Paper 010 - 2022

**Loan default predictions – A case study for Nigeria**

Wan Xinyu, Wang Yizhi, Yang Jingyuan, Singapore Management University

**ABSTRACT**

The world we are in right now is certainly experiencing a time of uncertainty. With Russian continuing their ‘Special military operation’ in Ukraine and the COVID-19 virus evolving into their new variants, many parts of the world continue to experience economic disruption due to these crises. (REINHART, C.2022) While economic recovery seems to be on the way, the effect is uneven. Recovery in terms of per capita income increase is far lower in low-income countries than advanced economy in 2021(IMF,2022), majority have yet to recover from the sharp decline in income from the crises mentioned earlier. Thus, in this paper, we will be studying the market of private debt, an area that has relatively less focus compared to public debts. Our paper aims to find a way to identify potential borrowers who may default their loan payment by using three supervised machine learning algorithms (Decision Tree, Bootstrap Forest, and Boosted Tree model) to aid us with the prediction. The goal is to assist financial institutions in ensuring that loans will continue to be given out despite the crises (avoiding a credit crunch!) but also that it will be only given out to those who have low default chance to avoid it from turning into bad debts. Since a large number of defaults in banks may lead to the banks seeking for bail outs from the government, our secondary goal here then is also to prevent the insolvency problem being transferred to the public sector when many governments are already facing budget strains due to the ongoing crises.

**INTRODUCTION**

Financial institutes such as banks make money in multiple ways. However, at the core of their business they are considered as lenders. (Corporate Finance Institute 2022) Therefore, loan collection is an essential job for banks because uncollectable loans would turn into non-performing loans or what is better known as bad debt after it exceeds the specified terms. Therefore, for banks to minimize losses on bad debts, there is a need for banks to examine the borrower’s profile stringently before deciding on disbursing the loan. Our task in this project is to develop a predictive model which will help banks in Nigeria to examine the borrower’s ability to return loans before the it is disbursed in the first place. According to CEIC data, Nigeria Non-Performing Loans was reported at 3.315 USD bn in Mar 2021. (CEIC Data. n.d.)

**OBJECTIVE**

In this paper, we will analyze the dataset containing the characteristics of the borrowers. We will then develop a predictive model which will help banks in Nigeria to examine the borrower’s ability to return loans before the it is disbursed. We will explore and evaluate the Decision Tree, Bootstrap Forest, and Boosted Tree model to identify the most suitable model that can be used as prediction. We will also list the variables that are important to identify the potential bad debtors.

**LITERATURE REVIEW**

We reviewed studies done by others on similar topic of loan default, but the focus has been mainly on building a single model to fit the prediction. Tree-Based methods for loan approval Vaidya, A. 2017) focused their research on using decision trees for their prediction while the other (Alaradi, M & Hilal, S. 2020) focused on using logistic regression. Both articles focused on a single model approach in their forecast. Thus, in this paper we want to build on these standalone model predictions and create multiple models (Decision Tree, Bootstrap Forest, and Boosted Tree) using the same dataset to determine which model will give us the optimum result.

**DATA PREPARATION WORKFLOW(General)**

Text

Description automatically generated

**DATASETS**

The datasets used in this paper are provided by Data Science Nigeria. It contains the following:

* Training set and test set (each containing the following):
  + Demographics
  + Performance
  + Previous Loan

**DATA PROCESSING cum EXPLORATION**

Processing and modeling were performed in the training dataset. The final model is then extracted and apply to the test set. In the training set, demographic and previous loan data were joined to form a new data table through common variable ’CustomerID’. New variables are derived and those original variables which either have too many missing values or those that we consider as insignificant for our analysis were subsequently removed. Below are extracts of the steps that we took in generating new variables. The table later also shows the final variables that were used in our model. All data wrangling process were done on JMP PRO 16.

