Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 01 January-2026, Page No.-8665-8668

DOI: 10.47191/etj/v11i01.31, I.F. – 8.482

Loan Default Risk Classification Using Random Forest Based on Monthly Loan Performance Data

Fitria[1], Emy Iryanie[2], Heldalina[3], Muhammad Syahid Pebriadi[4], Anhar Khalid[5]

[1,2,3,4,5]Politeknik Negeri Banjarmasin, Jalan Brigjen Hasan Basri, Pangeran, Kecamatan Banjarmasin Utara, Kota Banjarmasin, Kalimantan Selatan, Indonesia

ABSTRACT: The risk of loan default is a major challenge in the financial industry because it has a direct impact on the stability of lending institutions. As loan performance data becomes increasingly available, machine learning-based approaches are becoming a potential alternative to improve the accuracy of credit risk assessments. This study aims to analyze the performance of the Random Forest model in classifying default risk based on debtors' monthly performance data. The dataset used consists of more than one million observational records with variables representing credit characteristics and payment behavior. To avoid data leakage, only variables that were available before the default occurred are used. The model evaluation was carried out using Area Under the Curve (AUC) and confusion matrix metrics, accompanied by threshold adjustments to increase the sensitivity of the detection of risky debtors. The results showed that the Random Forest model achieved an AUC value of 0.737, which indicates a good discriminating ability. Threshold adjustments increase the detection of risky debtors with the consequence of increasing false positives, so this approach has the potential to support decision-making in credit risk management.

KEYWORDS: Credit risk, default, machine learning, random forest, monthly performance data

I. INTRODUCTION

The risk of loan default is a crucial problem in credit management, especially for financial institutions operating in an increasingly competitive and dynamic business environment (1–4). The debtor's inability to meet payment obligations not only has the potential to cause financial losses but can also affect the operational stability and sustainability of financial institutions. Therefore, an accurate and adaptive credit risk assessment process is becoming an increasingly important need.

Conventional approaches to credit risk assessment generally rely on classic statistical methods and rule-based rules. Although these methods are relatively easy to implement, they often have limitations in capturing the non-linear and complex patterns found in large-scale credit data. Along with the development of information technology and the increasing volume of historical data on loans, machine learning methods are becoming widely used as an alternative to improve the quality of credit risk prediction.

Machine learning allows for more complex modeling of the relationship between credit variables and default risk, especially when the data is imbalanced, where the number of defaulting debtors is much lower than non-defaulting debtors (5–8). However, the application of machine learning in the context of credit risk also faces a few challenges, such as the potential for data leakage and the selection of inappropriate evaluation metrics. Over-perfect predictions often indicate leaked information from variables that directly represent default status, so the model does not really learn from the relevant risk patterns.

In addition, many previous studies still focused on the level of accuracy as a key indicator of model performance. In the context of credit risk, accuracy metrics alone are not representative enough, as errors in classifying defaulting debtors as non-at-risk debtors can have significant financial impacts. Therefore, an evaluation approach that emphasizes more on the model's ability to detect minority classes, such as the use of Area Under the Curve (AUC) and confusion matrix analysis.

Based on these problems, this study aims to analyze the performance of machine learning models in classifying loan default risk using debtors' monthly performance data. This study uses the Random Forest model with feature selection that considers the availability of information before default occurs. In addition, adjustments were made to the classification threshold to increase the sensitivity of the model to risky debtors. The results of the research are expected to contribute to the development of a more adaptive and realistic decision support system in credit risk management.

In this study, Random Forest was chosen as the main model because of its ability to handle non-linear relationships, resilience to overfitting, and its good performance on large-scale and unbalanced credit risk data.

II. RESEARCH METHODOLOGY

A. Research Design

This study uses a quantitative approach with a computational experiment method to classify the risk of loan default using machine learning. The research stages include pre-processing of data, feature selection, model training, and model performance evaluation.

B. Datasets and Data Sources

The dataset used in this study was obtained from the open data platform Kaggle (9), which provides data on the performance of debtors' monthly loans. The dataset is used as the basis for analysis in the development of a loan default risk classification model. The dataset used is in the form of monthly loan performance data consisting of 1,048,575 observations with 13 variables. The target variable is has_defaulted, which represents the debtor's default status.

