Loan Default Prediction Using Supervised Machine Learning Algorithms

Cohort 2 Group 1

Joyce Njeri, Christine Wasike, Alice Fatmata,
Ayo Oluwapamilerin, Gilbert Sibomana

April 30, 2021

Abstract

African financial deepening is beset by a high rate of loan defaults which encourages banks to hold liquid assets instead of lending. Financial institutions have large amounts of data on their borrowers, which can be used to predict the probability of borrowers defaulting their loan or not. With respect to recent development in the field of machine learning, there has been an interest in investigating if machine learning techniques can perform better quantification of the risk. The purpose of this paper is to analyze individual loan defaults in Africa and examine which method from a chosen set of machine learning techniques exhibits the best performance in default prediction with regards to chosen model evaluation parameters. The investigated techniques were Logistic Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, eXtreme Gradient Boosting (XGBoost), and Gradient Boosting Classifier among others. The study recommended the use of the XGBoost Classifier model in conjunction with a supervised machine learning approach in loan default prediction in financial institutions.

Keywords: Loan defaults; loan default prediction; Nigeria; financial development; Africa;

1 Introduction

1.1 Background

Sub-Saharan Africa remains one of the most financially under-developed regions in the world [1]. Banks in Africa complain that there is a lack of creditworthy borrowers while at the same time households and firms find finance as a major constraint in their activities. Recent research has shown that banks are deterred from lending by a very high rate of loan defaults [2]. Understanding the determinants of loan default rates in Africa seems, therefore, to hold the key to overcoming the obstacles to financial development in Africa. The main importance of financial institutions, particularly banks, are to safeguard the money kept by their clients and make it accessible when need arises. They also advance loans to their customers [3]. Accurate prediction of default risk in lending has been a crucial theme for banks and other lenders over a century to avoid huge losses that result from wrong decisions [4]. Modern days availability of large datasets and open source data, together with advances in computational and algorithmic data analytics techniques, have renewed interest in this risk prediction tasks. Customers in default means that they did not meet their contractual obligations and potentially might not be able to repay their loans, thus the interest of acquiring a model that can predict defaulted customers [5].

1.2 Purpose

The objective of this assignment is to investigate which method from a chosen set of machine learning techniques performs the best default prediction.

1.3 Scope

The scope of this paper is to implement and investigate how different supervised binary classification methods impact default prediction. The model evaluation techniques used in this project are limited to precision, F-score and accuracy score. The classifiers that will be implemented and studied are: Logistic

Regression, Decision Tree, Random Forest, Gaussian Naive Bayes, XGBoost, and Gradient Boosting Classifier among others. The project will be performed using internal data of SuperLender customers, a local digital lending company in Nigeria. The results presented in this thesis may therefore be biased towards the profile of SuperLender clients, specifically their location and behavioral factors.

2 Methodology

2.1 Formulation of a Binary Classification Problem

Binary classification refers to the case when the input to a model is classified to belong to one of two chosen categories. In this project, customers belong either to the non-default category or to the default category. The categories can therefore be modeled as a binary random variable:

$$Y \in \{0, 1\}$$

, where 0 is defined as non-default, while 1 corresponds to default. The random variable Y_{i} is the target variable and will take the value of y_{i}, where it corresponds to the ith observation in the data set. The rest of the information about the customers, such as loan amounts and payments in arrears can be modeled as the input variables. These variables are both real numbers and categories and are often referred to as *features* or *predictors*. With this setup, it makes it possible to fit a supervised machine learning model that relates the response to the features, with the objective of accurately predicting the response for future observations [6]. The main characteristic of supervised machine learning is that the target variable is *known* and therefore an inference between the target variable and the predictors can be made.

2.2 Feature Selection Methods

In the data given by SuperLender, features are presented as both continuous and categorical variables. Thus, in order to understand how features correlate with each other and the response variable, implemented methods for feature selection should process both continuous and categorical variables simultaneously. That is why we used the following methods: Feature selection with Kendall's Tau Coefficient Analysis and Recursive Feature Elimination (RFE).

2.2.1 Correlation Analysis

The Kendall's Tau rank correlation coefficient measures the ordinal association between two measured variables, thus, it is a univariate method of correlation analysis [7]. The Kendall's Tau rank correlation is a non-parametric test which means that it does not rely on any assumptions of distributions between the analyzed variables [8]. In a statistical hypothesis test, the null hypothesis implies an independence of X and Y and for large data sets, the distribution of the test can be approximated by a normal distribution with mean zero and variance [9].

