Thailand Statistician
July 2025; 23(3): 710-731
http://statassoc.or.th
Contributed paper

Machine Learning Approaches for Credit Risk Assessment in SMEs

Piyada Wongwiwat [a], Apisak Chairojwattana [b], and Wikanda Phaphan*[c, d]

[a] Department of Computer Science and Data Innovation, Faculty of Science and Technology, Suansunandha Rajabhat University, Bangkok, Thailand.

[b] Department of Mathematics, Faculty of Science, Burapha University, Chon Buri, Thailand.

[c] Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

[d] Research Group in Statistical Learning and Inference, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

*Corresponding author; e-mail: wikanda.p@sci.kmutnb.ac.th

Received: 8 September 2024
Revised: 5 May 2025
Accepted: 7 May 2025

Abstract

Small and medium enterprises (SMEs) often face difficulties securing loans from financial institutions due to perceived high credit risk. Machine learning-based credit risk assessment models (CRA) offer a promising approach to assessing borrower creditworthiness more objectively. This study compared the performance of four credit assessment techniques: decision trees, support vector machines (SVMs), logistic regression, and adaptive boosting model. All models were evaluated using K-fold cross-validation implemented in Python. Decision trees and adaptive boosting demonstrated comparable recall scores, offering a trade-off between overall accuracy and FN rate. Due to inherent data imbalance with a higher prevalence of regular loans compared to non-performing loans (NPLs), the dataset was balanced to address potential biases. Following data balancing, adaptive boosting achieved the highest overall accuracy, making it the most suitable model for predicting SME loan defaults in this study.

Keywords: Credit risk models, decision tree, support vector machine, logistic regression, adaptive boosting model, AdaBoost, NPL, status of installment payments, creditworthiness.

1. Introduction

Small and medium-sized enterprises (SMEs) constitute a substantial portion of the global economy and workforce, and their financial liquidity is critical for survival (Abdulsaleh et al. 2013; Lu et al. 2022). While profitable, growing SMEs require increased investment, making them attractive lending targets for financial institutions. However, economic downturns, such as the COVID-19 pandemic, reduce consumer purchasing power and consequently, SME revenues. Since expenses like employment, rent, and interest remain relatively fixed, SMEs with limited capital face heightened bankruptcy risk. This increases SME demand for external financing at a time when

financial institutions tighten lending criteria to mitigate non-performing loan (NPL) risk (Kithinji et al. 2010; Akwaa-Selyi et al. 2015). Therefore, accurate assessment of SME debt repayment capacity is essential, enabling financial institutions to manage risk and facilitating funding access for creditworthy borrowers.

A key challenge in this assessment is the limited availability of comprehensive financial data from SMEs, which often lack detailed records of income, expenses, and input costs. This information gap compels financial institutions to rely on basic indicators such as firm demographics and debt history (Kithinji et al. 2010; Alfaro et al. 2012; McCann et al. 2012), potentially hindering effective credit risk evaluation. Consequently, robust credit risk assessment is crucial for financial institutions to inform lending decisions, minimize losses, and enhance profitability (Lu et al. 2022; Liu et al. 2010; Wang et al. 2011). Credit risk assessment involves evaluating various factors to determine the likelihood of a borrower defaulting on their debt. Credit scoring models are a key tool within this process, providing a numerical representation of a borrower's creditworthiness based on their characteristics.

The literature on SME credit risk assessment highlights several key themes. Traditional approaches have often relied on statistical models like logistic regression and discriminant analysis (Kithinji et al. 2010; Liu et al. 2010; Li et al. 2020; So et al. 2007). Logistic regression uses a sigmoid function to estimate the probability of default based on financial ratios and other relevant variables, while discriminant analysis classifies borrowers into different risk categories by identifying factors that distinguish between them by maximizing the separability among the predetermined groups. These models aim to quantify credit risk, often resulting in a credit score. However, they may struggle to capture the complexities inherent in SME finances, which often involve non-linear relationships and qualitative factors (So et al. 2007; Dahiya et al. 2015; Gunnarsson et al. 2021; Ptak-Chmielewska and Matuszyk 2020). Furthermore, the effectiveness of these models depends on careful feature selection and data preprocessing, as these steps are crucial for transforming raw data into a suitable format for analysis.