C. Data Pre-processing

Data pre-processing includes numerical data type conversion, missing value handling using median imputation, and data normalization with StandardScaler. This step aims to improve the data quality and stability of the model training process.

D. Feature Selection

To avoid potential data leakage, this study only uses features that were available before the default occurred. The features used include months_on_book, credit_score, credit_score_delta_3m, revolving_utilization, revolving_util_delta_3m, and scheduled_emi.

E. Modeling and Evaluation

This study focuses on the analysis of the Random Forest model as the main approach in the classification of loan default risk based on monthly performance data (10–12). The data was divided into training data and test data with an 80:20 ratio using stratified sampling. The model's performance evaluation was conducted using the Area Under the Curve (AUC) and the confusion matrix, with a primary focus on the default class recall. In addition, classification threshold adjustments were made to increase the sensitivity of the model in detecting at-risk debtors.

F. Research Flow

In general, the research flow starts from raw data processing, selection of secure features, training of Random Forest models, to evaluation of model performance based on metrics relevant to the context of credit risk as shown in figure 1.

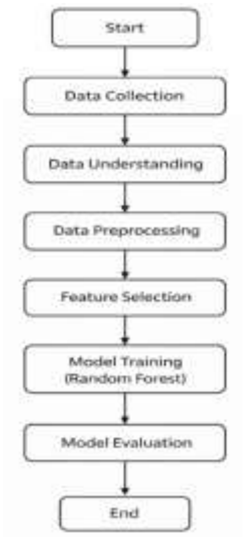

Figure 1. Research flow of loan default risk classification using machine learning

III.RESULTS AND DISCUSSION

A. Evaluate Model Performance Using ROC Curves

The initial evaluation of the performance of the Random Forest model was carried out using the Receiver Operating Characteristic (ROC) curve to assess the model's ability to distinguish debtors who have the potential to default and those who do not. The ROC curve describes the relationship between true positive rates and false positive rates at various classification threshold values, thus providing a comprehensive picture of the model's discriminating capabilities independently of a given threshold value (13,14).

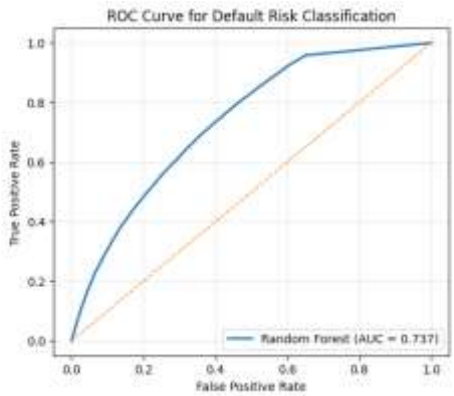

Figure 2. The ROC curve of the Random Forest model in the classification of loan default risk is based on monthly performance data.

Figure 2 shows the ROC curve of the Random Forest model resulting from the test data. Based on the test results, the model achieved an Area Under the Curve (AUC) value of 0.737. The AUC value indicates that the model has a good discriminating ability in classifying the risk of loan default. In general, an AUC value above 0.7 can be categorized as acceptable performance in the context of credit risk classification, especially in datasets with an unbalanced class distribution.

The AUC value obtained also shows that the machine learning approach based on monthly performance data can capture the initial pattern of credit risk more effectively than the random approach. In addition, a moderate and not too high AUC value reflects that the model does not experience overfitting and does not utilize leaked information (data leakage), so the results obtained are realistic and scientifically accountable.

B. Confusion Matrix Analysis and Threshold Adjustment

To analyze the impact of classification decisions in more detail, the evaluation was continued using a confusion matrix (14,15). At this stage, the classification threshold was adjusted from the default value of 0.5 to 0.2. The threshold adjustment aims to increase the sensitivity of the model in detecting defaulting debtors, considering that false negative errors (defaulting debtors who are predicted to be not at risk) have greater financial implications than false positive errors.

