UNIVERSITY OF EDINBURGH
Business School

2023-24

Predictive Analytics and Modelling of Data

CMSE114282023

Credit Scoring System: Predictive Approach to Mitigate Customer Risk for Financial Institutions

GROUP 10

Word count: 3734

ABSTRACT

Risk management plays a crucial role in the banking industry. In terms of preventing losses and profitability, loan granting is one of the aspects that financial institutions focus on. It is necessary to distinguish the loan defaulters, thereby avoiding the risk of bad loans. The report aims to provide a guideline for banking institutions to evaluate their customers' credit scores. Two different techniques, including logistic regression and random forest, are utilized, which allow the banking industry to identify good or bad customers accurately. The proposed work demonstrates a result which is related to the given Brazilian customers' dataset. In conclusion, among these two machine learning approaches, random forest found to be the most effective one which is suitable for imbalanced datasets. According to the F1-score, recall value and the confusion matrix of the grouped result, the adjusting threshold method is proven to perform better. The analysis report offers preventive methods and practical insights that benefit the financial industry in managing credit risk when making decisions.

INTRODUCTION

In an era where financial institutions navigate the delicate balance between risk and opportunity, credit risk management stands at the forefront of strategic imperatives. This report encapsulates a consultancy endeavor aimed at revolutionizing the bank's credit scoring process. Accurate and predictive credit scoring models help maximize the risk-adjusted return of a financial institution. As the financial landscape evolves, so too must the tools and methodologies employed to assess the default risk of customers.

Harnessing the power of predictive modeling, the study delves into a synthetic dataset spanning one year of credit card customer interactions. Credit scoring systems are generally used to estimate loan default probability. The prime objective of this research work is to unveil a tailored predictive model to find out the best score for all the existing customers. This model is not merely a statistical exercise; it is a strategic enabler, designed to foresight informed and calibrated credit decisions since assessing credit risk is a complex and critical task for banks.

As we embark on this journey, our focus extends beyond the technical intricacies. We delve into actionable insights gleaned from the analysis, advocating for a dynamic approach that aligns with the bank's risk appetite, customer demographics and ever-changing market dynamics. This report serves as a roadmap, presenting a synthesis of technical excellence and pragmatic recommendations, with the ultimate goal of fortifying the bank's resilience in the face of credit risk.

PROBLEM STATEMENT

Addressing the imperative of credit risk modeling, the study has undertaken the task of developing a predictive model for a banking institution to evaluate the probability of customer default. Recognizing the profound implications of credit risk on a bank's capital and financial stability, our objective is to refine the credit scoring process—a pivotal instrument for informed decision-making regarding credit approval or denial.

The foundational dataset for our analysis comprises synthetic data pertaining to credit card customers over a one-year period. Credit scoring, as a nuanced practice, manifests in varying approaches across banks, characterized by distinct models, features, and approval thresholds. Our mandate is to navigate this intricate landscape and construct a predictive model, drawing upon methodologies and techniques acquired through our academic curriculum.

This report is poised to serve as a comprehensive guide for the bank, presenting and justifying the proposed model and its resultant findings. Through a meticulous analysis of the dataset, our aim is to offer insights that transcend mere predictive accuracy. The focal point lies in furnishing actionable recommendations to optimize the credit scoring process. Acknowledging the bank as the end-user, the report aligns with their vested interests, emphasizing considerations such as risk aversion, the current customer portfolio, and the strategic implications inherent in the proposed model.

Commencing with an exploration of the predictive model's robustness, this study not only highlights its efficacy but also elucidates the ensuing business implications. Recommendations are tailored to enhance the bank's credit evaluation procedures, potentially advocating adjustments in lending strategies based on customer features. By assuming the perspective of the bank, our report aspires to meet not only technical standards but also to provide strategic counsel for augmenting the overall credit risk management process.

