MSI Journal of
Multidisciplinary Research (MSIJMR)

IDENTIFYING KEY FACTORS OF MICROLOAN DEFAULTS IN PNG USING PREDICTIVE MODELLING TECHNIQUES

Francis Abei[1], Meysam Khoshnava[2], Jeffrey Ambelye[3], Mohsen Aghaeiboorkheili[4*]

[1, 3, 4*] School of Mathematics and Computer Science, Papua New Guinea University of Technology, Lae, Morobe Province, Papua New Guinea.

[2]School of Architecture and Construction Management, Papua New Guinea University of Technology, Lae, Morobe Province, Papua New Guinea.

*** Correspondence:** Mohsen Aghaeiboorkheili

The authors declare that no funding was received for this work.

OPEN ACCESS

Received: 25-March-2026

Accepted: 31-May-2026

Published: 05-June-2026

This article is published in the **MSI Journal of Multidisciplinary Research (MSIJMR) ISSN 3049-0669 (Online)**

The journal is managed and published by MSI Publishers.

Volume: 3, Issue: 6 (June-2026)

ABSTRACT: One of the key issues faced by financial institutions in regards to their clients, is the problem of loan defaults by their clients, and this greatly affects the sustainability of various financial institutions in PNG. This paper investigates the causes, consequences, and potential solutions to this problem. The study aims to identify the factors that cause the development of such a problem and contribute to it growing. Some of the most common factors involved include age, income, employment status, past repayment records, the reason for taking the loan, the amount borrowed, the terms of the loan agreement, the interest rate on the loan, and the level of education of the borrower. This paper addresses problems that arise as a result of high loan default rates in terms of losses for financial institutions and other lending organizations, lack of access to credit for low-income earners, and issues that are related to financial stability. The methodology implements logistic regression, random forest, and XGBoost models, which are all coded and simulated using RStudio. The research paper suggests practical solutions which include better screening of borrowers through predictive models, better credit

monitoring, and financial literacy programs. In PNG, there is a need for all stakeholders involved to come up with a solution that addresses this issue. Other key areas discussed in this paper include credit risk management, which involves data analytics that aim to reduce the chances of loan defaults and improve portfolio performance, ensuring that financial services are available to all our people.

Keywords: *Loan Default; XGBoost; Logistic Regression; Random Forest; feature importance*

Introduction

Microfinance has emerged as a very important instrument to help promote financial inclusion for low-income people in developing countries, particularly in PNG, where access to formal credit is still very limited (Ledgerwood, 2013). By letting people take small loans, financial institutions encourage entrepreneurship and help reduce poverty. But these individuals, many of them do not have collateral or credit histories, hence if they default on a loan, the lenders make a loss from these loan defaults. (Mersland & Strøm, 2010)Some recent developments in data analytics provide promising solutions that allow lenders to take advantage of some attributes of a borrower such as age, income, employment status, payment history, purpose of the loan they took, the loan amount, etc., to predict the risk of a loan default more accurately (Lessmann et al., 2015). Machine learning methods, in contrast to the rule-based approach of traditional lending, can detect complex relationships and provide interpretable insights into the factors that lead to loan defaults (Hassan, et al., 2024).

Financial institutions, despite the possibility of having large records of borrower data, still resort to subjective or simplistic ways of assessing credit risk, which results in suboptimal lending decisions, higher losses, and limited access to credit for other creditworthy individuals (Berger & Frame, 2007). Some key questions are usually left unanswered, and without systematic, data-driven analysis, our financial institutions will not be able to optimize their lending criteria or be able to implement strategies that help mitigate risk and high default rates that compromise the financial stability of the institution and limit its outreach into underserved communities.

This research paper uses data analytics to show and identify the primary factors that have an impact in loan defaults among individual borrowers within PNG. The specific objectives that this paper aims to accomplish are:

1. Quantify the impact of nine key factors that play a role in loan defaults, and these are; age, income, employment status, repayment history, loan purpose, loan amount, loan terms, interest rate, and the level of education of the individual. These factors will be simulated through default probability using logistic regression, Random Forest, and XGBoost models.
2. Compare the performance of each model and the feature importance rankings across algorithms to determine the most reliable factor that can be used to predict loan default risk.
3. Conduct simulation analyses to demonstrate how variations in borrower profiles affect predicted loan default probabilities.
4. Provide actionable recommendations for financial institutions that will help enhance their credit decision-making and help reduce their losses associated with their lending activities.