2.2.2 Recursive Feature Elimination

RFE is a multivariate method of variable selection [10], which performs a backward selection of predictors [11]. The algorithm starts with computing an importance score for each predictor in the whole data set. Let S be a sequence of the number of variables to include in the model. For each iteration, the number of S_{i} predictors which have been top-ranked are retained in the model [12]. Further, the importance scores are computed again and the performance is reassessed. The tuning parameter for RFE is the subset size and the subset of S_{i} predictors with the highest importance score which is then used for fitting the final model. Thus, the performance criteria is optimized by the subset size with regards to the performed importance ranking.

2.3 Handling Sample Imbalance

In the data provided, a heavy class imbalance exhibits. An imbalanced data set contains observations where the classes of the response variable are not approximately equally represented. The imbalance causes a problem when training machine learning algorithms since one of the categories is almost absent, hence poor predictions of new observations of the minority class are expected. In order to increase the performance of the algorithms there are different sampling techniques that can be used. One of

them is called Synthetic Minority Oversampling Technique (SMOTE). SMOTE is a sampling algorithm that creates synthetic data points for the minority class rather than sampling existing observations with replacement [14]. The method is applied on the continuous parameters for each sample in the minority class.

2.4 Model Evaluation Techniques

2.4.1 Confusion Matrix

One common way to evaluate the performance of a model with binary responses is to use a confusion matrix. The observed cases of default are defined as positives and non-default as negatives [13]. From a confusion matrix there are certain metrics that can be taken into consideration. The most common metric is accuracy which is defined as the fraction of the total number of correct classifications and the total number of observations. In terms of business sense, the aim is to achieve a trade-off between losing money on non-performing customers and opportunity cost caused by declining of a potentially performing customer. Thus, there is a high relevance to analyze how sensitivity and precision are affected by various methods applied, as sensitivity is a measure of how many defaulted customers are captured by the model, while precision relates to the potential opportunity cost. Since sensitivity and precision are of equal importance in this project, a trade-off between these metrics is considered. The F-score is the weighted harmonic average of precision and sensitivity [14].

2.4.2 Area Under the Receiver Operator Characteristic Curve

Another way to evaluate results from the models is to analyze the Receiver Operator Characteristic (ROC) curve and its Area Under the Curve (AUC). The ROC curve uses the false positive fraction in order to describe the trade-offs between sensitivity and (1-specificity). A ROC curve can also be summarized by AUC score, which represents an index for ROC curve [6].

2.5 Cross-validation

In order to prevent using the same information in the training phase and the evaluation phase of models, which makes the results less reliable, the data is divided into training set, validation set and test set [15]. The training set and validation set are used for finding the best model and the test set is only used for calculating the prediction performance of the best model. The test data will therefore be held out until the best model is obtained [14]. Choosing the best model from a set of models can be done by a method called K-fold Cross-validation (CV).

3 Data Preprocessing and Variable Selection

The design adopts both quantitative and qualitative approaches or methods in a single study [16].

3.1 Data Description

The data that was used for this study was obtained from SuperLender, a local digital lending company in Nigeria. A random sample of 18,183 loan applicants whose loans had been approved by 18 banks over the period 2016 - 2017 was obtained. Information derived from these variables were what type of accounts a customer possessed and how these particular accounts behaved over a certain time span. The time span of data was merged with a default flag indicating if the customer was in default or not.

3.2 Data Preprocessing

The first step was to filter the data by cleaning it through python's library pandas. Missing values were treated by doing a complete case analysis. The data was then coded for easy analysis. The coding involved identification of non-performing loans or a loan default with a value '0' for 'Bad', and a performing loan with a value '1' for 'Good'. Equivalent number of dummy variables were created for the purposes of coding independent variables. The clean data was then used for analysis and generation of descriptive statistics and also fit the models.

3.3 Variable Selection by Correlation Analysis

In order to reduce the number of variables and treat multicollinearity, correlation analysis for the features against response and features against features was performed. Some variables have strong correlation, which should be treated. The next step in this case would be to choose a threshold for correlation allowed in the model. The threshold was chosen to 0.7, because pairwise correlation higher than 0.7 leads to unstable estimates and multicollinearity [17]. After the threshold was decided, the variable which correlates the most with the response variable was kept.