THEORETICAL FRAMEWORK AND METHODOLOGIES

Data Selection

The meticulous data selection for an analysis is crucial to ensure the quality and relevance of the information used in the model. Also, it is necessary to explore the structure and understand the types of the data. The dataset initially consisted of 53 independent variables, each potentially contributing to the prediction of credit default, with the dependent variable being TARGET_LABEL_BAD=1. These variables covered a wide range of factors, from demographic information to financial indicators. Variables for the analysis were selected through a dual criterion approach based on the degree of relationship between independent and dependent variables. This involved assessing the statistical significance of each variable's impact on credit default likelihood.

One of the primary considerations during the data selection process was dealing with missing values. There were certain variables that had a substantial number of missing values. For example, the 'EDUCATION_LEVELI' variable had 49,175 missing values out of a total 50,000 data points. Despite its potential relationship with the dependent variable, a stringent approach was initiated and decided not to include variables with excessive missing values to ensure a robust analysis. Furthermore, the imputation in the dataset was taken into account to make the data meaningful. Firstly, the distribution of the 'MONTHS_IN_RESIDENCE' variable shows that it is skewed which makes median replacement suitable because it is less sensitive to these extreme values. Secondly, 'RESIDENCE_TYPE' and 'OCCUPATION_TYPE' variables are categorical therefore, mode is the best way to replace the missing values since mean and median lack meaningful interpretation in this specific context.

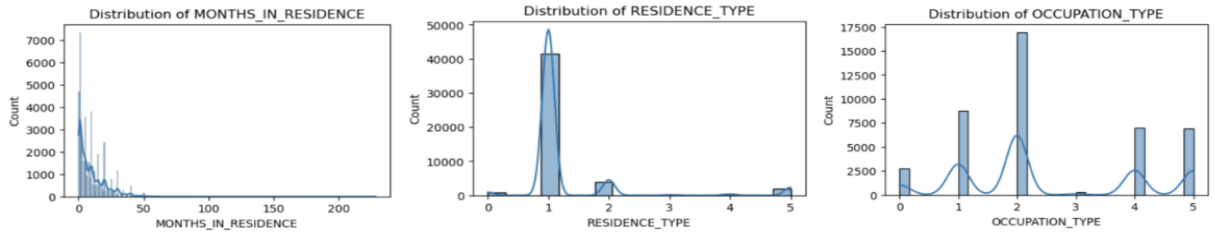

Figure1: Distributions of the selected variables

After handling the missing values the data is splitted into two nationalities, Brazilian and non-Brazilian to see whether it will be easier for a local citizen (Brazilian) to be approved by the bank. Subsequently, all the categorical variables are transformed into a format that effectively represents the independence of the categories without introducing any mathematical relationships between them, through one-hot encoding. Some variables were excluded from the analysis due to factors such as lack of available data, insignificance to the dependent variable or absence of meaningful insights. 'AGE' is one such variable provided. In the banking industry, it is logically

understood that people can open an account after turning 18. Due to that rationale, ages such as 6 or 7 were filtered out using the method above.

Finally, regarding the outliers, they are reserved for potential further research due to their meaning. More specifically, out of the initial 53 variables, a final selection of 22 was made. This balanced approach ensured a comprehensive representation of relevant factors while maintaining data integrity. Figure 2 indicates the correlation between the selected variables. For strong positive correlation there are:

1. QUANT_CARS and PERSONAL_ASSETS_VALUE
2. FLAG_DINERS and FLAG_AMERICAN_EXPRESS,
3. QUANT_BANKING_ACCOUNTS and QUANT_SPECIAL_BANKING_ACCOUNTS.