This study done has its main focus on individual-level loan default predictions using data that is simulated and tailored to represent the average borrower profile within PNG, incorporating the specified nine independent variables. Logistic regression, Random Forest, and XGBoost models are created in an R environment to simulate how these factors affect loan defaults, thus emphasizing interpretability through coefficients, and feature importance scores. The scope of these simulations does not include macroeconomic factors, portfolio-risk aggregation, and real-time deployment. The limitations of these simulations include the reliance on simulated data rather than data from a financial institution, potential omitted variables that could impact the probabilities of loan defaults, and having constraints that are generalized to PNG-specific finance contexts. The findings in this study should be interpreted as being indications that are of relative importance rather than absolute default probabilities.

The paper layout comprises of five chapters. The first chapter being introduction, which presents the background, problem statement, the objectives, scope, limitations and the structure of the paper. Chapter 2 is the literature review on loan default factors and predictive modeling techniques. Chapter 3 gives details on the methodology used to run the simulations, it also details the specifications of each model, and the procedures run. Chapter 4 discusses the results from the simulations, compares the models, and gives the overall analyses of the simulations. The final chapter of this paper discusses the findings and its implications for the financial institutions in PNG, and also presents recommendations for future research.

Literature Review

Microfinance is essential for financial inclusion, especially in developing economies where there are numerous individuals and small businesses who are excluded from formal banking. Financial institutions provide small loans to individuals to help support entrepreneurship, generate income, and to reduce poverty in the society. The issue is that the people taking out these loans often have no collateral, no stable jobs, and no credit history, so it is hard to know if they are able to pay back the loan in full. The sustainability of these financial institutions is also dependent on the repayment performances of these small loans; hence the prediction of loan defaults has become a major focus in credit risk research (Ledgerwood, 2013; Mersland & Strom, 2010). This literature review points out the why it is necessary to identify the characteristics of the loan and the person taking the loan and use these characteristics to help estimate the risk of a loan default before the loan is approved.

Logistic regression is the modelling technique that is used to help predict loan defaults because it fits binary outcome and its coefficients are easy to interpret. Various variables such as income, age, employment status, and repayment history have been identified as frequently used in logistic regression models for this kind of problem by research in microfinance and credit scoring (Lessmann, Baesens, Seow, & Thomas, 2015). The principal advantage of this method is that it creates transparency, which is useful to lenders and policy makers. However, it is limited in the case of non-linear relationships or when some individual characteristics interact in a complex way because it does not fully capture the underlying patterns behind the risk of loan defaults (Hosmer Jr, Lemeshow, & Sturdivant, 2013).

The more recent studies are starting to move towards machine learning methods, such as Random Forest and XGBoost. These methods are very useful in terms of how they accommodate nonlinear effects and interactions without strong statistical assumptions. Random forest improves classification by combining multiple decision trees, whereas, XGBoost builds accurate predictive models using gradient boosting (Breiman, 2001; Chen & Guestrin, 2016). In credit risk research, these methods often outperform simpler models in terms of predictive accuracy, but their results are not that easy to interpret, and thus feature importance and other tools are needed to make their results more meaningful for lending decisions (Lessmann, Baesens, Seow, & Thomas, 2015).

This literature review consistently identifies several important predictors of loan defaults. Characteristics of the borrower such as age, income, employment status, education, and repayment history are often associated with repayment capacity and financial stability. Older people are sometimes considered less risky, while higher income, financial stability and a strong record of repayments are also associated with lower risks of loan defaults (Murdoch, 1999; Berger & Frame, 2007). Loan-related factors are also important, such as loan amount, repayment terms, the interest rate, and the loan purpose. All these affect the repayment burden and the stress on the individual. Larger loans, higher interest rates, and loans that do not generate immediate returns can increase the risk of a loan default, showing that loan defaults are shaped by borrower capacity, financial behavior, and how the loan is designed, all together (Goddard, 2009; Miller, Ledgewood, Earne, & Nelson, 2013).

Recent works emphasize that accuracy is not enough, especially when loan default cases are much less frequent than non-default cases. Currently, researchers use methods such as precision recall, F1-score, AUC, and tools that help facilitate interpretation such as feature importance and coefficient analysis to ensure models that predict and explain the risk that is involved (Saito & Rehmsmeier, 2015). However, a big gap still exists, as there are many studies are on general banking and not microloans, and few compare logistic regression, random forest, and XGBoost in one framework. There is also very little research that has been done in PNG on this problem, which needs a methodologically sound and locally relevant study.