More recently, machine learning techniques have emerged as promising tools, capable of handling large datasets and incorporating non-financial information to improve predictive accuracy (Wang et al. 2011; Li et al. 2020; Dahiya et al. 2015). These techniques, including ensemble methods like random forests and gradient boosting, are designed to improve predictive performance by reducing overfitting and variance. A key consideration in applying these techniques is the selection of relevant features, as this can significantly impact model performance (Dahiya et al. 2015). The machine learning techniques used in this article include: decision trees, which offer interpretability by providing a flowchart-like structure for decision-making, where internal nodes represent features, branches represent decisions, and leaf nodes represent outcomes (Quinlan 1986); support vector machines (SVMs), which are employed for classification by identifying optimal boundaries (hyperplanes) between different risk categories, maximizing the margin between the classes (Cortes and Vapnik, 1995); logistic regression, which employs a sigmoid function to predict default probability based on financial ratios and other pertinent variables (Kithinji et al. 2010; Liu et al. 2010; Li et al. 2020; So et al. 2007); and adaptive boosting (AdaBoost), which combines multiple weak learners, such as decision stumps, into a strong ensemble, where AdaBoost assigns weights to each data point and iteratively trains classifiers, focusing on the misclassified instances from previous iterations (Freund and Schapire, 1997).

The objective of this article is to apply a credit analysis system employing four popular machine learning techniques, including decision tree techniques, support vector machines, logistic regression, and the Adaptive Boosting model, and to compare the efficiency of borrower classification using these techniques via the K-Fold cross-validation method, utilizing the Python programming language.

2. Materials and Methods

2.1. Data

2.1.1 Data acquisition

This study utilized data provided by the Fund for the Promotion of Family and Handicraft Industries in Thailand, Department of Industrial Promotion, Ministry of Industry. The dataset comprises borrower information from the revolving fund. The borrowers include both individual people and legal entities; however, the analysis in this study is restricted to individual borrowers due to their prevalence among the Fund's customer base and their higher proportion of non-performing loans (NPLs).

The selection of variables for analyzing the determinants of default and bad debt was guided by a review of existing literature and data completeness. To ensure the robustness of the analysis, only variables with complete information for a majority of borrowers were included. The consistency of borrower information was also examined to identify and address any abnormal values. Descriptive statistics for the variables used in this analysis are presented in Tables 1 and 2. The relationships between key features are explored through correlation analysis, and the results are visualized in Figure 1.

Table 1 Qualitative features and descriptions

Qualitative Feature Name	Description	Value
Age	Age (years) at the time of loan application.	Real Numbers
Member	Number of family members at the time of loan application.	Real Numbers
IN_debt	Total debt with all formal financial institutions that are not related to business at the time of loan application.	Real Numbers
OUT_debt	Total amount of informal debt that is not related to business at the time of loan application.	Real Numbers
Num_Emp	Number of employees.	Real Numbers
Exprenses_Y	Expenses per year (Baht).	Real Numbers
Ratio_CCus-TCus	Rate of change in total employee expenses during the 3 years before the time of loan application (%).	Real Numbers
Growth_rate_3Y	Average annual growth rate of income for the past 3 years at the time of loan application (% per year).	Real Numbers
Net-income_LY	Last year's net profit rate at the time of loan application (%).	Real Numbers
IN_debt_BUS	The total amount of debt borrowed from financial institutions in the system for use in business at the time of loan application.	Real Numbers
OUT_debt_BUS	Total debt of informal loans borrowed for use in business at the time of loan application.	Real Numbers

Table 1 (Continued)

Qualitative Feature Name	Description	Value
Loan_Amount	The loan amount received at the time of loan application.	Real Numbers
period	The period for repaying the loan at the time of loan application.	Real Numbers
Target (Dependent variable)	Current installment payment status.	0, 1 (1 is considered to be NPLs.)

A preliminary dataset comprising 1,014 observations was analyzed, revealing a classification of 272 entities as non-performing loans (NPLs) and 742 as non-NPLs. Descriptive statistics indicate a mean borrower age of 44.62 years. The mean number of employees per firm was 6.56, with a maximum of 201 employees observed in the largest entity. The average business operating expenditure was calculated at 708,148.10 Baht, while the mean 3-year operating growth rate was 21.38%. Furthermore, the average year-over-year net income growth rate was 23.51%. The financing structure of these businesses involved both formal (financial institutions) and informal lending sources for operational purposes. The mean formal loan amount was 253,024.65 Baht, with the average amount specifically allocated for business operations being 407,685.90 Baht. The maximum formal loan obtained reached 19,834,276.00 Thai Baht. Informal borrowing for business operations averaged 2,393.02 Baht, with a recorded maximum of 1,300,000 Baht. Loans not directly related to business operations exhibited mean values of 587,794.06 Thai Baht for formal sources and 16,617.84 Baht for informal sources, as detailed in Table 2.

