ANALYSIS OF LOAN DELINQUENCY PREDICTION BASED ON MULTINOMIAL LOGISTIC REGRESSION AND RANDOM FOREST

H. A. P. L. J. Hettiarachchi [*] ***and Ruwan Punchi-Manage***
Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Sri Lanka

The banking sector plays a critical role in fostering economic growth by extending credit to individuals and businesses; however, effective loan portfolio management remains a persistent challenge due to the risks associated with Non-Performing Assets (NPAs). Rising NPAs, often driven by economic downturns, inadequate risk assessment, and external shocks, pose a significant threat to financial stability. To address these challenges, this study explores predictive modeling approaches for loan delinquency classification by employing the Multinomial Logistic Regression Model and the Random Forest model applied to 43,644 loan records from a Sri Lankan bank. The analysis categorizes loan performance into four levels, ranging from A0 (performing loans) to D0 (severely delinquent loans), using several financial and demographic variables. Multinomial logistic regression shows that interest rate and loan age are the most influential predictors. The model indicates that a 1% rise in interest rate increases the risk of delinquency by 5.7–21.1%, while each additional month of loan age amplifies the likelihood by 27–76%. Delays in recovery also significantly elevate risk for severely delinquent loans, with each additional day of delay associated with a 1.2% increase in default probability. The model demonstrates excellent discriminatory power at the performance extremes (AUC: 0.987 for A0 and 0.996 for D0). The Random Forest model considers the loan status as a binary (current vs. delinquent) variable. An 80:20 training-testing split was used for data analysis. The performance of the model was evaluated using a confusion matrix, AUC-ROC curves, and accuracy metrics on a testing set. The Random Forest model pidenterforms better in overall predictive accuracy, with a low 1.55% out-of-bag error rate. It achieves 99.4% accuracy in classifying current loans and 96.6% for delinquent loans. Variable importance analysis confirms loan age, recovery date, and interest rate as dominant predictors. Our study is cross-sectional, with predictor variables measured at a defined observation point for each loan. Class imbalance is a limitation; to address this, we plan to apply class weighting and evaluate model performance using accuracy metrics in future work. Collectively, the models underscore the predictive strength of time-dependent variables (loan age and recovery delays) and financial indicators (interest rates, outstanding amounts). While multinomial logistic regression offers nuanced insight into risk progression across multiple categories, Random Forest delivers robust binary classification performance.

Keywords: loan delinquency, multinomial logistic regression, random forest, loan age, interest rates, recovery timeline, risk modeling, imbalanced data

[*]*Corresponding Author:s19386@sci.pdn.ac.lk*

ANALYSIS OF LOAN DELINQUENCY PREDICTION BASED ON MULTINOMIAL LOGISTIC REGRESSION AND RANDOM FOREST

H. A. P. L. J. Hettiarachchi and Ruwan Punchi-Manage***
Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Sri Lanka.

INTRODUCTION

The global banking sector faces persistent challenges in managing loan portfolios effectively, particularly in developing economies where economic volatility and rapid credit expansion often outpace risk management capabilities. Loan delinquency shows an alarming upward trend in recent years (Central Bank of Sri Lanka, 2023). This study examines loan delinquency in Sri Lankan banks, which has emerged as a critical threat to financial stability. Our research uses two powerful predictive modeling approaches, multinomial logistic regression (MLR) and Random Forest (RF) (Liaw & Wiener, 2002) to address the loan delinquency classification (Breiman, 2001). The analysis is threefold: First, comparing the predictive performance of parametric (MLR) and non-parametric (RF) approaches in handling the hierarchical nature of loan delinquency. Second, identifying and quantifying the impact of key risk drivers across different delinquency stages, with a particular focus on financial indicators such as interest rates, outstanding amounts, and temporal factors, including loan age and recovery timelines. Third, evaluating each model's capability in handling the inherent class imbalance problem in loan portfolios, where performing loans typically outnumber delinquent ones. The objectives of this study are to: (1) examine how financial variables (interest rates, outstanding amounts) interact with temporal factors (loan age, recovery delays) to influence delinquency risk progression. (2) determine which modeling approach better captures the transitional nature of intermediate delinquency categories; (3) evaluate whether ensemble methods can effectively bridge the interpretability–accuracy trade–off in risk prediction models.