Methodology

This paper focuses on the factors that influence loan defaults, from the approach of a quantitative predicting model. The emphasis is on how borrower characteristics and loan conditions influence the probability that a borrower fails to repay a loan. The goal is to build models that can predict both loan defaults, and explain what the key drivers of this issue are.

The study requires a data set that includes a binary outcome of default, along with key explanatory variables such as age, income, employment status, repayment history, the purpose of the loan, the loan amount, the terms of the loan, the interest rate, and the level of education of the individual. The paper also highlights what factors are strongly associated with loan defaults and clearly be explains the impact of these factors on the loan default probability.

This objective has been set up as a supervised classification problem where we have compared three algorithms; namely logistic regression, random forest, and XGBoost. Now, to elaborate further on the techniques that we will be using, we assume that the logistic regression model acts as a baseline; we assume it as a simple approach, and for dealing with complexity, we take the help of the other two models which are; random forest and XGBoost.

The analysis is done in RStudio, a platform for data preparation, statistical modelling, machine learning, and visualization. The execution of these models follows a sequence of data preparation and cleaning, data exploration, splitting data into training and testing sets, fitting of the three models, and evaluating these models using precision recall, feature importance and AUC. Then graphical outputs and feature importance measures are used to make the results more clear, understandable, and applicable.

Results and Discussion

Logistic Regression Graphs and Results:

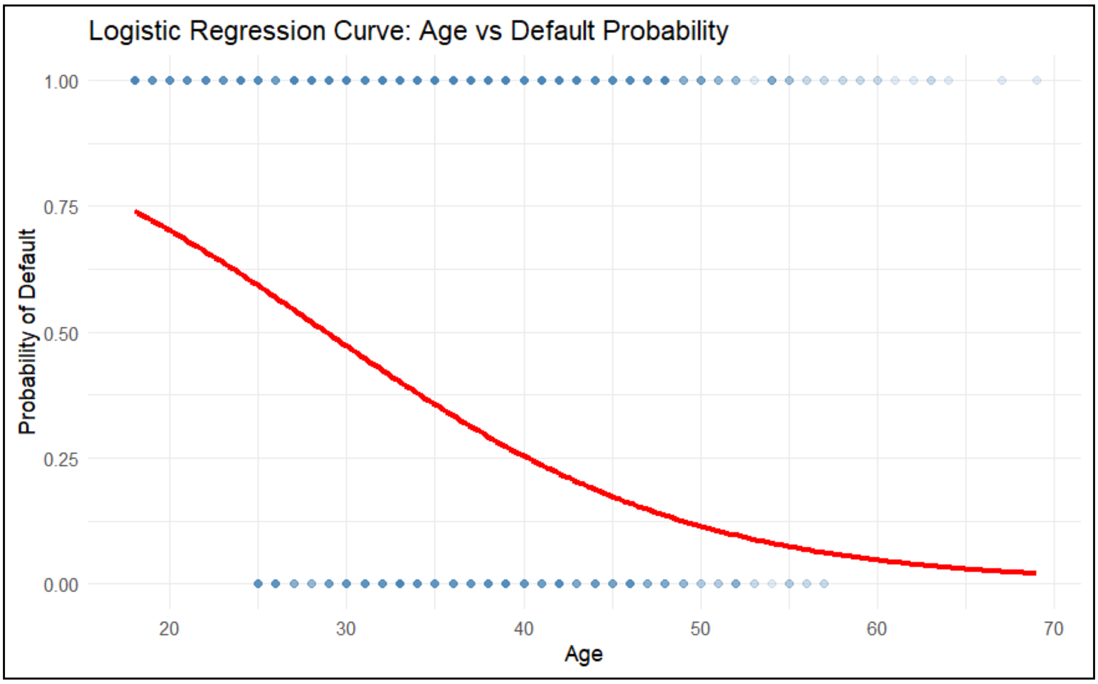

Figure 1: Age Vs. Default probability

- This is a logistic regression curve that shows the probability of loan default vs. age. This graph displays the predicted probability of defaulting on a loan as a function of the age of a borrower. The red curve is a strongly negatively correlated curve. The figure shows that younger individuals have a high predicted probability of not being able to repay the loan, this probability is around **0.75**. The probability declines gradually with age, and is close to 0.0 for borrowers in their 60s. This suggests that older borrowers are viewed as significantly more reliable by the model.