**1. Create new columns named *termduration*, *overdue or not*, *exceeded days.***

* The formula for *termduration:*
* The formula for *overdue or not*:

“1” refers to not overdue, “0” refers to overdue

* Formula for *exceeded days*:

**2. Group previous loan information by customerid**

Summary the sum and the mean of exceeded days, the sum of overdue or not (the numbers of loans that settled in time).

Graphical user interface, text, application

Description automatically generatedCalendar

Description automatically generated

As we have the total number of loans and the number of loans that settled in time, then create a new variable: good percent.

**3. Join data tables with primary key customerid.**

The predictors for our model:



**Trainmodel datatable (JMP)**

Generally, insignificant/unused variables, and variables with significant proportion of missing values should be removed. The ‘latitude’ and ‘longitude’ data are important geographical indicators. However, it is unsuitable in our analysis considering that we will be generating synthetic data later and the original data indicates location out of Nigeria as well. Variable ‘referred by’ with 94% missing value and ‘bank\_branch\_client’ with 99% missing value were excluded in the final dataset. Variables ‘level of education of client’ which identifies the education level has a missing value of 83%. However, we have decided to keep this variable despite the large number of missing value as we believe that it is a crucial determinant of a person’s ability to repay debts. Therefore, we retained it in the dataset and recoded as ‘Unknown’ in our dataset.

**MODEL PLANNING/BUILDING**

The original dataset provided has an unbalance number of good bad flag indicators where the ‘good’ exceed the ‘bad’ by 300%. Since our prediction result is binary (“bad” or “good”), the imbalance classification SMOTE plus Tomek sampling methods were applied to our imbalanced data.

Chart, pie chart

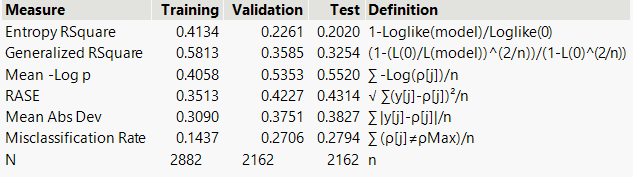
Description automatically generated

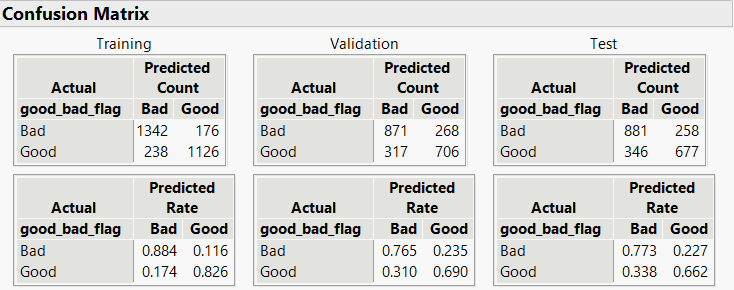
**BOOTSTRAP FOREST**

**INTRODUCTION**

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. [[1]](#footnote-2)

**OVERALL STATISTICS**





The report reveals that the Misclassification Rate of Training, Validation and Test data are 0.1437, 0.2706 and 0.2794 respectively. As observed from the confusion matrix, the true positive and true negative are above 0.7 and 0.6 respectively. This means that this model can capture borrowers who are forecasted to pay (and actually paid) and forecasted to default (and actually default) with relatively high accuracy while reducing the chance of borrowers who will default being classified as ‘credible borrowers’ while those who will pay as ‘risky borrowers.

**COLUMN CONTRIBUTIONS**

图表, 表格

描述已自动生成

The column contribution reveals that all the predictors are used to fit the model. **Mean (exceeded days)** was used most often, 1434 splits out of the 4663 splits.

**BOOSTED TREE**

**INTRODUCTION**

Boosted Regression Tree (BRT) models are a combination of two techniques: decision tree algorithms and boosting methods. Like Random Forest models, BRTs repeatedly fit many decision trees to improve the accuracy of the model.[[2]](#footnote-3)

**OVERALL STATISTICS**

文本

描述已自动生成

表格

描述已自动生成

The report reveals that the Misclassification Rate of Training, Validation and Test data are 0.2113, 0.2784 and 0.2798 respectively. As observed from the confusion matrix, the true positive and true negative are both above 0.7. This means that this model can also capture borrowers who are forecasted to pay (and actually pay) and forecasted to default (and actually default) with relatively high accuracy while reducing the chance of borrowers who will default being classified as ‘credible borrowers’ while those who will pay as ‘risky borrowers (with similar accuracy to Bootstrap Forest).

