Journal of Informatics and Web Engineering
Vol. 4 No. 3 (October 2025) eISSN: 2821-370X

Loan Default Prediction Using Machine Learning Algorithms

Zhi Zheng Kang[1], Sin Yin Teh[2]*, Samuel Yong Guang Tan[3], Wei Chien Ng[4]
[1,2,4]School of Management, Universiti Sains Malaysia, E43, Jalan Sasaran, Minden Heights, 11800 Gelugor, Pulau Pinang, Malaysia
[3]Department of Accountancy and Business at Tunku Abdul Rahman University of Management and Technology (TARUMT), Penang Branch, 77, Lorong Lembah Permai 3, 11200 Tanjung Bungah, Penang, Malaysia
corresponding author: (tehsyin@usm.my; ORCiD: 0000-0001-9439-4407)

Abstract – Financial institutions constantly face at the risk of default by borrowers which can result in significant financial losses. It is essential to develop an appropriate predictive model for loan default to reduce these risks and minimise financial losses. The objective of this study is to identify the most suitable machine learning model to predict loan default by comparing four models which are Random Forest, Decision Tree, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). Additionally, it also examines the key features influencing loan default prediction. The dataset used in this study is sourced from Kaggle and it consists of 148,670 rows with 34 features. As class imbalance is common in the model prediction, Synthetic Minority Over-sampling Technique (SMOTE) is applied during model training to enhance predictive performance. Model performance is evaluated using five significant assessment metrics: accuracy, precision, F1-score, recall, and the area under the receiver operating characteristic curve (ROC AUC). The outcomes indicate that LightGBM performs the best among the other models with the highest accuracy (0.9764), in addition to precision (0.9747) and recall (0.9503) scores. Feature importance analysis is conducted by using permutation importance. It identifies interest, credit type, interest rate spread, and upfront charges as the four most significant features of loan default. These findings provide useful information for financial institutions aiding risk assessment and decision-making to mitigate potential losses.

Keywords— Financial Inclusion, LightGBM, Loan Default, Machine Learning, XGBoost

Received: 9 March 2025; Accepted: 19 June 2025; Published: 16 October 2025

1. INTRODUCTION

Individuals rely on financial institutions for loans to address financial challenges, achieve personal goals or manage unforeseen expenses. Loans provide financial support that enables people to achieve their objective for purchasing a house, funding education, starting a business or paying off deft. Loan borrowing has become an essential economic activity in a dynamic and highly competitive financial landscape. Lending serves as a key revenue that generates business for financial institutions. However, despite its mutual benefits, it also has significant risks.

Published by MMU Press. URL: journals.mmupress.com/jiwe

According to [1], the global lending market has experienced dramatic growth and is projected to reach $2,165.09 billion in 2025. By 2029, it is expected to reach $115,985.39 billion with a compound annual growth rate (CAGR) of 7.1%. The strong economic conditions, business expansion opportunities, and global business operations are all driving more demand for lending services [2]. The globalization of business activities has further created new market opportunities for enterprises. Furthermore, financial technology innovation such as digital lending platforms and artificial intelligence (AI) driven credit evaluations have enhanced accessibility to loans for both individuals and businesses.

The primary risk associated with lending arises when borrowers fail to repay their loans on time. Credit risk assessment is used to evaluate a borrower's creditworthiness and is crucial to minimise the risk. The possibility of a borrower defaulting either partially or fully in the repayment of the loan and incurring losses for the lending institution is known as a credit risk. Financial institutions employ the 5Cs framework that is Character, Capital, Capacity, Collateral, and Conditions, to evaluate borrower characteristics and analyse the probability of loan default [3]. This approach aims to mitigate the risk associated with loan repayment. However, this approach is significantly dependent on the expertise and banking professionals' experience and makes evaluation process time-consuming. In addition, there is no absolute guarantee that approved applicants will fulfil their repayment obligations.

Before the adoption of AI, traditional credit risk management encountered several limitations with insufficient data analysis being a major challenge [4]. Banks and other financial institutions commonly relied on manual procedures and small datasets to assess creditworthiness and make lending decision which led to this issue. These traditional approaches used to utilise basic credit scoring models, repayment histories, and basic financial ratios that failed to capture the full scope of risk factors associated with the borrowers. Besides, the conventional manual credit assessment procedures were time-consuming and prone to human error and thus adding to lending inconsistency. Inadequate analysis of data increased the likelihood that risk evaluation would be erroneous or incomplete which had the potential to result in suboptimal credit assessment.

This research utilises various machine learning models to identify the optimal performing model in predicting loan defaults to predict loan default accurately. By evaluating the performance of different models, this study aims to define the optimal process for financial institutions to examine and control the risk of loan defaults. Other than model choice, this study also identifies important features that significantly impact the likelihood of loan default. Financial institutions can then understand these critical factors in order to select high-risk borrowers more effectively and enhance their risk assessment strategies. Lastly, the results and implications of this research will assist in minimizing financial losses and improving decision-making in lending, leading to more efficient and trustworthy lending schemes.