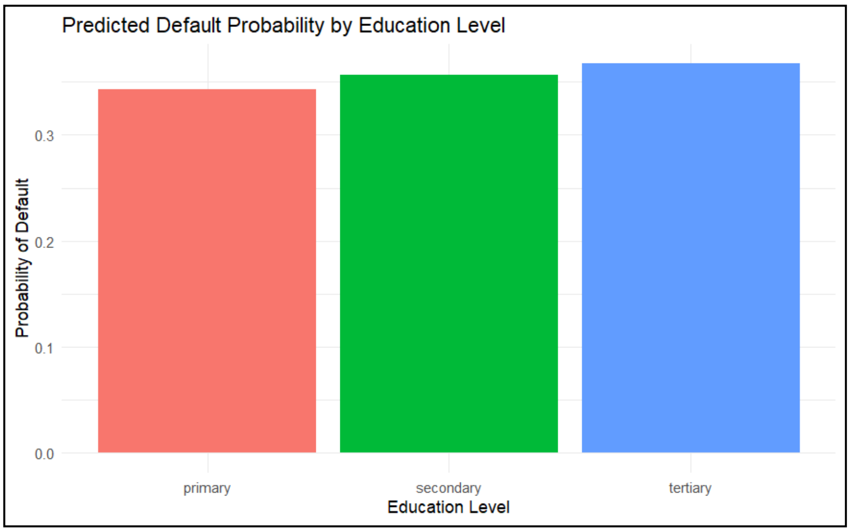

Figure 2: Default Probability by Education Level

- This is a bar chart. This graph compares three levels of education; primary, secondary, and tertiary. The differences between these categories are minimal. All levels fall within a narrow range between **0.34 and 0.37**. While tertiary education shows the highest predicted loan default probability and primary the lowest, the small variance indicates that education level is **not a dominant predictor** in this model.

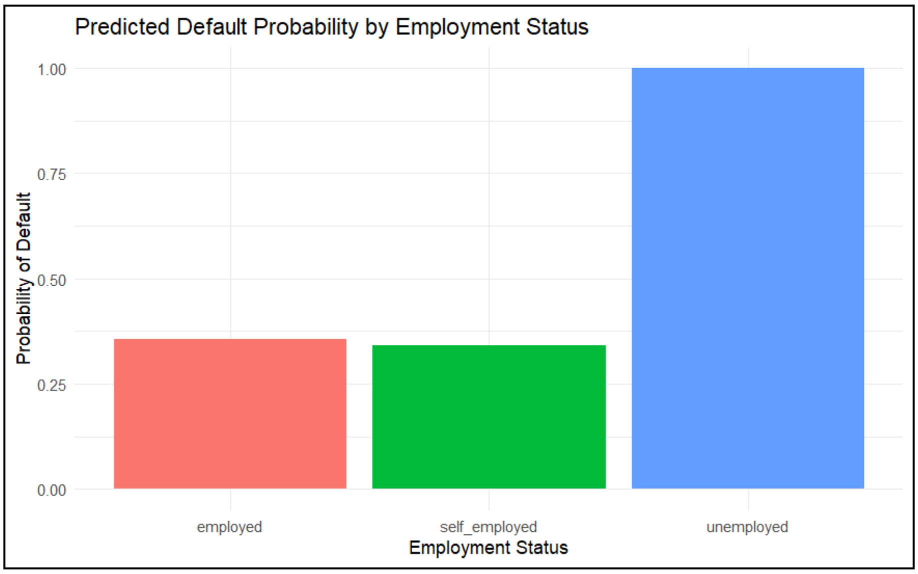

Figure 3: Default Probability by Employment Status

- This is a bar chart that examines the impact of being employed, self-employed, or unemployed. Those who are employed, or self-employed have a relatively low risk of loan defaults, around **0.35**. However, **unemployed individuals** have a high default rate.