Table 2 Descriptive statistics of quantitative features

Qualitative Feature Name	Count	Mean	SD	Min	Max
Age	1014	44.62	9.87	20.00	74.00
Member	1014	1.21	1.30	0.00	10.00
IN_debt	1014	587,794.03	1,280,888.00	0.00	19,834,276.00
OUT_debt	1014	16,617.84	257,638.80	0.00	7,214,728.00
Num_Emp	1014	6.56	9.31	0.00	201.00
Exprenses_Y	1014	708,148.10	904,281.70	0.00	18,540,000.00
Ratio_CCus-TCus	1014	52.36	23.39	0.00	225.00
Growth_rate_3Y	1014	21.38	18.41	-10.00	100.00
Net-income_LY	1014	23.51	17.15	-60.00	100.00
IN_debt_BUS	1014	407,685.90	1,033,702.00	0.00	19,834,276.00
OUT_debt_BUS	1014	2,393.02	43,475.00	0.00	1,300,000.00
Loan_Amount	1014	253,024.65	650,243.40	0.00	20,000,000.00
period	1014	56.70	16.50	0.00	116.00

2.1.2 Data exploration

Prior to further analysis, a data exploration process was conducted to understand the characteristics of the dataset. Figure 1 illustrates the Pearson correlation coefficients between each feature, with values ranging from -1 to 1. As shown in Figure 1, Age, Member, IN_debt, OUT_debt, Num_Emp, Expenses_Y, Net_income_LY, IN_debt_BUS, OUT_debt_BUS, and Loan_Amount exhibit minimal linear association with the target variable. Ratio_CCus_TCus, Growth_rate_3Y, and

period demonstrate a weak linear association with the target. While Num_Emp and Expenses_Y, IN_debt and IN_debt_BUS, and Net_income_LY and Growth_rate_3Y are correlated with each other.

The age distribution of loan applicants ranges from 20 to 75 years, with a concentration between 40 and 60 years. The distribution of loan applicant ages, along with a comparison between NPL and non-NPL groups, is visualized in Figure 2.

Analysis of business operations, illustrated in Figure 3, indicates that the majority of applicants are associated with small businesses, characterized by a workforce of fewer than 50 employees and operating expenses not exceeding 1 million baht. These businesses, however, tend to carry substantial non-business-related debt, with formal debt in financial institutions averaging approximately 1 million baht, and informal debt approximately 200,000 baht. Furthermore, it was observed that businesses with more than 50 employees tend to have higher expenses, consequently leading to a greater amount of borrowing from formal financial systems. This suggests that smaller organizations with lower operating expenses may be more susceptible to NPLs.

Figure 4 presents a visualization of borrower debt characteristics, specifically examining: total formal debt in financial institutions (excluding business-related debt), total informal debt (excluding business-related debt), formal business debt in financial institutions, and informal business debt. It was found that formal debt is correlated with formal debt unrelated to business operations. Conversely, it shows no correlation with informal debt. Furthermore, an increase in debt for business operations is associated with a tendency to borrow more from formal channels for other, non-business-related expenditures.

Figure 5 shows that the 3-year business growth rate for the NPLs group is lower, with a growth rate not exceeding 40%, and they exhibit negative annual net income, indicating business losses. Furthermore, Figure 5 illustrates a distinct pattern in debt utilization. Businesses with formal non-business debt also tend to utilize formal financing for their business operations. Conversely, businesses with informal non-business debt tend to rely on informal financing mechanisms for their business activities.

A comparative analysis of business financing sources indicates differences in growth and profitability. Businesses with formal financing demonstrate growth rates between 0% and 60% and show an increase in net income. In contrast, businesses with informal financing exhibit a more modest growth rate, averaging approximately 20%, and a lower increase in net income, around 25%. The distribution of net profit margins for both non-NPL and potential NPL businesses is shown in Figure 6, revealing contrasting patterns. Non-NPL businesses show a wider range of net profit margins (0% to 80%), while potential NPL businesses exhibit a narrower range (-15% to 40%).