**COLUMN CONTRIBUTIONS**

表格

描述已自动生成

The report reveals that 8 of the 9 predictors are used to fit the model, **termdays** was used most often, 180 splits out of the 1095 splits.

**DECISION TREE**

**INTRODUCTION**

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.[[3]](#footnote-4)

**OVERALL STATISTICS**

文本

描述已自动生成

表格

描述已自动生成

The report reveals that the Misclassification Rate of Training, Validation and Test data are 0.2793, 0.3016 and 0.3215 respectively. As observed from the confusion matrix, the true positive and true negative are both above 0.6. This means that this model can capture borrowers who are forecasted to pay (and actually pay) and forecasted to default (and actually default) with relatively moderate accuracy (comparatively to the other 2 models).

**COLUMN CONTRIBUTIONS**

图表

描述已自动生成

The report reveals that 8 of the 9 predictors are used to fit the model, **good percent** was used most often, 6 splits out of the 34 splits.

**DECISION TREE**

图形用户界面, Word

描述已自动生成

This figure shows the overview of the decision tree of our model.

**THE ROC CURVE**

The ROC graph is a curve that reflects the relationship between sensitivity and specificity. The abscissa X-axis is 1 – specificity, also known as false positive rate, the closer the X-axis is to zero, the higher the accuracy; The ordinate Y axis is called sensitivity, also known as the true positive rate (sensitivity), and the larger the Y axis, the better the accuracy.

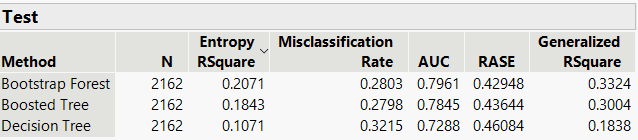
According to the position of the curve, the whole graph is divided into two parts. The area under the curve is called AUC (Area Under Curve), it is used to indicate the prediction accuracy. The higher the AUC value, that is, the larger the area under the curve, the higher the prediction accuracy. The second aspect of the curve is the area above the curve. The closer the curve is to the upper left corner (the smaller X and the larger Y), the higher the prediction accuracy.

Chart

Description automatically generated

The figure above shows the ROC curve for the 3 models that we have built: Decision Tree, Bootstrap Forest, and Boosted Tree. We notice that the ROC curve for Bootstrap Forest has the largest AUC, which means that the result is the most accurate, among the 2 models, in distinguishing the positive and negative classes

To further our analysis, JMP also have specific statistics about the performance of the algorithms as shown below.



Graphical user interface, application

Description automatically generated with medium confidence

Our focus on the above statistics table will be on the ROC and misclassification rate and the F1 score. We have explained about AUC earlier, the misclassification rate shows how often the confusion matrix is incorrect in predicting the actual positive and negative outputs and the F1 score is the harmonic mean of the precision and recall and it is a performance measure of test accuracy.

Since Bootstrap Forest has the lowest misclassification rate, highest TP and TN and the second highest F1 score, this paper will use consider it as the optimum model to apply for analysis. Using these 3 indictors, we rank our models as the following starting with the best:

**Bootstrap forest > Boosted Tree > Decision Tree**

**CONCLUSION AND FUTURE WORK**

**CONCLUSION**

In this paper, we explored the three types of models to be used for the prediction of loan repayment in Nigeria. Our findings show that the Bootstrap Forest gives the best result given that it produces the lowest misclassification rate of 0.2766 and highest AUC value of 0.7948.