Negative correlation:

1. NATIONALITY and QUANT_DEPENDANTS
2. FLAG_MASTER_CARD and FLAG_VISA
3. FLAG_EMAIL and FLAG_MASTER_CARD
4. AGE and MONTHs_IN_RESIDENT

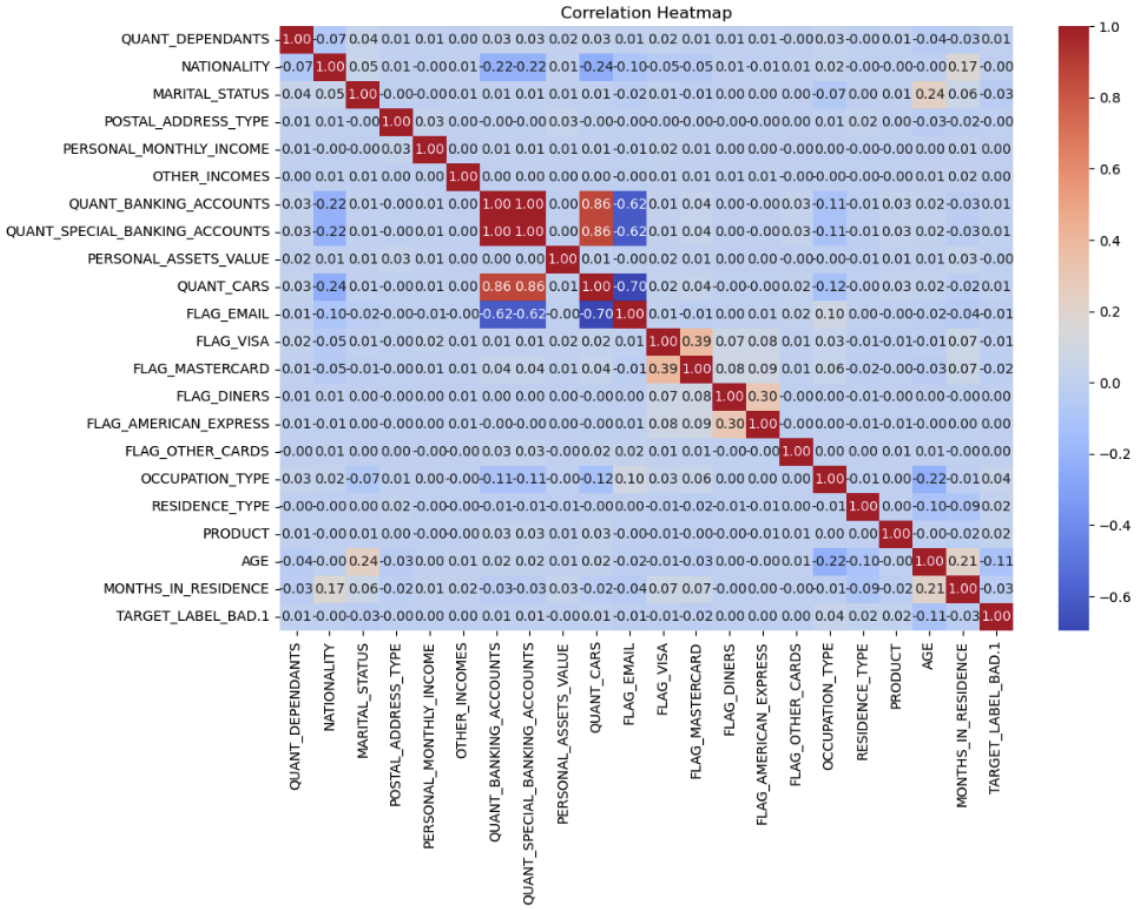

Figure 2: Correlation Heatmap of Each Variable

METHODOLOGIES

In the report, two machine learning models were utilized: Logistic Regression model and Random Forest model.

Logistic regression is a supervised machine learning model, widely used in classification problems, especially in binary classification (Kost et al., 2021). Its main advantages include simplicity, low computational costs, ease of implementation and understanding, and the ability to work on a small dataset. Logistic regression can directly estimate the probability of an event occurring, which is useful

for areas like credit scoring and fraud detection (Hosmer et al., 2013). And, the low costs of logistic regression make it a preferred learning method for many problems. In addition, logistic regression is very easy to interpret and can show which factor influences the prediction the most (Le Cessie and Van Houwelingen, 1994). Moreover, the model provides more reliable results in situations with less data. It can be very slow and hard to tackle the data when it is used on a large-scale dataset (Kost et al., 2021). In the report, due to the subject of the data which is credit scoring related and other advantages, logistics regression will be applied to the research.