2.1.3 Data preprocessing and feature selection

All 13 independent variables, including demographic (e.g., Age, Member), financial (e.g., IN_debt, OUT_debt, Net-income_LY), and business-related indicators (e.g., Num_Emp, Growth_rate_3Y, IN_debt_BUS), were retained for initial analysis, based on their relevance to SME creditworthiness as identified in the literature. Prior to model development, the dataset underwent a comprehensive preprocessing pipeline to address issues related to data quality, consistency, and format:

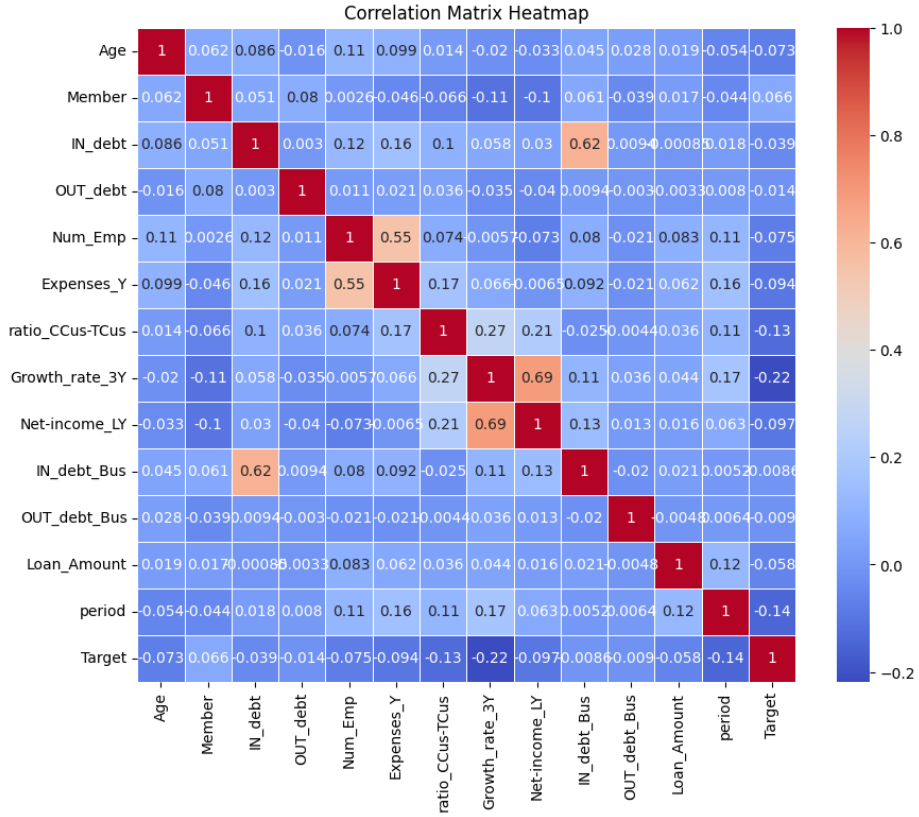


Figure 1 Correlation heat map of each features

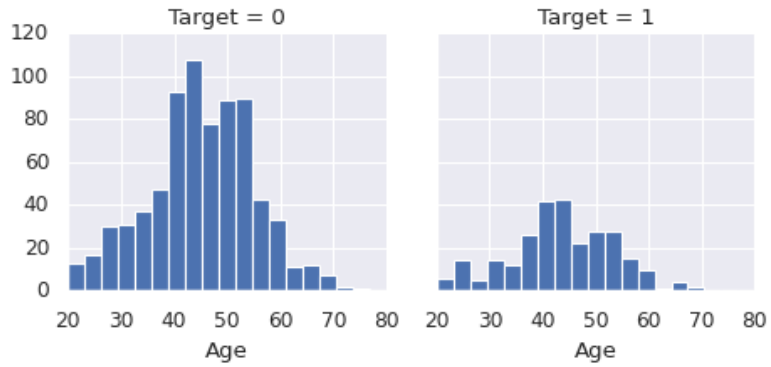


Figure 2 The data distribution for both NPL and non-NPL

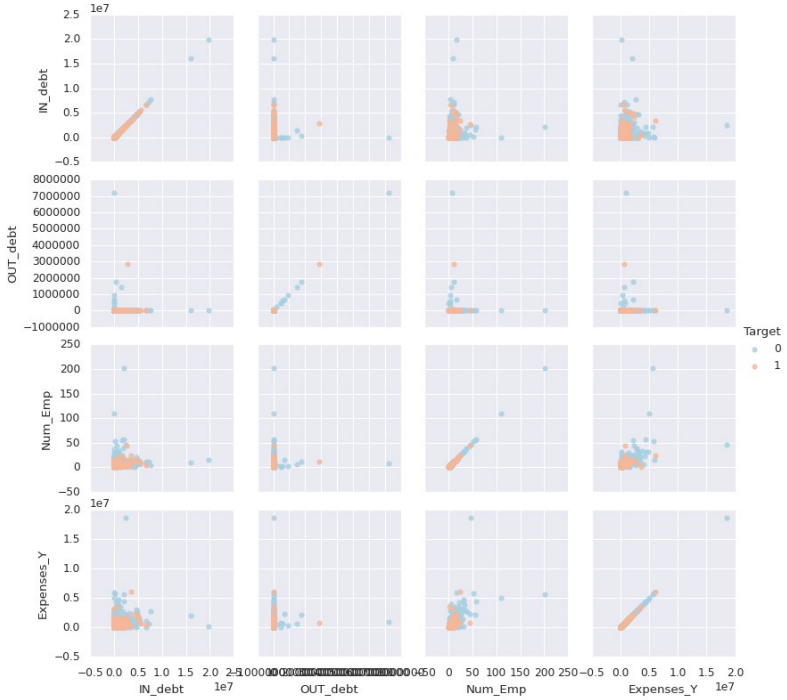


Figure 3 Scatter plot of the relationship of credit risk score and business operations

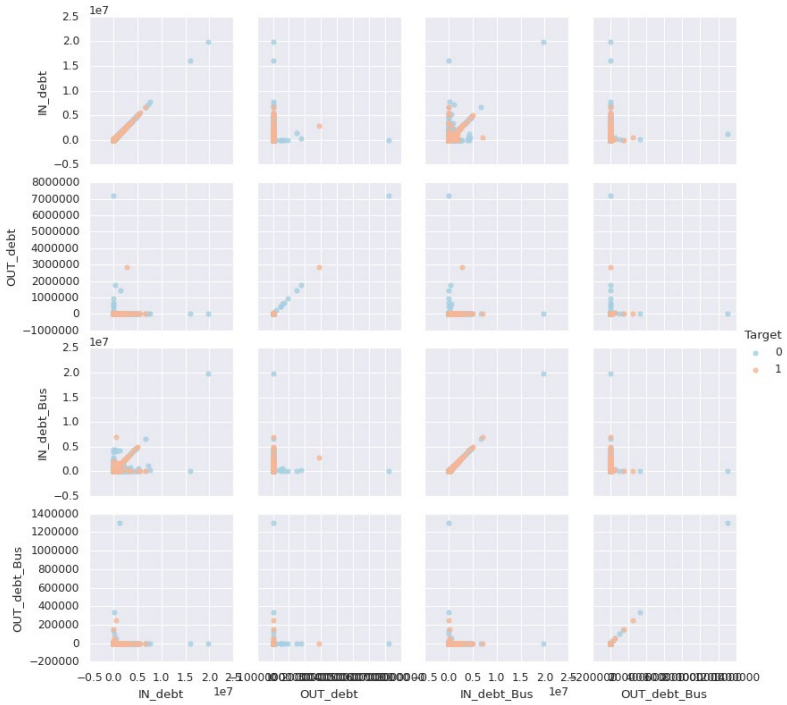


Figure 4 Scatter plot of the relationship of credit risk score and debt characteristics