However, we would want to caution readers that our model involved the use of imbalance data classification add-in for JMP where we used the combined SMOTE and Tomek sampling technique. The reason for applying the add-in is because the Good or Bad flag given in the Train dataset has imbalanced binary responses. Data for Good flag is approximately 300% more than the data for Bad flag.

Our team have tried to create the model with data provided after cleaning, however, the model returned showed a biased result where in the confusion matrix, the True Positive is extremely high (>0.9) and True negative is considerably low (<0.1). The overall AUC was considerable low as well.

Therefore, SMOTE is used to generate synthetic samples for the minority class (i.e., Bad Flag) while TOMEK is used to remove the majority class data with lowest Euclidean distance from minority class data.

However, this may have led to potential data leakage given that we synthesized additional data outside the given training dataset and Tomek may remove data from original set. Our group have taken some measures to address data leakage issue such as removing leaky variables (i.e. *customerid*, *systemloanid, etc.*)and testingour models on the holdout dataset but readers should be aware of the potential limitations in our approach.

**FUTURE WORK**

Our team invites interested parties to explore the topics/suggestions below:

In order to improve the accuracy of our model and predictions, it would be wise for more variables and differentiating factors of individuals to be considered. Some of which that can be considered are number of households, contribution to pension/retirement, expenditure on entertainment. As addressed in the above section on data leakage, more actual raw data for each variable can be collected (if available) to reduce the need for synthetic data generation.

To add on, the data provided for our analysis did not specify the type of loan that was involved. Since various types of loan are offered by the banks (mortgage, car, education, etc.), there may be a difference in default rate (for example, student loans may be more likely to default as they do not have income) which can uncover interesting insights and findings.

Another potential aspect that interested parties can continue is to extend the timeframe of the analysis and conduct a time series forecast. The findings would provide insights to whether time series events such as financial crisis (usually occurring every 10 years) or fluctuation in interest and inflation rate (also occurs in a cycle) would impact the repayment of loans by the individuals.

**REFERENCES**

How do banks make money*?* Corporate Finance Institute. (2022, November 8). Retrieved November 12, 2022, from https://corporatefinanceinstitute.com/resources/knowledge/finance/how-do-banks-make-money/

Nigeria non-performing loans. Nigeria Non-Performing Loans, 2007 – 2022 | CEIC Data. (n.d.). Retrieved November 12, 2022, from https://www.ceicdata.com/en/indicator/nigeria/nonperforming-loans

REINHART, C., & KLAPPER, L. (2022, May 5). *Private-debt risks are hiding in plain sight*. World Bank Blogs. Retrieved November 12, 2022, from https://blogs.worldbank.org/voices/private-debt-risks-are-hiding-plain-sight

World Economic Outlook, April 2022: War sets back the Global Recovery. IMF. (n.d.). Retrieved November 12, 2022, from https://www.imf.org/en/Publications/WEO/Issues/2022/04/19/world-economic-outlook-april-2022

Vaidya, A. (2017, July). Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

Alaradi, M., & Hilal, S. (2020, October). Tree-Based Methods for Loan Approval. In 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI) (pp. 1-6). IEEE.

**Contact Information**

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

Name: Wan Xinyu

E-mail: xinyu.wan.2022@mitb.smu.edu.sg

Name: Wang Yizhi

E-mail: yizhi.wang.2021@mitb.smu.edu.sg

Name: Yang Jingyuan

E-mail: jyyang.2022@mitb.smu.edu.sg

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS  
Institute Inc. in the USA and other countries. ® indicates USA registration.  
Other brands and product names are trademarks of their respective companies.

1. Random forest (2022, 11 08). In *Wikipedia*. https://en.wikipedia.org/wiki/Random\_forest [↑](#footnote-ref-2)
2. Boosted Regression Tree - Support – BCCVL (2022, 11 08) https://support.bccvl.org.au/support/solutions/articles/6000083202-boosted-regression-tree [↑](#footnote-ref-3)
3. 1.10. Decision Trees — scikit-learn 1.1.3 documentation [↑](#footnote-ref-4)