J. Fox and R. Andersen, *Department of Sociology, McMaster University*, 2005, 2–4.

M. Kuhn *et al.*, *Journal of statistical software*, 2008, **28**, 1–26.

H. Wickham, *URL https://CRAN. R-project. org/package= tidyverse. R package version*, **1**, 51.

wikipedia.org, *Towards Data Science*, 2015, 2–4.