Random Forest is a machine learning algorithm with various advantages, such as high predictive accuracy, applicability to imbalanced datasets, and numerous variable handling abilities. Initially, in terms of accuracy, the research discussed that precise credit score methods could significantly improve enterprises' future savings (Henley and Hand, 1997). Accuracy is considered to be a crucial factor in evaluating and choosing the model. Random forest combines multiple trees, which balance the bias of each individual tree. The other associated paper from Breiman (2001) shows that accurate categorisation might increase when the ensemble trees vote for the class with the higher popularity. Moreover, the concept is proven by Sweeney et al. (1987), who also indicates that Random forest techniques can deal with a large number of variables to avoid overfitting issues. Secondly, it is essential to decide on an applied method not only by its accuracy but also the dataset suitability. It has been demonstrated that a random forest classifier can be implemented successfully for imbalanced datasets and high-dimensional spaces (Pes, 2021). According to the previous paper, considering the dataset of cash loans granted, random forest is an effective classifier underlying the correlation-based feature selection method (Becker, 2020). Thirdly, it is necessary to consider the time cost when banking establishments estimate the default risk. Teles et al. (2021) find that random forest takes less time to build the model, which is beneficial in computerized environments. Consequently, random forest is favored in the proposed report to recognise customers due to the features above.

FINDINGS AND INSIGHTS (ANALYSIS)

Literature review

The main focus of this report is to obtain the recognition results of the default customers. In this subsection, the F1-score, recall, and confusion matrix are examined for each classifier. Several papers give information about the judgment criteria according to the numerical values from the model's result. The recall ratio is an indicator to check if the true value of a risk class can be predicted to belong to the class (Ampountolas, 2021). On the other hand, regarding the confusion matrix, Mailund (2017) proved that the accuracy of models could be evaluated by comparing actual and predicted values. More specifically, measuring the confusion matrix is important in the credit scoring section (Zeng, 2020). The other associate paper related to the application of the credit scoring system can also be found in Siddiqi (2006). By comparing the actual good and bad values with the predicted ones, the cutoff score is estimated to assess the accuracy.

Logistic Regression Model

Results Overview

The logistic regression model has yielded the following results:

1. Confusion Matrix:
    [True Negatives (TN): 7340 , False Positives (FP): 4]
    [False Negatives (FN): 2647 , True Positives (TP): 5]
    TN: 7340 customers correctly predicted good.
    FP: 4 good customers incorrectly predicted as bad.
    FN: 2647 bad customers incorrectly predicted as good.
    TP: 5 customers correctly predicted bad.

2. Classification Report:

Classification	Precision	Recall	F1- Score	Support Count
Class 0 (Good Customers)	73%	100%	85%	7344
Class 1 (Bad Customers)	56%	0%	0%	2652
Macro/Weighted Average	65%/69%	50%/73%	43%/62%	9996
Overall	73%			9996

Table 1: Logistic Regression Classification Report

Analysis

The confusion matrix and classification report reveal insights into the model's performance:

The model effectively identifies good customers (Class 0), as indicated by a 73% precision and 100% recall rate. However, it struggles significantly in accurately identifying bad customers (Class 1). This is evident from the extremely low recall rate (0%) and a low precision rate (56%), resulting in an F1-score of 0% for this class which is meaningless.

While the overall accuracy of the model is 73%, this figure is predominantly influenced by the high number of Good customers in the unbalanced dataset, thus masking the poor performance in predicting Bad customers.