3.4 Variable Selection by RFE

For the sake of getting a multivariate based variable selection, RFE was implemented. The variable selection by RFE was performed only on the behavioral variables with a Random Forest classifier with F-score as a scoring evaluation.

4 Results

4.1 Performance of the Methods

As mentioned in section 2.3, performance of the methods is evaluated by sensitivity, precision, F-score and ROC-score with the F-score as the primary metric.

In Table 4.1, it can be seen that the highest obtained F-score was for XGBoost. With regards to precision, sensitivity and accuracy, LR generated the best precision, while XGBoost had the best sensitivity, roc-score and accuracy.

Table 4.1: Performance of Methods with regards to chosen performance measurements

Variable Selection	Number of - Features	Model	Precision	Sensitivity	F-score	ROC-score	Accuracy
Correlation Analysis	30	Logistic Regression	0.985	0.567	0.888	0.687	0.801
	30	Decision Tree Classifier	0.768	0.298	0.798	0.561	0.691
	30	Random Forest Classifier	0.886	0.376	0.858	0.605	0.762
	30	Gaussian Naive Bayes	0.890	0.417	0.862	0.626	0.773
	30	XGBoost Classifier	0.970	0.606	**0.891**	0.715	0.810
	30	Gradient Boosting Classifier	0.959	0.534	0.885	0.674	0.803
RFE	7	XGBoost Classifier	0.746	0.705	0.817	0.804	0.802
	18	XGBoost Classifier	0.735	0.689	0.809	0.794	0.792
	25	XGBoost Classifier	0.733	0.688	0.807	0.792	0.789

Further, in Table 4.1, it is shown that in terms of variable selection methods, correlation analysis with Kendall's Tau performed better than RFE.

4.2 Results of the Best Performing Method

The results showed that XGBoost Classifier obtained the best result with respect to the chosen model evaluation metric. The model had an accuracy of 0.81. It showed a precision score of 0.97, and accuracy of 0.81.

5 Discussion

From section 4, it can be seen that the overall best performance was obtained with the machine learning technique XGBoost. In this project, F-score was chosen to be the main indicator of how well the model

performed. Thus, if the main performance indicator is to be changed (for example sensitivity, accuracy or precision), another conclusion can be made and further analysis should be performed.

5.1 Findings

Generally, results indicated that tree-based models showed a good performance on average. This is aligned with results from another study, which had the same conclusion when comparing performance between Artificial Neural Network (ANN) and tree-based methods [18]. However, in order to test if this relation exhibits generally, more tests should be done and different sets of classifiers should be studied.

5.1.1 Impact of SMOTE

On average, SMOTE did not have a significant positive impact on the F-score, but it had an impact on sensitivity and precision. It can be noted that implementation of SMOTE led to an increase in sensitivity and a decrease in precision.

5.1.2 Impact of Variable Selection Methods

After analyzing the results, the conclusion can be made that correlation analysis with Kendall's Tau performed better than RFE. That can be partially explained by the nature of variable selection methods. Kendall's Tau provides a pair-wise analysis of variables, while RFE is a multivariate selection method, which allows the analysis of a set of features simultaneously. In this case, it can be concluded that for this type of project, the pair-wise selection method is more suitable than the multivariate one.

Results indicated also that the number of features included in the model had an impact on the model performance. When using RFE as a variable selection method, there is a direct relation between an decrease in F-score and an increase in the number of variables used in the model. This is shown in Table 4.1. On the other hand, the increase is marginal. Considering the trade-off between the complexity of the model and obtaining the highest F-score, one solution in this case could be to define the highest number of variables allowed in the model or a threshold of the lowest F-score.

5.2 Conclusion

The overall results showed that XGBoost executed by correlation analysis with Kendall's Tau showed the best performance with regards to F-score. It can be concluded that SMOTE did not enhance the performance of the models remarkably in terms of F-score. When SMOTE was applied, sensitivity increased and precision decreased. On average, it was also concluded that the increased number of variables used in the models chosen by RFE had a direct relation with a decrease in F-score. However, this decrease was marginal and that should be taken into consideration. Further, it could be also concluded that tree-based methods showed on average a relatively good performance.