METHODOLOGY

Data set: The analysis utilized a cleaned dataset of approximately 43,644 loan records with 16 variables (e.g., interest rate, loan age, gender, etc.).

Multinomial Logistic Regression: The asset class variable has four levels. A multinomial logistic regression model was fitted to the data. The model takes the form: $ln\left[\frac{\Pr(Y_i=k|\boldsymbol{X})}{\Pr(Y_i=K|\boldsymbol{X})}\right] = \boldsymbol{\beta_k^T X}$; where $\boldsymbol{\beta_k}$ is the vector of coefficients for category k, and $P(Y_i = k|\boldsymbol{X})$ is the probability of observing the asset class category k given the predictors $\boldsymbol{X}$ (e.g., interest rate, loan age, gender, etc.). The class probabilities are given by: $P(Y_i = k|\boldsymbol{X}) = \frac{\exp(\boldsymbol{\beta_k^T X})}{\sum_{j=1}^{K}\exp(\boldsymbol{\beta_j^T X})}$, for $k = 1, 2, \dots, K-1$, and the

probability for the reference category K is $P(Y_i = K|\boldsymbol{X}) = \frac{1}{\sum_{j=1}^{K} \exp(\boldsymbol{\beta}_j^T \boldsymbol{X})}$, $\beta_K = 0$ for identifiability. The odd for a one-unit increase in the predictor X_j, comparing category i to the reference category K, is: $OR = \exp(\beta_{ji})$, where β_{ji} is the coefficient for the predictor X_j in category i (Venables & Ripley, 2002).

Random Forest: For the Random Forest classification method, we consider the loan status variable as binary (*Default* or *Non-Default*). A Random Forest algorithm is an ensemble supervised machine learning method known for its robustness and accuracy. The Random Forest builds multiple decision trees during training and makes predictions based on the majority vote among these trees, thereby reducing the risk of overfitting associated with individual decision trees. Each tree in the ensemble is trained on a bootstrap sample of the training data (sampling with replacement), and at every node split, a random subset of predictors is considered. This strategy (random feature selection at each node) helps reduce the correlation among the trees and enhances the model's ability to generalize to unseen data. Prediction for a new observation *X* is given by $\{\widehat{Y}\} = \{\text{majority vote}\}\{T_b(x)\}_{\{b=1\}}^{\{B\}}$ where T_b represents the b^{th} decision tree and *B* is the total number of trees (set to 500 in this model). Variable importance is measured by the mean decrease in Gini impurity, a metric that quantifies how much each variable contributes to improving the purity (homogeneity) of the target classes at each split across all trees. The class probability for a given class k is estimated as the proportion of trees that predict class *k*, given by: $P(Y = k \mid X) = \frac{1}{B}\sum_{\{b=1\}}^{\{B\}} I(T_b(x) = k)$, where I(·) is the indicator function that equals 1 if the condition is true, and 0 otherwise. The model's performance was evaluated using a confusion matrix, AUC-ROC curves and accuracy metric on a held-out test set comprising 20% of the dataset. These evaluation metrics were used to assess the model's ability to classify both default and non-default cases correctly. For binary classification, OOB error rates, precision-recall curves, and Matthew's correlation coefficients were used.

Computation: This study was performed using R software (R Core Team, 2023). An 80:20 training-testing split was used for data analysis. The *nnet* package (Ripley, 2002) was used for multinomial logistic regression analysis, while the *randomForest* package (Liaw & Wiener, 2002) was used for ensemble learning in classification tasks. Model evaluation metrics were computed using the *caret* package (Kuhn, 2008) and the *pROC* package (Robin *et al.*, 2011).

RESULTS AND DISCUSSION

The results from the multinomial logistic regression model reveal several important predictors of loan status across the categories of non-performing loans (A1), underperforming loans (B0), substandard loans (C0), and bad loans (D0), relative to the reference group. Interest rate is a consistently significant predictor across all categories, with positive coefficients indicating that higher interest rates

increase the likelihood of falling into riskier loan statuses. For instance, in category A1, each unit increase in interest rate raises the odds by approximately 5.3%. Loan age also shows a strong positive effect, with older loans being increasingly likely to fall into default-related categories; notably, the odds of being in D0 nearly double with each additional unit of loan age. Gender (being male) has a statistically significant but small effect, slightly increasing the odds across most categories. Outstanding loan amounts are positively associated with risk, suggesting that higher unpaid balances contribute to loan deterioration.