Analysis from the Bank's Perspective

From a banking perspective, the ability to accurately predict potential bad customers (Class 1) is crucial. This model's performance raises concerns due to:

- **Inadequate Identification of Risk:** The near-zero recall rate for bad customers implies that the model fails to identify almost all customers who are likely to default. This is a critical shortfall as it suggests that the bank can't realize most of the high-risk customers, potentially leading to unmitigated credit risks.
- **Bad Impact on customer-Bank Relationship:** It is correct only 56% of the time. The low precision could result in undue strain on customer relations, as many customers wrongly classified as potential bad might be subjected to unnecessary scrutiny or denial of services.

Recommendations

1. Use Different Classification Thresholds

The model's tendency to predict the majority class indicates a significant imbalance in the dataset. In this case, the proportion of non-defaulters (good customers) is much larger than that of defaulters (bad customers). Data imbalance can lead to models that are biased towards the majority class, as seen in the high accuracy for predicting good but poor performance in identifying bad.

Considering different classification thresholds, rather than the default 0.5, especially in the context of imbalanced data, might yield better results. And the optimal threshold for classification can be calculated using methods such as ROC Curve and Precision-Recall Curve analysis.

2.Assess the impact of each feature on the target variable

Irrelevant or weakly correlated features might be reducing the model's effectiveness. By introducing new features or the removal of less informative ones, the model could show better predictive power.

3.Choose Alternative Models

In a study comparing machine learning models for imbalanced data in predicting renal graft failures (Mulugeta et al., 2023), Random Forest is proved superior to logistic regression with an AUC-ROC of 0.84. Among all, Random Forest has the best calibration performance (Brier score of 0.045). Logistic regression, while suitable for balanced class distributions, was less effective in this imbalanced context. This is because of a common issue in medical datasets where "normal" instances far exceed "abnormal" ones.So, exploring other machine learning models, such as Random Forest or Gradient Boosting Machines, might be more effective in handling imbalanced data and complex relationships.

Random Forest Model

As the logistic regression result shows, it cannot provide useful insight into the targets. In contrast, based on the research, the random forest is one of the static models that is suitable for solving the data imbalance problem. Therefore, the consultant team conducted the random forest algorithm to improve the outcome.