The remainder of this paper is structured into the following. Section 2 contains a literature review of current works on loan default prediction and machine learning in risk assessment in finance. Section 3 outlines the research methodology, such as the data sources, data preprocessing techniques, and machine learning models that have been used in this study. In addition, it explains the assessment measures taken to measure model performance. Section 4 displays results and discussion after machine learning results have been interpreted and compared. It further displays the interpretation of significant features. The study concludes and summarises at last in Section 5 that also states its limitations and indicates the need for potential future research to be conducted.

2. LITERATURE REVIEW

2.1 Loan Default Prediction

Loan default prediction uses past data to predict whether a borrower will be unable to repay a loan. It is important for financial institutions to reduce the losses because loan defaults are highly correlated to profitability [5]. For example, the 2008 financial crisis was significantly influenced by extensive lending to individuals and businesses that were unable to meet their loan obligations [6]. In recent years, financial institutions have increasingly used machine learning methods to automate loan default prediction and have significantly enhancing both accuracy and efficiency. Now, financial institutions can assess the risk of the borrower with higher accuracy and reduce the likelihood of approving a loan to a high-risk individual by using advanced algorithms such as decision trees, ensemble models, and deep learning. Moreover, the rise of online shopping and mobile payments has enabled financial institutions to collect large amount of data to improve the capabilities of prediction. This large amount of real time transactional data allows financial models to adapt dynamically. It also able to identify risks and market trends more effectively. Financial institutions can thus improve their overall financial stability by implementing proactive risk management strategies.

2.2. Supervised Machine Learning in Loan Default Prediction

Supervised learning for binary classification is commonly used in loan default prediction. Binary classification categorises new observations into one of two classes and is used to train labelled data in a binary format, such as true/false or positive/negative. In this study, binary classification algorithms are used to forecast whether a borrower will default or successfully repay the loan. Given their widespread application in binary classification tasks, four supervised learning algorithms which are Decision Tree, Random Forest, XGBoost, and LightGBM, are discussed in the following sub-sections based on existing research studies. This study provides practical insights into the four model performances and appropriateness for credit risk assessment using a consistent workflow and relevant financial data.

2.2.1 Decision Tree

Decision Tree splits data into smaller subsets iteratively based on its attributes and this process continues until a particular stopping criterion is fulfilled [7]. According to [8], Decision Tree algorithm consists of inner nodes that represent branching structures, datasets that inform the decisions made by the algorithm, and leaf nodes that indicate final outcomes. The algorithm comprises two types of nodes: (1) decision nodes which make decisions and have multiple branches, and (2) leaf nodes which represent final outputs and do not branch further.

Decision Trees able to handle large datasets effectively due to their partitioning mechanism which systematically divides the dataset into smaller segments [7]. However, their applicability is often limited to straightforward attribute (value data) which may require modifications for more complex scenarios. In the study by [9], tree-structured methodologies were found to be the most effective due to their strong generalization capabilities and interpretability. Specifically, those incorporating boosting techniques and decision trees. Similarly, [10] used Naïve Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbours, and Random Forest for loan default prediction using a dataset from a European peer-to-peer (P2P) lending platform comprising 220,906 records and 112 features and the results show that Random Forest is the best machine learning model.

2.2.2 Random Forest

Random Forest is a versatile algorithm that is workable to both classification and regression problems. It integrates multiple classifiers to enhance model performance and address complex predictive challenges based on the concept of ensemble learning. It builds multiple decision trees and acts as a meta estimator by applying these trees to different subsets of the dataset during the training process [11]. Random Forest is a fast and effective model for handling large and imbalanced datasets. However, it often struggles to train effectively on diverse datasets, especially in regression tasks [12].

In a study by [13], they used XGBoost and Random Forest algorithms to develop a loan default prediction model using a dataset from Imperial College London which consists of 105,471 records and 778 features. The study used the variance threshold method for the elimination of features with low importance. The variance inflation factor helped with the assessment of multicollinearity within the data. The results showed that Random Forest had an accuracy of 0.90657 while XGBoost attained 0.90635. It can be concluded that the performance variation between both models stayed small and both algorithms suited loan default forecasting.

2.2.3 XGBoost

XGBoost is a boosting-based tree algorithm widely used due to its optimizations in four key areas: (i) a distributed weighted square graph method for segmentation point selection, (ii) enhanced handling of sparse data, (iii) an efficient cache-aware block data storage structure, and (iv) enhanced utilization of parallel and distributed computing [14]. XGBoost builds trees by splitting leaves within the same layer, reducing the likelihood of overfitting. However, excessive leaf nodes can lower the splitting gain and introduce additional computational overhead. Furthermore, extensive parameter tuning is required during model training, requiring continuous optimization of hyperparameters [15].