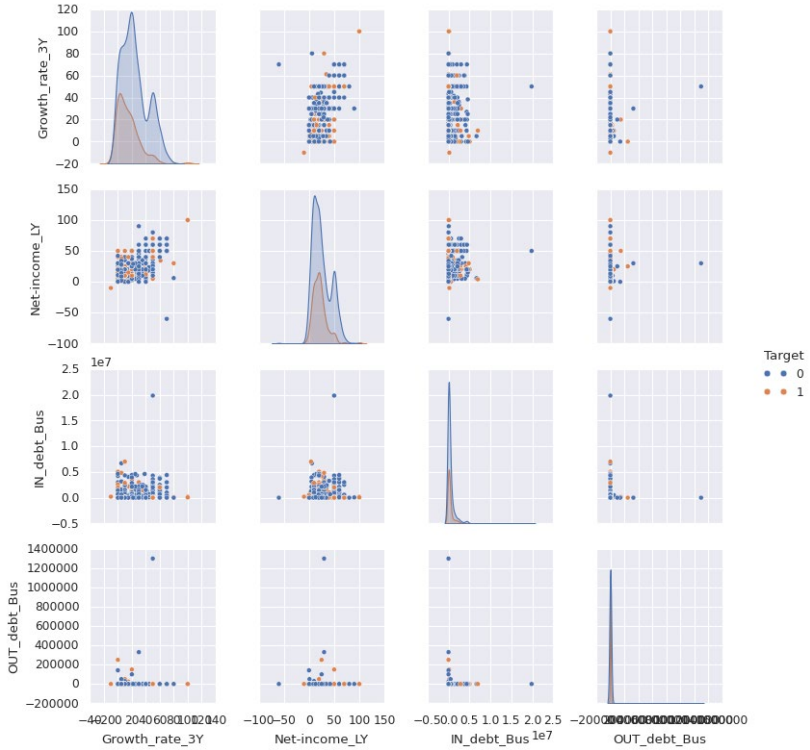


Figure 5 The formal debt in financial institutions

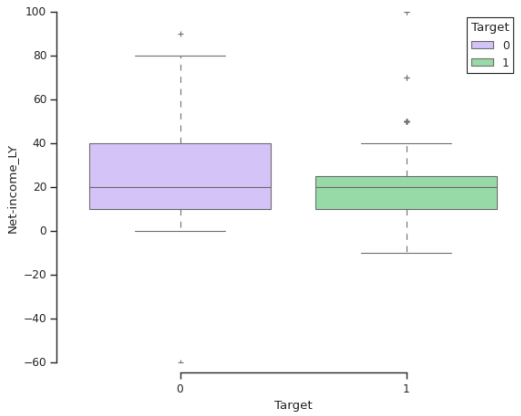


Figure 6 Box plots for relative risks for credit risk score

 Missing Values: The dataset was examined for missing values. Variables with missing rates below 5% were imputed using the median (for numerical variables) to preserve the underlying distribution, while records with excessive missing data were excluded to avoid introducing bias.

 Outlier Detection and Treatment: Outliers in continuous variables such as Expenses_Y, Loan_Amount, and Net_income_LY were detected using the IQR method and winsorized to reduce skewness without eliminating potentially important cases.

 Normalization and Scaling: To ensure comparability among features with different scales (e.g., Loan_Amount and Growth_rate_3Y), min-max normalization was applied to continuous variables prior to modeling, particularly for algorithms sensitive to feature scale, such as SVM and Logistic Regression.

 Categorical Encoding: The target variable, target (installment payment status), was binary encoded (0 for non-default, 1 for default). All input features were numerical; hence, no additional encoding was required.

2.2. Machine learning models and algorithms

This section outlines the machine learning models employed in this study for credit risk assessment. The selection of these models—decision trees, support vector machines (SVMs), logistic regression, and adaptive boosting (AdaBoost)—was predicated on their established efficacy in classification problems, their ability to handle the characteristics of credit risk data, and their widespread application in related financial modeling contexts. To ensure robust and reliable model evaluation, a rigorous 5-fold cross-validation procedure was implemented. The 5-fold cross-validation technique, detailed below, was consistently applied across all four machine learning models under investigation and across all experimental conditions, encompassing both the imbalanced and balanced datasets (achieved through undersampling and SMOTE oversampling). This uniform application of the cross-validation methodology is critical for enabling a fair comparison of model performance and minimizing potential bias.

A stratified 5-fold cross-validation protocol was adopted to preserve the original class distribution within each fold. This stratification is of particular importance in the context of imbalanced datasets, such as those encountered in credit risk assessment, as it guarantees that each fold accurately represents the proportion of both the majority (non-NPL) and minority (NPL) classes. The procedure was executed as follows:

1. Prior to fold creation, the complete dataset was randomly shuffled, employing a fixed random seed to ensure the reproducibility of the data partitioning.