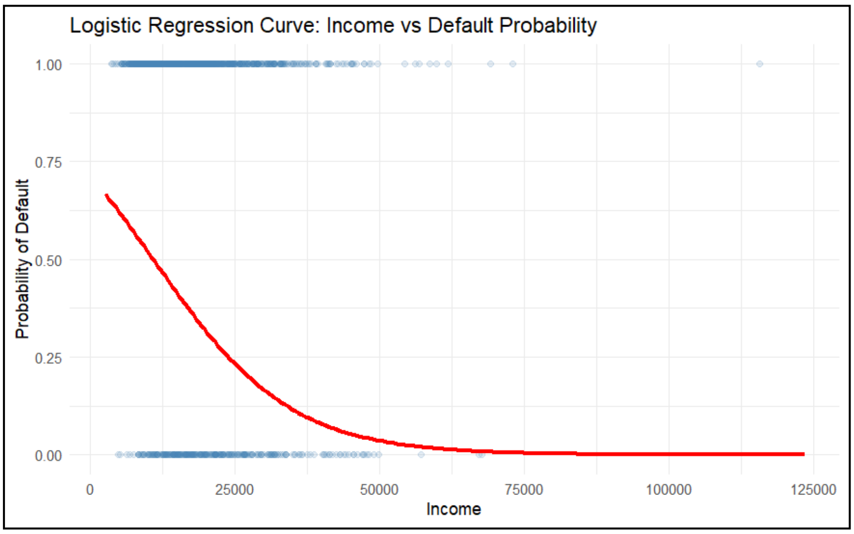

Figure 4: Income Vs. Default Probability

- This graph shows a logistic curve of income against loan default probability. This curve tracks how annual income affects the likelihood of loan defaults. There is a **sharp inverse relationship**. Individuals with very low incomes face a high risk of defaulting on a loan. As income increases, the probability of loan default drops rapidly hitting nearly **zero** once the income surpasses a certain limit. This identifies income as a major factor in being able to successfully repay a loan.

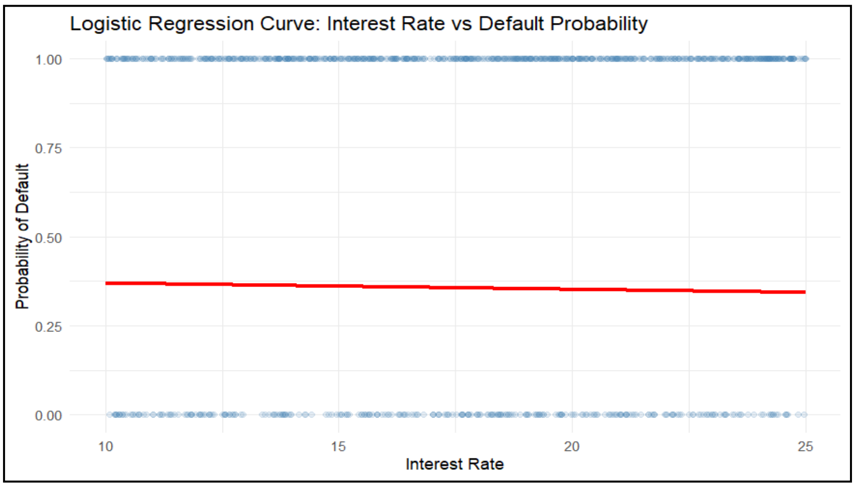

Figure 5: Interest Rate Vs. Default Probability

- This logistic curve analyzes if the cost of the loan (the interest rate) influences individuals to default on a loan. The red regression line is **nearly flat**, staying consistently around the **0.35 probability** across the entire 10% and 25% interest rate range. This suggests that, within this specific model, the interest rate itself does not significantly play a role in having an individual default on a loan.

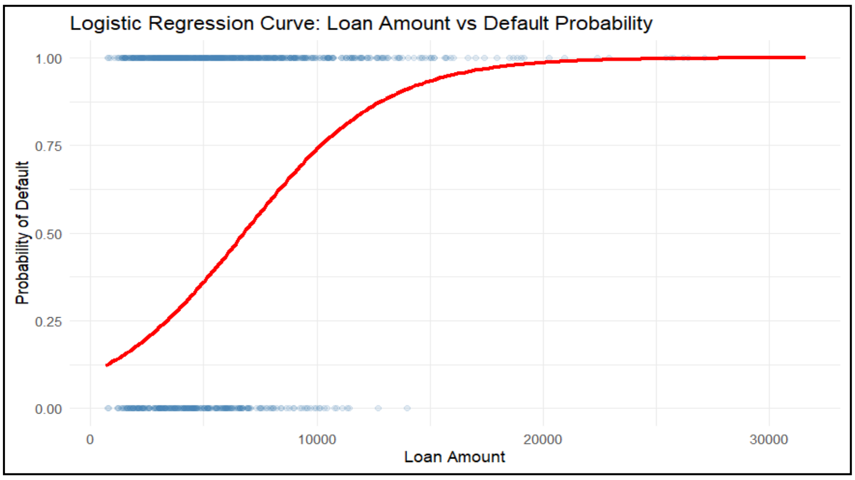

Figure 6: Loan Amount Vs. Default Probability

- This curve plots the total amount of a loan taken against the loan default probability. The model does show a **strong positive correlation** following an S-curve. Small loans have a low loan default probability, however, as the amount increases the risk climbs steeply for loans that are of large amounts.