In a study by [16], Decision Tree had the highest precision but the lowest Area Under the Curve (AUC) while Random Forest achieved the highest accuracy but the lowest recall. XGBoost was the best model because it demonstrated the highest recall and AUC. Similarly, [17] employed Logistic Regression, Random Forest, XGBoost, and Adaptive Boosting (AdaBoost) to analyse the application of loan default prediction through a dataset of borrowers from Kaggle. Their results indicated that XGBoost achieved the highest accuracy (93.26%) whereas Logistic Regression had the lowest accuracy (81.20%).

2.2.4 LightGBM

LightGBM is a gradient-boosting decision tree technique that identifies segmentation points and selects features for decision trees using the histogram algorithm [18, 19]. LightGBM uses gradient-based one-sided sampling (GOSS) to prioritise training on samples with low prediction performance while reducing the sample size in each iteration. LightGBM applies exclusive feature bundling to further lower computational complexity which groups mutually exclusive features within the data. This approach is particularly effective because most datasets are sparse and certain features exhibit mutual exclusivity. In other words, their non-zero values do not appear simultaneously.

In a study by [20], Decision Trees, XGBoost, LightGBM, and Logistic Regression were used to predict loan default. Out of the entire sample, 80% of the data is chosen at random to serve as the training set, with the remaining portion serving as the test set. The findings indicated that both XGBoost and LightGBM were more predictive than Logistic Regression and Decision Trees, with LightGBM slightly outperforming XGBoost. Cross-validation results demonstrated that LightGBM exhibited stable performance across training and test sets with minimal variation in accuracy and AUC. Additionally, the findings suggested that no overfitting was observed.

3. RESEARCH METHODOLOGY

The dataset is collected from Kaggle through the link of https://www.kaggle.com/datasets/yasserh/loan-default-dataset. It consists of 148670 rows and 34 columns of data. The details regarding the features and data type are presented in Table 1. The process of model building consists of some steps, and these steps are conducted accordingly based on the flow chart in Figure 1. The dataset contains some missing values, and certain unnecessary columns are removed. These preprocessing steps are discussed in Section 3.1.

Table 1. Feature Description and Data Type

Features	Description	Data Type
ID	Client loan application ID	integer
Year	Year of loan application	integer
Loan_limit	Conforming of loan status	string
Gender	Gender	string
Approv_in_adv	Loan preapproval status	string
Loan_type	Type of loan	string
Loan_purpose	Purpose of loan	string
Credit_Worthiness	Credit worthiness	string
Open_credt	Whether the applicant has any open credit accounts	string
Business or commercial	Whether the loan is for business/ commercial or personal purposes	string
Loan_amount	Amount of money being borrowed	integer
Rate_of_interest	Interest rate charged on the loan	integer
Interest_rate_spread	Different between on interest rate on the loan and a benchmark interest rate	integer
Upfront_charges	Initial charges associated with securing the loan	integer
Term	Duration of the loan in months	integer
Neg_ammortization	Whether the loan allows for negative amortization	string
Interest_only	Whether the loan has an interest-only payment option	string
Lump_sum_payment	Whether a lump sum payment is required at the end of the loan term	string
Property value	Value of property being financed	integer

Construction_type	Type pf construction	string
Occupancy_type	Type of occupancy	string
Secured_by	Type of collateral securing the loan	string
Total_units	Number of units in the property being financed	string
Income	Applicant's annual income	integer
Credit_type	Applicant's type of credit	string
Credit_score	Applicant's credit score	integer
Co-applicant_credit_type	Co-applicant's type of credit	string
Age	Age of applicant	string
Submission_of_application	How the application was submitted	string
LTV	Loan-to-value ratio	integer
Region	Geographic region where the property is located	string
Security_Type	Type of security or collateral backing the loan	string
Dtir1	Debt-to-income ratio	integer