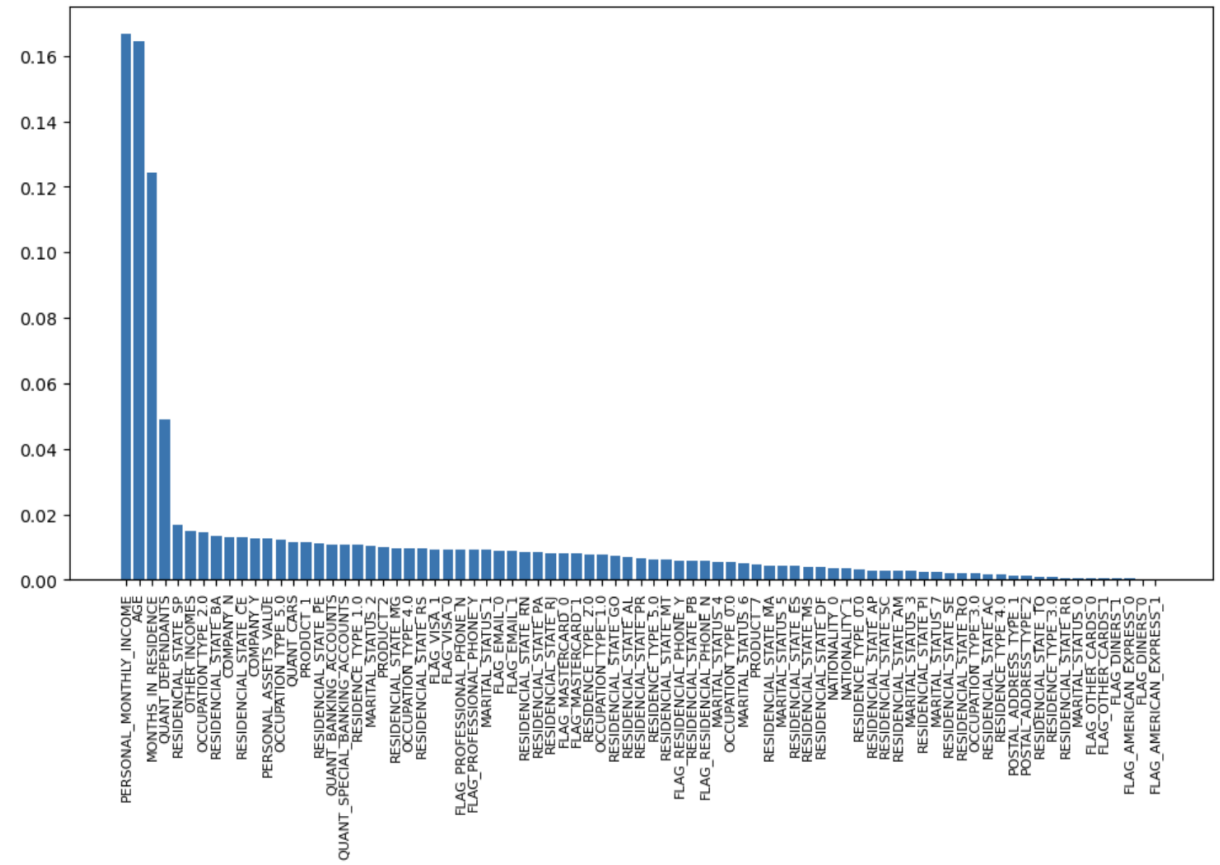

Figure 3: Variables Importance Result

In the first step, all variables are selected and grouped into the top 15, top 10 and top 5 by their importance (see Figure 3). Also, it is observable that there is an imbalance problem in the variable "TARGET_LABEL_BAD=1" of data. Consequently, three processing methods are implemented to deal with the imbalance data problem: Adjusting the weight, Altering the thresholds and Applying a balanced random forest.

Weight Adjustment

The ratio of 0 and 1 in "TARGET_LABEL_BAD=1" is assumed to be imbalanced initially. However, the result of the Python code shows that the best weight ratio is 1:1. The outcome differs from the actual dataset which means other alternative approaches are needed to use.

Adjusting Threshold Method

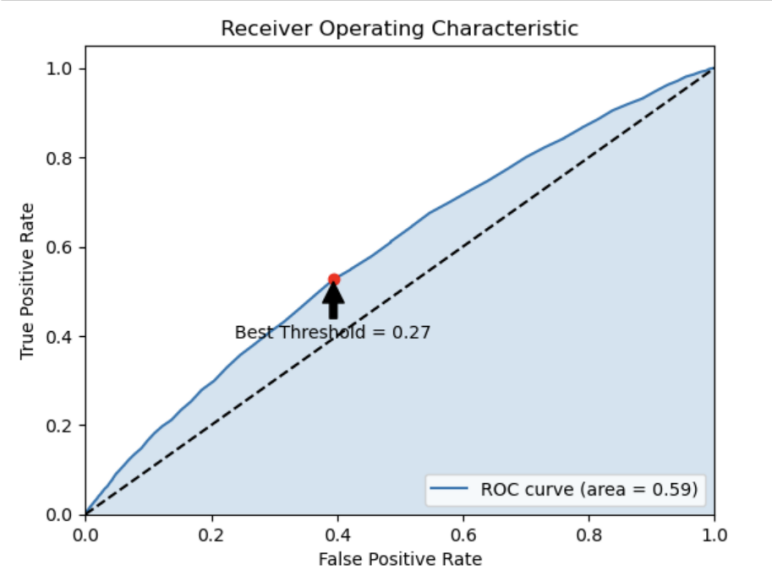

Figure 4: ROC Curve of Adjusting Threshold Method

Adjusting the threshold method is utilised to reach the purpose of improving the model performance. In the context of adjusting the threshold method, the ROC curve plays a crucial role in finding the best threshold value. There is the optimal balance of sensitivity with the point closest to the top-left corner. It is considered the best solution for larger true positives and smaller false positives. The figure 4 shows the ROC curve of the model. Noticeably, the best threshold value is selected as 0.27 underlying the upper-left corner point.