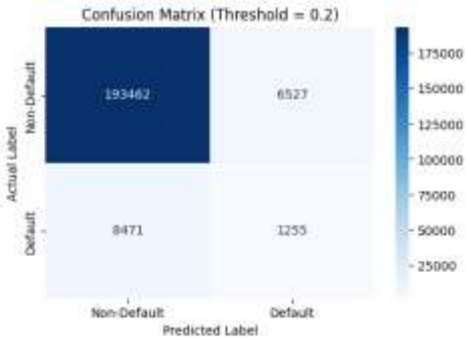

Figure 3. Confusion matrix of the Random Forest model with a threshold of 0.2 in the classification of loan default risk.

Figure 3 shows the confusion matrix of the Random Forest model with a threshold of 0.2. Based on these results, the model managed to correctly classify most non-defaulting debtors, which is indicated by a high true negative value. In addition, the model is also able to detect a few defaulting debtors, although there are still risky debtors who have not yet been identified.

The results of the evaluation showed that the threshold adjustment increased the recall value in the default class to around 0.13, with a precision value of 0.16. This increase in recall suggests that the model is becoming more sensitive in detecting at-risk debtors, although the consequence is an increase in the number of false positives. In the context of credit risk management, this condition is acceptable because it is better to identify risky debtors excessively than to fail to detect debtors who have the potential to cause losses.

C. Discussion and Implications

The results show that the Random Forest model with a selection of features that avoid data leakage can provide realistic classification performance on large-scale credit risk data that is in line with previous research (16–19). The use of AUC as the primary metric provides a more representative picture than accuracy alone, especially under unbalanced data conditions. The relatively high accuracy in this study is not used as the main indicator, because it can be misleading in the context of the dominance of the non-default class.

Classification threshold adjustments have proven to be an important strategy in improving the practical usability of the model. By lowering the threshold, the model becomes more responsive to the default class, making it more suitable for use as a decision support tool in credit risk assessment systems (20–23). Although the default class recall is still moderate, these results reflect real challenges in predicting credit risk based on historical data and open opportunities for further development.

Overall, the results of this study confirm that a machine learning approach based on monthly performance data has the potential to be applied in credit risk management systems. However, the model's performance can still be improved through the exploration of additional features, the use of more explicit temporal approaches, or the application of further data balancing techniques in subsequent studies.

IV. CONCLUSIONS

This study has analyzed the performance of the machine learning model in classifying loan default risk based on debtors' monthly performance data. By using the Random Forest model and feature selection that avoids potential data leakage, this study produces a classification model that is realistic and scientifically accountable.

The results of the evaluation showed that the Random Forest model was able to achieve an Area Under the Curve (AUC) value of 0.737, which indicates a good discriminating ability between defaulting and non-defaulting debtors. Adjustment of classification thresholds has been shown to increase the sensitivity of the model in detecting at-risk debtors, although it is followed by an increase in false positive errors. This approach is considered more appropriate in the context of credit risk management, where errors in identifying defaulting debtors have greater financial implications.

Overall, the results of this study show that a machine learning approach based on monthly performance data has the potential to be used as a decision-making support tool in credit risk assessment systems. The developed model can help financial institutions in monitoring credit risk in a more adaptive and data-driven manner. Further research is suggested to explore a more explicit temporal approach, the use of additional features, as well as the application of data balancing methods to improve the performance of default risk detection.

ACKNOWLEDGMENT

The author would like to thank Politeknik Negeri Banjarmasin for the support and facilities provided so that this research can be carried out and completed properly. The author also expressed his appreciation to the manager of the Kaggle platform who has provided the dataset as an open data source used in this study. In addition, the author would like to thank colleagues for their input and discussions that contributed to the refinement of the research and writing of this article.