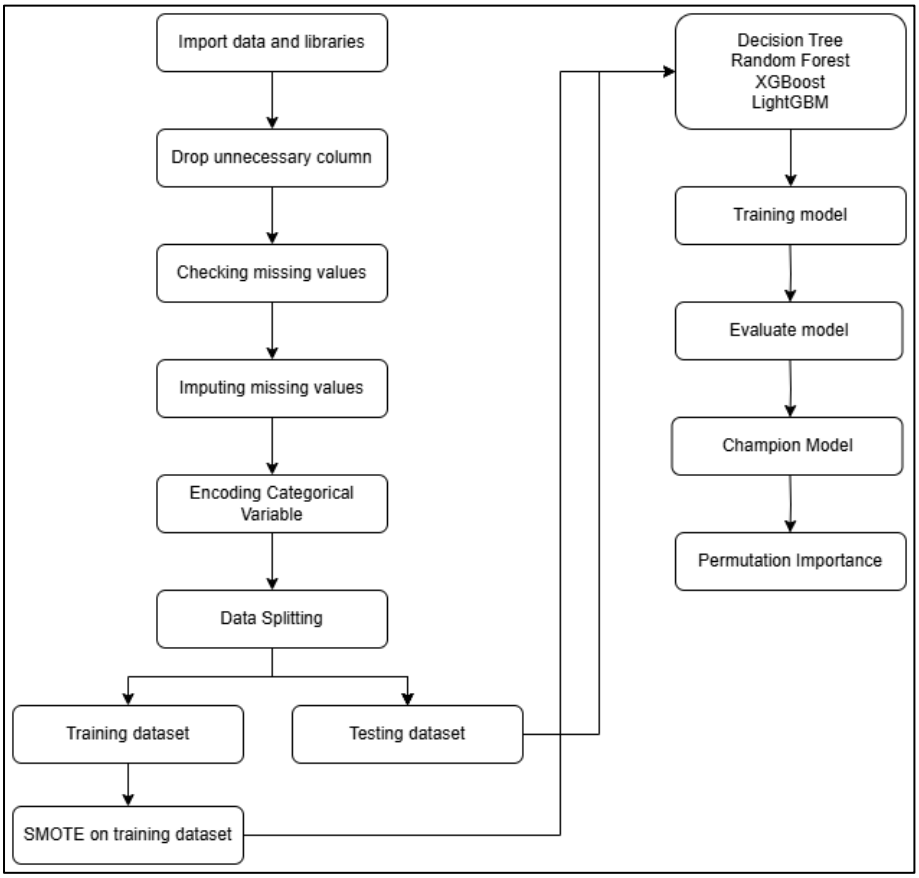

Figure 1. Flow Chart of the Machine Learning Workflow

3.1 Data Preprocessing

A crucial stage in data preparation is data preprocessing which converts raw data into a clear and useful format to make sure the data is reliable, consistent, and prepared for additional analysis or modelling. It entails data transformation, data integration, and data cleaning. It will focus on reducing noise in the data through the optimization of number of features through techniques like feature selection or transformation, and it also manage missing values using imputation or removal and addressing outliers and normalizing data values through scaling transformation [21].

In the study by [22], K-Nearest Neighbours Imputer (KNNImputer) is utilised to address missing numerical values. KNNImputer is a common technique that used to impute missing values and is often preferred over conventional imputation methods. It estimates and fills in missing values based on the nearest neighbours and identified through the Euclidean distance matrix, prioritizing non-missing values while disregarding missing ones [23]. While for missing categorical data, the mode imputation technique is used to replace missing values with the most frequently occurring categorical values within the dataset and ensures minimal disruption to data distribution.

3.2 Data Splitting

In machine learning, data splitting is an essential step to ensures that the models are effectively trained and tested. In this study, data splitting is performed using train and test with 80:20 ratio [24]. That is, 80% of data is utilised to train the machine learning models while 20% is reserved for testing purposes. The model learns patterns, relationships and decision rules from the training data during training. The test set is then utilised to evaluate the model's performance so that it will generalise well to new data. Overfitting can be prevented and ensure that the model is able to make valid predictions on unseen data rather than memorise the training data when the data is split appropriately. Furthermore, it is a balanced evaluation and allows researchers who can optimise models to deliver better performance when the train-test ratios are maintained in an appropriate way.

3.3 Synthetic Minority Over-sampling Technique (SMOTE)

A data augmentation technique called SMOTE has been employed extensively to handle class imbalance in datasets which is a prevalent concern in machine learning classification issues. Imbalanced data arises when a class is significantly larger or smaller than other classes and results in biased model predictions where the majority class is favoured while the minority class is often misclassified. The presence of imbalanced data can negatively impact the ability of models to generalise as the model may fail to learn meaningful patterns from the minority class. Although it will have good accuracy, but it is a bad prediction for the minority class as it might overly focus on the majority class. This issue will cause the model built is not robust or reliable for all scenarios. SMOTE addresses this issue by creating synthetic data points by interpolating between existing minority instances and their k-nearest neighbours. This maximises the potential of the model to learn patterns in minority classes at the expense of reducing risk of overfitting, that typically occurs in random oversampling techniques [25].

3.4 Data Modelling

During the data modelling stage, the selected models are built and trained to learn patterns from the dataset to develop predictive models. The random state is set at 42 so that consistency splits of the dataset are achieved during different runs and experiments. This parameter enhances the reliability and replicability of outcomes to prevent changes in model performance as a result of random fluctuations in data partitioning.

Researchers can compare different models reasonably and tune hyperparameters with greater certainty by maintaining a consistent random state. Furthermore, training and test sets can be kept constant across iterations and enable reasonable approximation of the accuracy predicted by each model and how generalizable each model is by creating random state [26].

3.5 Permutation Feature Importance

Permutation feature importance is a conventional metric that is used to evaluate a feature's contribution to overall model performance [27]. It measures the decrease in predictive performance of a model when values of a specific feature are randomly shuffled and break the relationship between the feature and the target variable [28]. It is employed in this study once the beat machine learning has been determined. This is done by importing permutation_importance from sklearn.inspection. Features are ranked based on their mean importance scores. The greater the mean score, the more significant the contribution to the model's prediction. A lower mean score means little or no contribution. This

method provides an interpretable measure of feature relevance aiding in the model refinement by identifying the most significant variables.