5.3 Recommendations

The potential future work for this project will be a further development of the model by deepening analysis on variables used in the models as well as creating new variables in order to make better predictions. Data available for the scope of this thesis has constraints in terms of geographical breadth of Nigeria's clients. It should be considered that the behaviour of customers influences the results of this research. It means that the behaviour of clients outside Nigeria may or may not follow the same pattern and therefore one should make additional analysis and obtain a geographically-broader data set if the objective is to have a model unbiased of the geographical location.

Further, it would be interesting to apply other variable selection methods. An alternative for dimensionality reduction could be Principal Component Analysis (PCA). It would also be interesting to make a study concerning what metrics are the most relevant for this type of the problem. In this project the main metric that all evaluations were analyzed by was F-score, because the aim was to achieve a trade-off between the sensitivity and the precision. If a deeper analysis could be performed regarding the most relevant metric for this type of problem, then potentially a weight function could be implemented if one of the metrics explored turned out to be of more importance.

References

[1] P. Honohan and T. Beck, "Making Finance Work for Africa", 2007. Available: https://www.researchgate.net/Making$_{F}inance_{W}ork_{f}or_{A}frica$.[$Accessed 24 April 2021$].

[2] S. Iftikhar, "The impact of financial reforms on bank's interest margins", Journal of Financial Economic Policy, vol. 8, no. 1, pp. 120-138, 2016. Available: https://www.researchgate.net/deref/http

[3] Divino JA, Lima ES, Orrillo J. Interest rates and default in unsecured loan markets. Quantitative Finance. 2013;13(12):1925-1934.

[4] Lahsana A, Anion R, Wah T. Credit scoring models using soft computing methods: A survey. International Arab Journal of Information Technology. 2010;7(2):115-123.

[5] L. C. Thomas, E. N. C. Tong, and C. Mues. "Mixture cure models in credit scoring: If and when borrowers default". In: European Journal of Operational Research 218 (2012), pp. 132–139.

[6] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. An Introduction to Statistical Learning. eng. Springer Series in Statistics. New York, NY: Springer New York, 2013, p. 26. isbn: 978-1-4614-7138-7.

[7] Xavier Benoit Gust Lucile D'journo. The use of correlation functions in thoracic surgery research. https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC4387406/. [Accessed 24 April 2021].

[8] Jean D. Gibbons. Nonparametric Measures of Association. Thousand Oaks, California: SAGE Publications, Inc., url: https://methods.sagepub.com/ book/nonparametric-measures-of-association.

[9] R.B. Nelsen. Kendall tau metric. eng. Encyclopedia of Mathematics. 2001. isbn: 978-1-55608-010-4.

[10] Recursive Feature Elimination (RFE). https://www.brainvoyager.com/bv/doc/UsersGuide/MVPA/RecursiveFeatureElimination.html. [Accessed 27 April 2021].

[11] Kjell Kuhn Max Johnson. Feature Engineering and Selection: A Practical Approach for Predictive Models. eng. Feature Engineering and Selection: A Practical Approach for Predictive Models. 2019. isbn: 9781138079229.

[12] Feature Analysis Visualizer. Recursive Feature Elimination. https://www.scikit-yb.org/en/latest/api/features/rfecv.html. [Accessed 27 April 2021].

[13] James Finance. "Machine Learning in Credit Risk Modeling: Efficiency should not come at the expense of Explainability". In: (July 2017).

[14] Cyril Goutte and Eric Gaussier. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation". In: European Conference on Information Retrieval. Springer. 2005, pp. 345–359.

[15] Stuart J. Russell and Peter Norvig. Artificial Intelligence: A Modern Approach. eng. Upper Saddle River, New Jersey 07458: Pearson Education, Inc., 2010.

[16] Tashakkori A, Teddie C. (Eds). The handbook of mixed methods in social and behavioural research, sage. Thousand Oaks, CA; 2003.

[17] Michael Bowles. Machine Learning in Python: Essential Techniques for Predictive Analysis. eng. 2015. isbn: 9781118961742.

[18] Peter Martey Addo, Dominique Guegan, and Bertrand Hassani. "Credit Risk Analysis Using Machine and Deep Learning Models". In: Risks 6 (2018).