	All variables	***Top 15 variables***	***Top 10 variables***	***Top5 variables***
Confusion Matrix:	[[5463 5626] [1388 2517]]	[[5494 5595] [1673 2232]]	[[6208 4881] [1991 1914]]	[[6592 4497] [2204 1701]]
Recall	**0.64**	**0.57**	**0.49**	**0.44**
F1 Score	**0.42**	**0.38**	**0.36**	**0.34**
ROC-AUC	**0.59**	**0.59**	**0.54**	**0.53**

Table 2 : Adjusting Threshold Method Result between each Variable Category

The empirical result (see Table 2) shows the recall, F1 Score and the Roc-Auc of the model after conducting the adjust threshold method. Compared to the performance of various variable categories, the recall value of implementing a model including "all variables" is greater than others. This means the model correctly identified 64% of actual positive cases. In terms of a 0.42 F1 score, it is suggested that the model performs moderate precision and recall for the task. F1 1 means the perfect precision and recall, whereas 0 is the worst. Also, among all categories, the highest ROC value is 0.59. A ROC-AUC value closer to 1 indicates a stronger predictive capability of the model. What's more, the model's predictive performance is evaluated by the confusion matrix. These values represent the true negative (TN), false positive (FP), false negatives (FN) and true positive (TP). To focus on the objective of credit scoring, the FN and TP are two of the most important indicators for detection. It is necessary to pursue the result with the lower FN and TP to confirm the accurate identification. Lower FN value can prevent financial institutions from losing money. As for the FP value, it might affect the possibility of their profitability. Take the first column, the "all variables" result, as an example.

They demonstrate the meaning of the confusion matrix:

True Negative (TN): 5463 customers correctly predicted good.
False Positive (FP): 5626 good customers incorrectly predicted as bad.
False Negatives (FN): 1388 bad customers incorrectly predicted as good.
True Positive (TP): 2517 customers correctly predicted bad.

Based on the confusion matrix consequence above, the model with all variables performs better. Although the FP value relative is high, it has a relatively good overall performance, effectively reaching the risk-preventing goal in the credit scoring section.

Balanced Random Forest

	All variables	*Top 15 variables*	*Top 10 variables*	*Top5 variables*
Confusion Matrix:	[[6360 4729] [1614 2291]]	[[6153 4936] [1863 2042]]	[[6117 4972] [1859 2046]]	[[6051 5038] [1907 1998]]
Recall	**0.5867**	**0.5229**	**0.5239**	**0.5117**
F1 Score	**0.4194**	**0.3753**	**0.3746**	**0.3652**
ROC-AUC	**0.5801**	**0.5389**	**0.5378**	**0.5287**

Table 3 : Balanced Random Forest result between each Variable Category

The variant machine learning algorithm, balanced random forest, is applied to the dataset, with the dataset of one class outweighing the others. It is designed to distinguish the minority class and reduce towards the majority class, leading to fairer performance evaluation. The balanced machine learning result presents different numerable indicators for the variable subsets to achieve the best model configuration. According to Table 3 , the best overall performance occurs with all variables. It indicates a better F1 score of 0.4194. An approximately 0.58 recall value represents the model's ability to help identify the credit defaulters, thus helping the bank industry avoid potential losses. Furthermore, the lowest 1614 FN confusion matrix value shows the result considering all variables as well. The case of the wrong prediction of recognising bad customers as good ones might lower the categories compared to others.