4. RESULTS AND DISCUSSIONS

4.1 Model Evaluation

Model performance is evaluated to evaluate the efficiency of the model performance. Accuracy (1), Precision (2), recall (3), and F1 score (4) are used as performance measures in this study. Each measure provides different insights into how well the model is performing when handling loan default prediction classification tasks.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1\ Score = 2 \times \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

where True Positive (TP) represents a correctly identified defaulter, meaning the model successfully the loan default of a borrower. True Negative (TN) denotes a correctly identified non-defaulter indicating that the model successfully classifies a borrower who will repay the loan as not defaulting. False Positive (FP) occurs when a non-defaulter is incorrectly classified as a defaulter, which could lead to unnecessary loan rejections or higher interest rates for borrowers who are creditworthy. In contrast, False Negative (FN) happens when a defaulter is wrongly classified as a non-defaulter that can bring a significant risk to lenders. The financial institution might approve loans for borrowers who are likely to default. In financial applications, reducing FN is crucial because misclassifying high risk borrowers as low risk can lead to financial losses. It is necessary to balance these metrics to develop a reliable loan default prediction model.

The Receiver Operating Characteristic (ROC) curve is also used in this study to evaluate the model's performance. It is a graphical representation of a machine learning model's effectiveness, illustrating the relationship between the TP rate and the FP rate. The ROC curve helps analyse a model's trade-off between sensitivity and specificity across different thresholds It enables the comparison of multiple models on the same dataset to determine which performs better. The Area Under the ROC Curve (AUC) is a key metric used to evaluate model's overall performance in differentiating between positive and negative classes. The AUC value lies between 0 and 1. Better model performance is indicated by a larger value.

4.2 Comparison of Performance Metrics

The overall performance metrics of the models are demonstrated in Table 2. In addition, Figure 2 presents a line chart that visualises the comparative analysis of the machine learning models across 5 key performance metrics such as the accuracy, prevision, recall, F1 score and ROC AUC. Different machine learning models can perform differently such as poorly, well or excellent on the same task. This mainly due to the difference in their learning algorithms and architecture mechanisms [29]. Based on the results, XGBoost and LightGBM outperform Decision Tree and Random Forest. LightGBM achieves the highest accuracy (0.9764), precision (0.9747), and F1 score (0.9503) which indicate

that it has a strong overall classification effectiveness. Meanwhile, XGBoost records the highest recall (0.9307) and ROC AUC (0.9917) which suggest that it is slightly better at identifying actual defaulters and distinguishing between classes. These findings highlight the strong predictive capabilities of both XGBoost and LightGBM in loan default prediction.

Table 2. Performance Metrics of Machine Learning Algorithms

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Decision Tree	0.9408	0.8672	0.8939	0.8804	0.9250
Random Forest	0.9394	0.8933	0.8530	0.8727	0.9778
XGBoost	0.9760	0.9694	**0.9307**	0.9496	**0.9917**
LightGBM	**0.9764**	**0.9747**	0.9271	**0.9503**	0.9910

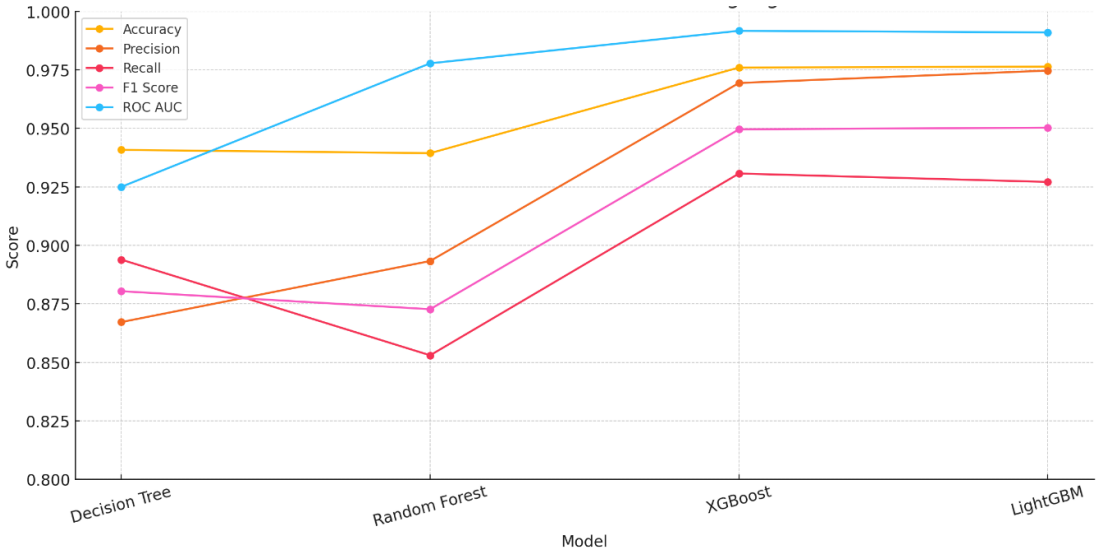

Figure 2. Line Chart of the Performance Metrics

Among the traditional tree-based models, Decision Tree has the lowest accuracy (0.9408) but maintains a relatively balanced performance across all metrics. Random Forest as an ensemble method improves upon Decision Tree's precision (0.8933) and achieves a much higher ROC AUC (0.9778). This showing that Random Forest can better differentiate between loan defaulters and non-defaulters. However, its recall (0.8530) is lower compared to XGBoost and LightGBM. This means XGBoost may fail to identify some defaulters.