REFERENCES

1. Xu J, Lu Z, Xie Y. Loan Default Prediction of Chinese P2P Market: A Machine Learning Methodology. Scientific Reports. 2021.
2. Avgeri E, Psillaki M. Factors Determining Default In P2P Lending. Journal of Economic Studies. 2023.
3. Miled KBH, Landolsi M. Risk Assessment for Reimbursement of Microfinance Institutions. International Journal of Professional Business Review. 2023.
4. Shi S, Tse R, Luo W, d'Addona S, Pau G. Machine Learning-Driven Credit Risk: A Systemic Review. Neural Computing and Applications. 2022.
5. Abdullah M, Chowdhury MAF, Uddin A, Moudud-Ul-Huq S. Forecasting Nonperforming Loans Using Machine Learning. Journal of Forecasting. 2023.
6. Rao C, Liu Y, Goh M. Credit Risk Assessment Mechanism of Personal Auto Loan Based on PSO-XGBoost Model. Complex \& Intelligent Systems. 2022.
7. Zhang Y, Jian-xiong L, Qian Y, Shu L. Machine Learning and Financial Inclusion: Evidence From Credit Risk Assessment of Small-Business Loans in China. The Singapore Economic Review. 2025.
8. Chen YR, Leu J, Huang SA, Wang JT, Takada J. Predicting Default Risk on Peer-to-Peer Lending Imbalanced Datasets. Ieee Access. 2021.
9. Bank Customer payment behavior [Internet]. [cited 2026 Jan 17]. Available from: https://www.kaggle.com/datasets/deepak915/bank-customer-payment-behavior
10. Doko F, Kalajdziski S, Mishkovski I. Credit Risk Model Based on Central Bank Credit Registry Data. Journal of Risk and Financial Management. 2021.
11. Loef B, Wong A, Janssen N, Strak M, Hoekstra J, Picavet HSJ, et al. Using Random Forest to Identify Longitudinal Predictors of Health in a 30-Year Cohort Study. Scientific Reports. 2022.
12. Li G, Ma H, Liu R, Shen M, Zhang K. A Two-Stage Hybrid Default Discriminant Model Based on Deep Forest. Entropy. 2021.
13. Wang G, Lu C, Solomon OM, Gu Y, Ling Y, Xu F, et al. Construction and Evaluation of a Machine Learning-Based Predictive Model for Enteral Nutrition Feeding Intolerance Risk in ICU Patients. Frontiers in Nutrition. 2025.
14. Walther F, Heinrich L, Schmitt J, Eberlein-Gonska M, Roessler M. Prediction of Inpatient Pressure Ulcers Based on Routine Healthcare Data Using Machine Learning Methodology. Scientific Reports. 2022.
15. Jauhiainen S, Kauppi J, Krosshaug T, Bahr R, Bartsch J, Äyrämö S. Predicting ACL Injury Using Machine Learning on Data From an Extensive Screening Test Battery of 880 Female Elite Athletes. The American Journal of Sports Medicine. 2022.
16. Ghosh P, Azam S, Jonkman M, Karim A, Shamrat FMJM, Ignatious E, et al. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. Ieee Access. 2021.
17. Xie X, Zhang J, Luo Y, Gu J, Li Y. Enterprise Credit Risk Portrait and Evaluation From the Perspective of the Supply Chain. International Transactions in Operational Research. 2023.
18. Feng J, Ye J, Qi GF, Hong L, Wang F, Liu S, et al. A Comparative Analysis of Eight Machine Learning Models for the Prediction of Lateral Lymph Node Metastasis in Patients With Papillary Thyroid Carcinoma. Frontiers in Endocrinology. 2022.
19. Tan J, Huang ES, Yang H, Wan H, Zhang Q. Risk Factors Associated With Severe Progression of Parkinson's Disease: Random Forest and Logistic Regression Models. Frontiers in Neurology. 2025.
20. Ko PC, Lin PC, Do HT, Huang YF. P2P Lending Default Prediction Based on AI and Statistical Models. Entropy. 2022.
21. Lyócsa Š, Vašaničová P, Misheva BH, Vateha MD. Default or Profit Scoring Credit Systems? Evidence From European and US Peer-to-Peer Lending Markets. Financial Innovation. 2022.
22. Nguyen L, Ahsan M, Haider J. Reimagining Peer-to-Peer Lending Sustainability: Unveiling Predictive Insights With Innovative Machine Learning Approaches for Loan Default Anticipation. Fintech. 2024.
23. Hvidberg KB. Field of Study and Financial Problems: How Economics Reduces the Risk of Default. Review of Financial Studies. 2023.