Table 1: Multinomial Logistic Regression Model (Coefficients and Significance Level)

Variable	A1		B0		C0		D0	
(Intercept)	–4.532	***	–6.672	***	–7.669	***	–10.152	***
Interest Rate (X_1)	0.052	***	0.134	***	0.164	***	0.211	***
Price Granted (X_2)	4.5×10^{-7}	***	5.3×10^{-7}	***	3.1×10^{-7}		1.5×10^{-7}	
Loan Age (X_3)	0.189	***	0.277	***	0.419	***	0.6728	***
Gender Male (X_4)	0.023	***	0.082	***	0.002	***	0.0675	***
Outstanding (X_5)	2.1×10^{-6}	***	2.4×10^{-6}	***	1.9×10^{-6}	***	1.4×10^{-6}	**
Price Due (X_6)	-5.7×10^{-7}		1.2×10^{-6}	***	6.3×10^{-7}		-2.6×10^{-7}	**
Interest Due (X_7)	6.7×10^{-6}		2.6×10^{-5}	***	2.7×10^{-5}	***	2.2×10^{-5}	***
Charge Due (X_8)	2.5×10^{-4}	***	-1.7×10^{-4}	**	0.0001	**	0.0003	***
Recovery Date (X_9)	7.8×10^{-3}	***	1.34×10^{-2}	***	0.013	***	0.014	***
Due Installment (X_{10})	0.233	***	0.173	***	0.098	***	0.066	***

Due installments are among the most influential variables, with each additional due installment raising the odds of A1 by about 26.3%. Other financial indicators, such as interest due and charge due, are also strong predictors, especially in categories B0 through D0, indicating that accumulation of unpaid amounts plays a critical role in default classification. Although primary loan amount granted (X_2) is statistically significant in some categories, the effect size is negligible, implying limited practical relevance. The odds ratios derived from the model help quantify these impacts, highlighting the importance of payment behavior, outstanding liabilities, and loan duration in predicting loan status transitions. Some predictor variables (e.g., Loan Age and Interest Rate) are time-dependent. In our study, they were measured at a defined observation point for each loan. Therefore, the study does not capture full temporal dynamics (i.e., it is not longitudinal) and is considered cross-sectional. This design allows the models to use these features without violating independence assumptions. Although the dataset contains 16 variables, not all were used in the analysis; for example, variables such as occupation and bank location have many categories in the contingency tables, which would complicate modeling due to numerous missing values in some categories.

The confusion matrix in Table 2 provides insights into the classification performance of the multinomial logistic regression model across the five loan status categories: A0, A1, B0, C0, and D0. The model performs well in correctly classifying most A0 and D0 cases, with 5,658 true positives for A0 and 2,245 for D0, indicating strong predictive power for these two categories. However, misclassifications are evident, particularly between A1 and A0, where 275 actual A1 cases are incorrectly classified as A0, suggesting some overlap or similarity in predictor patterns between these categories. Additionally, a moderate number of D0 cases are misclassified as A0 (36) or A1 (29), which may point to shared characteristics with lower-risk groups. The model struggles the most with minority classes like B0 and C0, where correct classifications are extremely limited (only 1 for B0 and none for C0), and most instances are misclassified as D0 or other categories. These minority classes, such as B0 and C0, suffer from severe misclassification, indicating unresolved challenges with class imbalance and overlapping predictor patterns. This imbalance highlights the model's difficulty in handling underrepresented classes, likely due to class imbalance in the dataset. Overall, the model demonstrates high accuracy for dominant categories.

Table 2: Confusion matrix

	Actual				
Predicted		A0	A1	B0	C0	D0
	A0	5658	275	109	39	38
	A1	29	89	16	33	11
	B0	2	0	1	0	0
	C0	0	1	3	0	2
	D0	36	29	26	87	2245

The RF model demonstrated exceptional capability in binary classification, achieving a precision-recall AUC of 0.983 and 97.8% balanced accuracy on 20% of testing data. Variable importance analysis (Fig. 2) strongly aligned with MLR findings, identifying: Loan Age (mean decrease accuracy = 42.7) as the most influential predictor, interest rate trajectory patterns (36.2), and recovery delay duration (29.8) as key risk indicators. The model exhibited a conservative bias (133 false positives vs 407 false negatives), making it particularly suitable for risk-averse early warning systems where minimizing missed delinquencies is prioritized (Table 2). The Random Forest model exhibited exceptionally high predictive performance, achieving perfect classification accuracy (100%) on the 20% test set, correctly identifying all 3,606 Default and 5,123 Non-Default cases without a single misclassification. While this result suggests remarkable model performance, it raises critical concerns about overfitting, or an overly deterministic definition of default that may artificially simplify the classification task. The rule-based construction of the target variable (Loan Status)—which includes strict thresholds such as three or more due installments, loans that are fully outstanding, or loans overdue for more than 12 months—may have created a clear-cut separation between Default and Non-Default cases, limiting real-world variability.