Overall, the results suggest that XGBoost and LightGBM are the most effective models. LightGBM excelling in precision and F1 score, making it the best choice for loan default prediction in this study. However, if the primary goal is to maximise recall and identify as many defaulters as possible, XGBoost would be preferable due to its superior recall and ROC AUC score.

4.3 Champion Model

LightGBM was chosen as the champion model in this study since it performed better than other models, as presented in Table 2. Its accuracy, precision, and F1 score being high demonstrate its good predictive performance. It can make the most reliable model for accurately identifying loan defaulters. In addition, LightGBM's able to handle large datasets efficiently and its optimised gradient-boosting framework contributes to its effectiveness. Its faster training speed and lower computational cost make it ideal for real-world financial applications.

4.4 Feature Importance Analysis

From Table 3, the mean score for each variable in the LightGBM champion model is displayed in descending order. The results indicate that the four most significant features for loan default prediction are the interest rate, credit type, spread of interest rate and upfront charges, with mean scores of 0.283319, 0.073586, 0.008695, and 0.008281, respectively. Open credit is identified as the least important feature, with a mean score of -0.000011.

Table 3. List of the Feature with Feature Importance Score

Feature	**Score**
rate_of_interest	0.283319
credit_type	0.073586
interest_rate_spread	0.008695
upfront_charges	0.008281
business_or_commercial	0.005296
negative_amortization	0.003998
ltv	0.003950
submission_of_application	0.003275
lump_sum_payment	0.002848
loan_amount	0.002672
income	0.002044
term	0.001573
dtir	0.001280
property_value	0.001098
occupancy_type	0.000966
region	0.000887
co_applicant_credit_type	0.000343
approved_in_advance	0.000141
age	0.000132
interest_only	0.000113
credit_worthiness	0.000058
credit_score	0.000049
loan_limit	0.000010
construction_type	0.000009
total_units	0.000006
secured_by	0.000000
security_type	0.000000
open_credit	-0.000011

The rate of interest is the most significant feature because higher interest rates increase borrowers' repayment obligations. This making them more likely to miss payments and default. Credit type in the dataset refers to different scoring sources, each using distinct models and datasets to assess creditworthiness. Some financial institutions apply stricter credit reporting while others incorporate alternative credit data. Interest rate spread which reflects the gap between market rates and the borrower's interest rate, serves as an indicator of perceived risk. The higher spreads suggest lenders view the borrower as a greater risk, increasing default likelihood. Upfront charges such as processing fees or administrative costs can contribute to financial strain, further elevating default risk if borrowers struggle to manage these costs at loan initiation. In contrast, open credit is the least important predictor of loan default. Merely having open credit accounts does not necessarily indicate financial distress or an inability to repay a loan. The number of open credit accounts alone does not provide sufficient insight into credit utilization, debt levels, or repayment history, making it a less significant feature in predicting loan default.

5. CONCLUSION

In conclusion, the LightGBM model is identified as the most suitable model for predicting loan default, as it achieves the highest accuracy, precision, and F1 score. This indicates its superior predictive capability compared to other models. The results also highlight that the most influential feature in loan default prediction is the rate of interest, followed by credit type, interest rate spread, and upfront charges. Other significant features in descending orders include business or commercial use, negative amortization, loan-to-value ratio, submission of application, lump sum payment, loan amount, income, term, debt-to-income ratio, property value, occupancy type, and region. Moreover, co-applicant credit types, approved in advance, age, interest-only loans, creditworthiness, credit score, loan limit, construction type, total units, secured by, and security type also contribute to the prediction. The least important feature in determining loan default is open credit, suggesting that the presence of an open credit account alone does not strongly indicate default risk.

As demonstrated through the study results, LightGBM is an effective model to make loan default prediction and can be relied upon for use as a machine learning model in this purpose. Furthermore, the results obtained from the permutation feature importance provide financial institutions useful information regarding identification of those customers who have higher chances of loan default. These results can aid banks in having a better understanding of credit risk, improving their lending processes, and implementing proactive actions to lower default rates. As a result, enables financial institutions to optimise risk management, enhance financial stability, and maximise profitability.