2. The shuffled dataset was then partitioned into five mutually exclusive folds, with each fold containing approximately 20% of the data, while maintaining the overall class distribution.

3. The model training and evaluation process was iterated five times. In each iteration, a single fold was designated as the validation set, and the remaining four folds were used to train the model.

4. Following training, the model's predictive performance was assessed on the validation set, and the chosen evaluation metrics (accuracy, precision, recall, and F-measure) were computed for that specific fold.

5. The final performance metrics for each model were calculated by averaging the metrics obtained across the five validation folds. This averaging procedure provides a more stable and generalized estimate of the model's performance, reducing the impact of any single, potentially idiosyncratic, data split.

The specific hyperparameters for each model were tuned using grid search in conjunction with this 5-fold cross-validation process, as detailed below:

2.2.1 Decision trees

Decision trees are non-parametric supervised learning algorithms utilized for both classification and regression tasks. The model operates by partitioning the feature space into a set of decision rules. Specifically, the value of the response variable is determined through a series of pre-specified rules derived from the predictor variables. This process can be conceptualized as a piecewise constant

approximation (Bhowmik 2011; Srisuradetchai and Panichkitkosolkul 2022). Decision trees were chosen for their interpretability and ability to capture non-linear relationships between variables, which is important in understanding the complex factors that influence credit risk. For the decision tree model, key hyperparameters such as maximum depth and minimum samples split were optimized using grid search with 5-fold cross-validation to prevent overfitting and improve generalization performance.

2.2.2 Support vector machine (SVM)

Support vector machines (SVMs) are effective machine learning tools applicable to a variety of tasks. SVMs are particularly advantageous in scenarios involving high-dimensional data, demonstrating robustness to outliers, and offering flexibility in accommodating diverse data types through the utilization of kernel functions. Classification in SVM is based on the distance between a new data point and a separating hyperplane. Specifically, a new data point is assigned to the class corresponding to the nearest hyperplane (Vapnik 1998; Zhang 2001; Zhang et al. 2015; Zan et al. 2004). SVMs were selected for their capacity to handle non-linear relationships through kernel functions and their effectiveness in classification tasks with potentially high-dimensional financial data. The SVM model was tuned using grid search with 5-fold cross-validation to determine the optimal values for the regularization parameter (C) and the kernel coefficient (gamma) to maximize the margin and minimize classification errors.

2.2.3. Logistic regression

Logistic regression is a machine learning method employed for modeling qualitative response variables using either quantitative or qualitative independent variables. The methodology models the data by a sigmoid curve representing the probability of the event of interest. The dependent variable, denoted as Y, is defined as $Y = 1$, indicating the occurrence of the event of interest, and $Y = 0$, indicating otherwise (Harrell 2001; Simmachan et al. 2023). Despite its linearity, logistic regression serves as a robust baseline model and provides easily interpretable probabilities, making it valuable for understanding the likelihood of default. Its widespread use in credit scoring also facilitates comparison with existing studies. Hyperparameter tuning for logistic regression involved optimizing the regularization strength using 5-fold cross-validation to prevent overfitting and improve model stability.

2.2.4 Adaptive boosting model (AdaBoost)

Adaptive boosting (AdaBoost) constructs a predictive model based on an ensemble of multiple weak predictive models, typically decision tree models. AdaBoost iteratively adjusts the weights assigned to each model based on the number of incorrect predictions. This process continues until a predetermined level of accuracy is achieved (Simmachan et al., 2023). AdaBoost was chosen for its ability to improve the performance of weak classifiers and its effectiveness in handling complex classification problems. As an ensemble method, AdaBoost can reduce both bias and variance, leading to more accurate and robust credit risk assessments. For the AdaBoost model, the number of estimators and the learning rate were optimized through grid search with 5-fold cross-validation to balance model complexity and predictive accuracy.

2.3. Confusion matrix and evaluation metrics

The evaluation of machine learning models is crucial for assessing their effectiveness and reliability, particularly in classification tasks. This section details the confusion matrix and several key metrics used to quantify model performance, providing a foundation for understanding how well a model distinguishes between different classes

2.3.1 Confusion matrix

Following the application of a predictive model, the predicted outcomes are compared with the actual values. This comparison results in four possible categories, which are commonly represented in a confusion matrix, as detailed in Table 3 (Botchey et al. 2020). The confusion matrix provides a comprehensive view of a classification model's performance by illustrating the counts of correct and incorrect predictions for each class. These categories are defined as follows:

 True Positive (TP): A TP represents a case where the model correctly predicts the positive class. In the context of credit risk, for example, a TP would be a loan that was correctly predicted to default.