The ROC curve analysis further supports this observation, with Area Under the Curve (AUC) scores ranging from 0.917 to 0.995 across all classes, indicating near-perfect separability, especially for classes like D0 (AUC = 0.995) and A0 (AUC = 0.986) (Fig. 1). The variable importance plot reveals that *Loan Age* and *Due Installments* are the most influential predictors, with Gini importance values significantly higher than others, followed by *Price Due*, *Charge Due*, and *Interest Due*. Demographic and less directly default-linked features like *Gender* and *Interest Rate* contribute minimally. Overall, although the model demonstrates strong internal performance, its lack of misclassifications, extreme AUC values, and reliance on a rule-derived target variable suggest the need for external validation on more nuanced, real-world data where defaults and non-defaults exhibit more overlap and uncertainty.

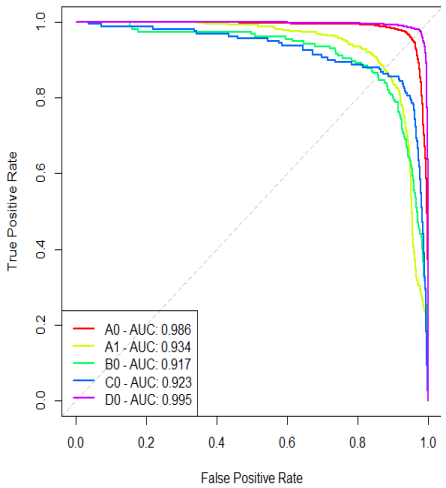

Figure 1: ROC curves

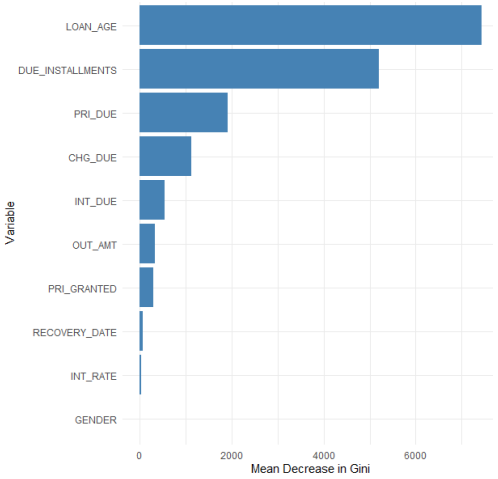

Figure 2: Variable importance of RF

CONCLUSIONS/RECOMMENDATIONS

This study advances both academic research and banking practice by establishing temporal dynamics as critical predictors of loan delinquency and demonstrating the value of hybrid modeling approaches. The findings provide a robust framework for handling hierarchical delinquency categories while offering actionable insights for risk management. Key theoretical contributions include the identification of non-linear interest rate effects, loan age thresholds, and seasonal patterns that shape delinquency risk. The comparative analysis reveals that while Random Forest excels in operational screening, multinomial logistic regression offers superior interpretability for risk progression, highlighting the benefits of a combined approach. We acknowledge that class imbalance is a limitation. To address this, we plan to apply resampling or class weighting and evaluate model performance using the previously mentioned metrics, such as the confusion matrix, AUC, and precision, in future work. For practitioners, we recommend tiered

monitoring systems based on the 12-month loan age threshold and dynamic pricing models that account for the 15% APR risk cliff. Regulatory bodies should incorporate these findings into stress-testing frameworks and develop standards for model interpretability in credit risk assessment.

REFERENCES:

Central Bank of Sri Lanka. (2023). Annual report 2023. Central Bank of Sri Lanka.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1–26.
Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.
Ripley, B. D. (2002). nnet: Feed-forward Neural Networks and Multinomial Log-Linear Models. R package version 7.3-1.
Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(77), 1–8.
Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.