While this study provides valuable insights into loan default prediction, it has certain limitations that can be addressed in future research. First, the dataset contains missing values with some features having up to 36,439 missing entries. If not handled properly, this could bias the analysis. Future studies can explore more advanced statistical techniques or machine learning-based imputation methods to improve data quality. Second, the dataset is imbalanced with a ratio of 112,031:36,639 between the majority and minority classes. Although SMOTE was applied in this study, alternative resampling methods such as SMOTE-Tomek, SMOTE-ENN, and KMeans-SMOTE could be investigated to enhance class balance further. Third, while this study uses various machine learning models, it does not look into ensemble techniques, which mix multiple models to improve performance. Future study could look into ensemble methods like stacking or hybrid models, which integrate multiple classifiers to increase prediction accuracy and stability. Finally, the study focuses on specific machine learning models. Future research can use other models like as AdaBoost and artificial neural networks to improve forecast accuracy and robustness.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

This work was supported by the Universiti Sains Malaysia, Incentive Graduate on Time (Grant Number: R502-KR-GOT001-0000001269-K134).

AUTHOR CONTRIBUTIONS

Zhi Zheng Kang: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation;
Sin Yin Teh: Project Administration, Supervision, Writing, Corrections;
Samuel Yong Guang Tan: Review & Editing;
Wei Chien Ng: English Proofreading.

CONFLICT OF INTERESTS

No conflicts of interest were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

REFERENCES

[1] "Lending global market report 2025," The Business Research Company. [Online]. Available: https://www.thebusinessresearchcompany.com/report/lending-global-market-report

[2] "Commercial lending market size, share, and growth analysis." SkyQuest Technology Group. [Online]. Available: https://www.skyquestt.com/report/commercial-lending-market

[3] S. A. Aziz, R. Jayanti, and A. Dinaseviani, "The role of bank and startup fintech P2P lending in supporting financial credit for Indonesian farmers," *Jurnal Perspektif Pembiayaan dan Pembangunan Daerah,* vol. 12, no. 1, pp. 47-66, 2024, doi: 10.22437/ppd.v12i1.23575.

[4] "The role of AI in credit risk management." JurisTech. [Online]. Available: https://juristech.net/juristech/the-role-of-ai-in-credit-risk-management/ (accessed.

[5] E. B. Ntiamoah, E. Oteng, B. Opoku, and A. Siaw, "Loan default rate and its impact on profitability in financial institutions," *Research Journal of Finance and accounting,* vol. 5, no. 14, pp. 67-72, 2014.

[6] V. Ivashina, D. Scharfstein, "Bank lending during the financial crisis of 2008," *Journal of Financial Economics,* vol. 97, no. 3, pp. 319-338, 2010, doi: 10.1016/j.jfineco.2009.12.001.

[7] D. S. Nkambule, B. Twala, and J. H. C. Pretorius, "Effective machine learning techniques for dealing with poor credit data," *Risks,* vol. 12, no. 11, 2024, doi: 10.3390/risks12110172.

[8] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decision Analytics Journal,* vol. 3, 2022, doi: 10.1016/j.dajour.2022.100071.

[9] A. K. C, H. S. Samreen, T. GP, P. S, and Y. B, "Forecasting loan suitability with machine learning," *Journal of Emerging Technologies and Innovative Research,* vol. 11, no. 4, 2024.

[10] V. Padimi, V. Sravan, and D. D. Ningombam, "Applying machine learning techniques to maximize the performance of loan default prediction," *Journal of Neutrosophic and Fuzzy Systems,* pp. 44-56, 2022, doi: 10.54216/jnfs.020204.

[11] L. Sathish kumar, V. Pandimurugan, D. Usha, M. Nageswara Guptha, and M. S. Hema, "Random forest tree classification algorithm for predicating loan," *Materials Today: Proceedings,* vol. 57, pp. 2216-2222, 2022, doi: 10.1016/j.matpr.2021.12.322.

[12] W.-W. Tay, S.-C. Chong, and L.-Y. Chong, "DDoS attack detection with machine learning," *Journal of Informatics and Web Engineering,* vol. 3, no. 3, pp. 190-207, 2024, doi: 10.33093/jiwe.2024.3.3.12.

[13] W. Wu, "Machine learning approaches to predict loan default," *Intelligent Information Management,* vol. 14, no. 05, pp. 157-164, 2022, doi: 10.4236/iim.2022.145011.

[14] J. Gao, W. Sun, X. Sui, and A. Farouk, "Research on default prediction for credit card users based on XGBoost-LSTM Model," *Discrete Dynamics in Nature and Society,* vol. 2021, pp. 1-13, 2021, doi: 10.1155/2021/5080472.

[15] Y. Cheng, "Research on credit strategy based on XGBoost Algorithm and optimization problem" *Journal of Physics: Conference Series,* 2021, doi: 10.1088/1742-6596/1865/4/042137.

[16] Y. Zhou, "Loan default prediction based on machine learning methods," in *Proceedings of the 3rd International Conference on Big Data Economy and Information Management, BDEIM*, pp. 2-3, 2022.

[17] J. Gu and J. Lin, "Research on loan default prediction based on logistic regression, RandomForest, XGBoost and AdaBoost," *SHS Web of Conferences,* vol. 181, 2024, doi: 10.1051/shsconf/202418102008.

[18] J. Liu, Y. Gao, and F. Hu, "A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM," *Computers & Security,* vol. 106, 2021, doi: 10.1016/j.cose.2021.102289.