To sum up, random forest is an applicable machine learning technique to assess customers' credit scores with the given dataset. On the other hand, the adjusting threshold method offers a desirable result among the three approaches. For the models with numerous decision trees, selecting the best tree might stimulate the performance of the original model. Nevertheless, there remains scope for further enhancement because of the subjective interpretation. To maintain profitability and risk resilience, the banking facilities might take not only the False Negatives (FN) but also the False Positive (FP) of the confusion matrix into account. FP can be a criterion which leads to different performance with another variable category. Also, if the consultant group can get more information about the penalty coefficient, error costs and default cost, it is possible to estimate a better performance.

CONCLUSION

In conclusion, this study oversees the methods used to predict the credit scores of customers of a financial institution to recognize loan defaulters. Credit scoring helps organizations to have an aligned environment to estimate the risks associated with defaulting. It is said that if a bank undergoes the practice of scoring their customers, they would never use other judgemental-based systems (Lewis, 1992).

This research study reveals that the predictive model, Random Forest, is best suitable for this particular case because of the particular variations and distribution of the customers' dataset. This analysis plays a crucial role in the banking sector for lenders or investors to form important customer-based decisions as well as to prevent banks from huge financial losses. Out of all the models presented above which includes decision trees and random forest predicted the value with the best accuracy.

The analysis established would help the stakeholders in improving their financial conditions by utilizing these results wisely. Even with the persisting limitations like the unavailability of some crucial data, errors and missing values, this study remains a basic framework for further modifications and future research by comparing the results of random forest values with different prediction models like GBM or genetic algorithms.

REFERENCES

1. Ampountolas, A., Nyarko Nde, T., & Constantinescu, C. (2021). A machine learning approach for micro-credit scoring. Risks, 9(3), p. 50.
2. Becker, J., Ziemba, P., Radomska-Zalas, A.. (2020). customer evaluation decision models in the credit scoring tasks. *Procedia Computer Science*, 176, pp. 3301-3309.
3. Breiman, L. (2001). Random forests. *Machine learning*, 45, pp. 5-32.
4. Henley, W. E. & Hand, D.J. (1997). Construction of a k-nearest-neighbour credit-scoring system. *IMA Journal of Management Mathematics*, 8(4), pp. 305-321.
5. Kost, S., Rheinbach, O., & Schaeben, H. (2021). Using logistic regression model selection towards interpretable machine learning in mineral prospectivity modeling. *Geochemistry*, 81(4), p. 125826.
6. Le Cessie, S., & Van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1), pp. 95-108.
7. Lemeshow, S., Sturdivant, R. X., & Hosmer Jr, D. W. (2013). Applied logistic regression. *John Wiley & Sons*.
8. Lewis, E. M. (1992). An introduction to credit scoring, Athena Press, San Rafael, CA.
9. Mailund, T. (2017). Beginning Data Science in R. California: Apress.
10. Mulugeta, G., Zewotir, T., Tegegne, A. S., Juhar, L. H., & Muleta, M. B. (2023). Classification of imbalanced data using machine learning algorithms to predict the risk of renal graft failures in Ethiopia. *BMC Medical Informatics and Decision Making*, 23(1), pp. 1-17.
11. Siddiqi, N. (2012). Credit risk scorecards: developing and implementing intelligent credit scoring (Vol. 3). John Wiley & Sons.
12. Sweeney, D.J., Williams, T.A. and Anderson, D.R., (1987). Statistics for business and economics. West Publishing Company.
13. Teles, G., Rodrigues, J. J., Rabelo, R. A., & Kozlov, S. A. (2021). Comparative study of support vector machines and random forests machine learning algorithms on credit operation. *Software: Practice and Experience*, 51(12), pp. 2492-2500.
14. Pes, B., 2021. Learning from high-dimensional and class-imbalanced datasets using random forests. *Information*, 12(8), p.286.
15. Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. Communications in Statistics-Theory and Methods, 49(9), pp. 2080-2093.