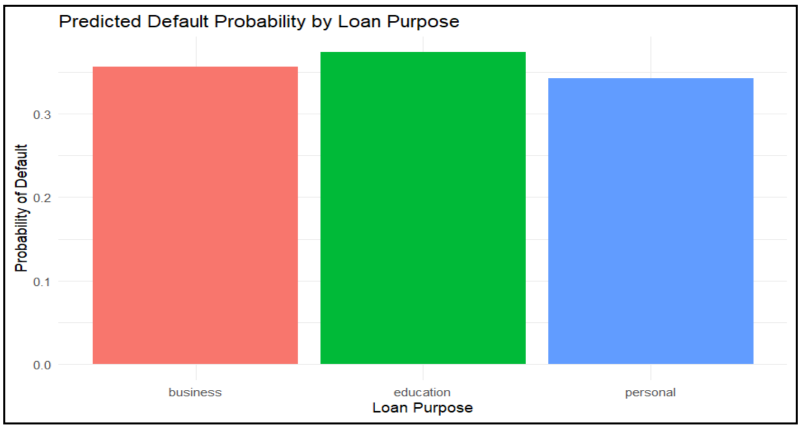

Figure 7: Default Probability by Loan Purpose

- This bar chart compares three types of loan purposes; business, education, and personal. The results appear to be relatively uniform. All three purposes show a predicted default probability of approximately **0.35**. Loans taken for education seem to have a slightly higher risk than those for business or personal use, but the difference is statistically small.

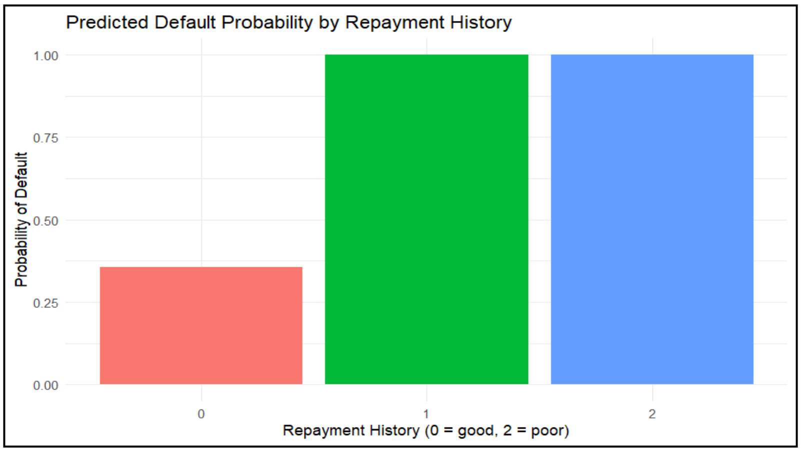

Figure 8: Default Probability by Repayment History

- This chart compares individuals with "good" repayment history (level 0) against those with very "bad" repayment history (levels 1 and 2). This is one of the most significant predictors in this model. Individuals with a good repayment history have a manageable default probability of roughly **0.35**. In contrast, any individual that borrows and is a category 1 or 2, is assigned a **1.00 predicted probability of defaulting on a loan**. This highlights that past behavior is a primary factor that plays an important role in the model.

The results shown in these various graphs indicate that **age, income, employment status, the loan amount, and repayment history** are dominant factors that influence loan defaults in this model. Older individuals that take out a loan, and those with higher income are predicted to have a lower probability of defaulting on a loan, while those individuals who are unemployed and those individuals with poor repayment histories are classified as individuals that are highly likely to be unable to repay loans. Loan amount is also a factor to consider as a strong positive predictor, with large loans associated with having a higher loan default risk. On the other hand, the factors such as the level of education, the purpose of the loan, and interest rate associated with the loan, tend to have minimal influence on loan defaults, as their predicted probabilities remain relatively stable across categories or ranges.

Thus, this analysis demonstrates that this logistic regression model places emphasis on the capacity of the borrower and their repayment behavior rather than on the type of loan or how much it costs to take out the loan. Repayment history is the most important factor, as it shows that past borrowing behavior can also influence future repayment performance. Employment status and income are factors which also play a major role as they reflect the individual's ability to generate cashflow and be able to make repayments. In summary, this demonstrates that loan defaults are primarily based on the need for financial stability and past repayment history. The amount of money borrowed also influences the likelihood of a loan default.