[19] X. Hao, Z. Zhang, Q. Xu, G. Huang, and K. Wang, "Prediction of f-CaO content in cement clinker: A novel prediction method based on LightGBM and Bayesian optimization," *Chemometrics and Intelligent Laboratory Systems,* vol. 220, 2022, doi: 10.1016/j.chemolab.2021.104461.

[20] X. Zhu, Q. Chu, X. Song, P. Hu, and L. Peng, "Explainable prediction of loan default based on machine learning models," *Data Science and Management,* vol. 6, no. 3, pp. 123-133, 2023, doi: 10.1016/j.dsm.2023.04.003.

[21] S. Albahra *et al.*, "Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts," *Seminars in Diagnostic Pathology,* vol. 40, no. 2, pp. 71-87, Mar 2023, doi: 10.1053/j.semdp.2023.02.002.

[22] A. Juna *et al.*, "Water quality prediction using KNN imputer and multilayer perceptron," *Water,* vol. 14, no. 17, 2022, doi: 10.3390/w14172592.

[23] A. Altamimi *et al.*, "An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques," *BMC Medical Research Methodology,* vol. 24, no. 1, pp. 221, Sep 27 2024, doi: 10.1186/s12874-024-02324-0.

[24] Q. H. Nguyen *et al.*, "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil," *Mathematical Problems in Engineering,* vol. 2021, pp. 1-15, 2021, doi: 10.1155/2021/4832864.

[25] D. Ojo, M. Al-Mhiqani, H. Al-Aqrabi, and T. Al-Shehari, "Evaluation of machine learning algorithm and SMOTE for Insider threat detection," in *International Symposium on Intelligent Computing Systems*, Springer, pp. 303-318, 2024.

[26] N. L. S. S. Seemakurthi, "Parkinson's disease detection using existing machine learning algorithms," Bachelor of Science in Computer Science, Blekinge Institute of Technology, 2024.

[27] H. Chen, L. Yang, and Q. Wu, "Enhancing land cover mapping and monitoring: an interactive and explainable machine learning approach using Google Earth Engine," *Remote Sensing,* vol. 15, no. 18, 2023, doi: 10.3390/rs15184585.

[28] M. Peplinski, B. Dilkina, M. Chen, S. J. Silva, G. A. Ban-Weiss, and K. T. Sanders, "A machine learning framework to estimate residential electricity demand based on smart meter electricity, climate, building characteristics, and socioeconomic datasets," *Applied Energy,* vol. 357, 2024, doi: 10.1016/j.apenergy.2023.122413.

[29] S. A. Lashari, M. M. Khan, A. Khan, S. Salahuddin, and M. N. Ata, "Comparative evaluation of machine learning models for mobile phone price prediction: Assessing accuracy, robustness, and generalization performance," *Journal of Informatics and Web Engineering,* vol. 3, no. 3, pp. 147-163, 2024, doi: 10.33093/jiwe.2024.3.3.9.

BIOGRAPHIES OF AUTHORS

	**Zhi Zheng Kang** is a master student of Universiti Sains Malaysia (USM) with a degree in Business Analytics. He has developed expertise in data analysis and business intelligence, which he applies in his current role at MYwave Sdn. Bhd., a company specializing in human capital management solutions. Based in Alor Setar, Malaysia, Kang has built a growing professional network and actively engages in analytics-driven problem-solving. His LinkedIn profile reflects his career journey and industry connections. His research focuses on machines learning. He can be contacted at email: kzheng975@gmail.com.
	**Sin Yin Teh** is an Associate Professor of Operations and Business Analytics at the School of Management, Universiti Sains Malaysia, Malaysia. She specialises in Statistical Quality Control, Operations Management, and Business Analytics. Teh has published over 100 papers in international journals and conference proceedings, including ISI Q1-ranked journals. Her research covers areas such as Robust Statistics, Data Mining, and the Theory of Inventive Problem Solving (TRIZ). She has received several academic awards and actively collaborates with industry partners on analytics-driven projects. She can be contacted at email: tehsyin@usm.my.
	**Samuel Yong Guang Tan** is a lecturer in the Department of Accountancy and Business at Tunku Abdul Rahman University of Management and Technology (TARUMT), Penang Branch, Malaysia. He holds a Bachelor of Commerce from the University of Wollongong and a Master of Business Analytics from Universiti Sains Malaysia (USM). His academic expertise lies in business analytics and commerce-related fields. He is actively involved in teaching and research within these areas. He can be contacted at email: ygtan@tarc.edu.my.
	**Wei Chien Ng** is a lecturer in the Business Analytics department at Universiti Sains Malaysia (USM), Malaysia. He was an Associate Dean and Senior Lecturer at Tunku Abdul Rahman University of Management and Technology (TAR UMT). With expertise in finance, data analytics, and TRIZ methodology, he has conducted various seminars and workshops for companies and government agencies. His research collaborations include projects with Wawasan Open University, Sanmina-SCI Systems, and CREST. He has presented at international conferences and industry events. He can be contacted at email: ngweichien@usm.my.