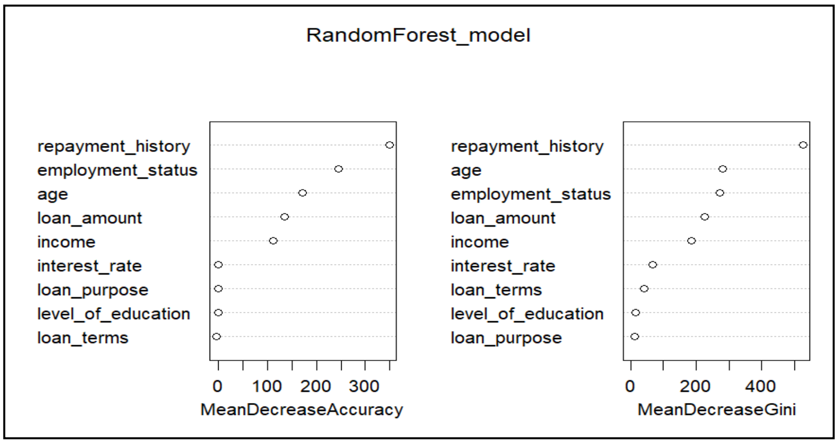

Figure 9: Random Forest Model

Random Forest Model

According to the results from the Random Forest variable importance plot, the most significant variable which plays a major role in loan defaults in this case study is **repayment history**. There are also other important factors to consider, they are not as significant as repayment history, but also have some influence in loan defaults, and these are employment status and age. Therefore, for this model, the main factors that determine whether an individual defaults on a loan are repayment history, age, and employment status. In summary, these variables are the main factors that help the model identify potential cases of loan defaults.

As observed in Figure 9, income and loan amount are moderately significant when determining and making predictions using this model; however, these two variables are less significant when compared to the factors; repayment history, and employment status. The implication is while financial capability is an important factor to also consider, the behavior of the borrower is a more important variable to help predict loan defaults when compared to loan design parameters such as interest rates, the loan purpose, level of education, and the loan term.

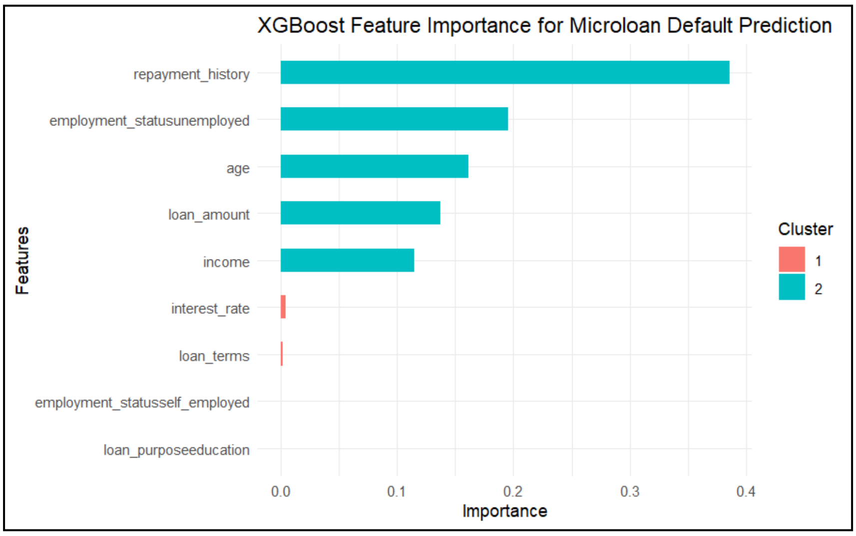

Figure 10: XGBoost Model – Feature Importance Graph

XGBoost Model

The feature importance graph of the XGBoost model shows that the factor, **repayment history** is by far the most influential factor in determining loan defaults. This is followed by **unemployment status**, **age**, **the amount of the loan**, and

income. The model places emphasis on these specific borrower characteristics, especially repayment history, when predicting whether an individual repays the loan or not. Based on this result, these factors are described as **dominant drivers** of risk in the XGBoost model.

On the other hand, the other factors which are **interest rate**, **loan terms**, **self-employment status**, and **the purpose of the loan**, appear to have very little influence in loan defaults. The scores that they have, indicate that they have minimal influence when compared with the borrower's financial stability and their repayment history. Overall, this model has demonstrated that loan default is primarily explained by the borrower's prior repayment history and their socioeconomic condition rather than by how the loan is designed.

Discussion

The models that have been used in this study have all given the same conclusion that repayment history is the most important thing when it comes to predicting loan defaults. Looking at the graphs from all three models, logistic regression, Random Forest, and XGBoost, these graphs indicate that factors such as **employment status, age, income, and the loan amount**, are also factors to consider when determining whether an individual defaults on a loan or not. Other factors such as **interest rate, purpose of the loan, the terms of the loan,** and **the level of education of an individual** are not that important when assessing loan default risk.

When an individual fails to repay a loan, the reason for that happening would have to do with how stable their finances are, their repayment history, and their socioeconomic status. The results that have been collected from the models imply that financial institutions should pay close attention to the reliability of an individual and how the individual has repaid their loans in the past, when trying to carry out credit risk assessment. The results indicate that loan defaults are mostly caused due to the individual characteristics of the borrower rather than about how the loan is designed.

Conclusion

This paper sought to determine the variables that strongly influence loan defaults. This investigation was done through logistic regression, random forest, and XGBoost modelling. The results revealed that past repayment records have a strong impact on determining whether a loan defaults or not. Employment status, age, income and loan size are also other significant factors, while the role of interest rates, the purpose of the loan, loan terms, and the level of education of the individual is insignificant. In general, it appears that the characteristics of borrowers as well as their economic situation plays a more significant role in loan defaults than the actual loan structure.

The aims of this research were successfully accomplished through constructing an analytical framework for predicting and comparing the results of these three models to determine which one gives a good comprehensive insight into the issue. It is evident that logistic regression provides good results that can be easily interpreted, whereas random forest and XGBoost reinforce those results and helps discover more about patterns within the dataset. Altogether, these models deliver consistent results, making this study valid and useful for future practical practices.

The study done has its limitations, since the dataset used is a simulation and not derived from any financial institutions. Other important variables, such as family size, previous debts and savings, were not included, and actual data might give different results due to missing data and imbalance. Further studies will have to utilize real data on loans, consider other factors related to the borrower and the financial institutions, apply modern machine learning algorithms, and analyze various borrower populations and geographic areas for more accurate results.

References

1. Berg, T. (2018). Got rejected? Real effects of not getting a loan. The Review of Financial Studies, 31, 4912–4957. https://academic.oup.com/rfs/article-abstract/31/12/4912/4962209

2. Berger, A. N., & Frame, W. S. (2007). Small business credit scoring and credit availability. Journal of small business management, 45, 5–22. https://www.tandfonline.com/doi/abs/10.1111/j.1540-627X.2007.00195.x
3. Breiman, L. (2001). Random forests. Machine learning, 45, 5–32. https://link.springer.com/article/10.1023/a:1010933404324
4. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, (pp. 785–794). https://dl.acm.org/doi/abs/10.1145/2939672.2939785
5. Goddard, G. J. (2009). The Economics of Microfinance by B. Armendariz & J. Morduch: (2007). MIT Press, 360 pages, rev. paperback ed. ISBN-10: 0262512017, ISBN-13: 978–0262512015. The Economics of Microfinance by B. Armendariz & J. Morduch: (2007). MIT Press, 360 pages, rev. paperback ed. ISBN-10: 0262512017, ISBN-13: 978–0262512015. Taylor & Francis.
6. Hassan, M. A., Mansur, U. M., Jha, R., Fahim, M. F., Mahesh, T. R., & others. (2024). Interpretable machine learning models for credit risk assessment. 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), (pp. 361–365).
7. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. John Wiley & Sons. https://scholar.google.com/citations?user=8Sjt6C0AAAAJ&hl=en&oi=sra
8. Ledgerwood, J. (1998). Microfinance handbook: An institutional and financial perspective. World Bank Publications.
9. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European journal of operational research, 247, 124–136. https://www.sciencedirect.com/science/article/abs/pii/S0377221715004208

10. Mersland, R., & Strøm, R. Ø. (2010). Microfinance mission drift? World development, 38, 28–36. https://www.sciencedirect.com/science/article/abs/pii/S0305750X09000990

11. Miller, C., Ledgewood, D., Earne, J., & Nelson, C. (2013). The Microfinance Handbook; A Financial Market System Perspective.

12. Morduch, J. (1999). The microfinance promise. Journal of economic literature, 37, 1569–1614. https://www.aeaweb.org/articles?id=10.1257/jel.37.4.1569

13. Rahman, A. (1999). Micro-credit initiatives for equitable and sustainable development: who pays? World development, 27, 67–82. https://www.sciencedirect.com/science/article/abs/pii/S0305750X98001053

14. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10, e0118432. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432&trk=public_post_comment-text