 True Negative (TN): A TN represents a case where the model correctly predicts the negative class. In credit risk, this would be a loan that was correctly predicted to be repaid.

 False Positive (FP): A FP, also known as a Type I error, represents a case where the model incorrectly predicts the positive class. In this scenario, the prediction is positive, but the actual class is negative. For instance, a loan predicted to default that was actually repaid.

 False Negative (FN): A FN, also known as a Type II error, represents a case where the model incorrectly predicts the negative class. In this scenario, the prediction is negative, but the actual class is positive. For instance, a loan predicted to be repaid that actually defaulted.

To provide a systematic evaluation of machine learning model performance, several established metrics are employed. This research utilizes accuracy, precision, recall, and F-measure for the evaluation of classification tasks (Vaklili et al. 2020). These metrics offer different perspectives on the model's ability to correctly classify instances, with a particular emphasis on its performance with respect to the positive class (e.g., loan default).

- True Positive (TP): an outcome where the model correctly predicts the positive class.
- True Negative (TN): an outcome where the model correctly predicts the negative class.
- False Positive (FP): an outcome where the model falsely predicts the positive class. In other words, the prediction provides true while the actual is not true.
- False Negative (FN): an outcome where the model falsely predicts the negative class. In other words, the prediction is false while the actual is true.

To systematically evaluate our machine learning models, there are ubiquitous tools which numerically evaluate the performance of the model. The evaluation metrics for classification tasks we apply in this research are accuracy, precision, recall, and F-measure (Vaklili et al. 2020).

2.3.2 Accuracy

Accuracy is a fundamental evaluation metric, representing the proportion of correctly classified observations out of the total number of observations. While intuitive, accuracy can be misleading in cases where the classes are imbalanced, meaning one class has significantly more instances than the other:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{\text{number of correctly classified data}}{\text{total number of observations}}. \quad (1)$$

Table 3 Confusion matrix

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

2.3.3 Precision

Precision quantifies the proportion of correctly predicted positive cases among all instances predicted as positive. It reflects the reliability of the positive predictions. In the context of credit risk, precision indicates how often a loan predicted to default actually defaulted. A high precision score suggests that when the model predicts a loan will default, it is likely to be correct:

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{number of correctly classified data as positives}}{\text{total number of positive classified observations}}. \quad (2)$$

2.3.4 Recall

Recall measures the proportion of actual positive cases that are correctly predicted as positive. It indicates the model's ability to identify all relevant positive instances. In credit risk, recall indicates how many of the loans that actually defaulted were correctly identified by the model. A high recall score suggests that the model is effective at finding most of the defaulting loans:

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{number of correctly classified data}}{\text{total number of actual positives}}. \quad (3)$$

2.3.5 F-measure

The F-measure, also known as the F1-score, represents the harmonic mean of precision and recall. It provides a balanced measure of a model's accuracy, particularly useful when the class distribution is imbalanced. The F-measure gives more weight to lower values, thus penalizing models that have significant differences between precision and recall. A high F-measure indicates that the model has both high precision and high recall:

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}. \quad (4)$$

2.4. Experiments

This section details the experimental procedure employed to evaluate the effectiveness of the proposed machine learning models in a credit risk assessment context. The experiments were conducted under three scenarios: an imbalanced scenario reflecting the actual distribution of the target variable, a balanced scenario achieved through undersampling, and a balanced scenario achieved through oversampling with SMOTE. These techniques were selected for their widespread use in addressing class imbalance and their complementary approaches to the problem.

Undersampling reduces the dominance of the majority class, while SMOTE actively generates new instances of the minority class, preserving more information than simple duplication.

2.4.1 Imbalanced scenario

The "Target" response variable exhibits an imbalanced distribution, with 272 observations in the non-performing loans (NPLs) class (Y = 1) and 742 observations in the non-NPLs class (Y = 0). This class imbalance presents a challenge for classification models, as they may exhibit bias towards the majority class. The experimental procedure for this scenario comprises two primary phases: data preprocessing and modeling, followed by model evaluation and analysis.

2.4.1.1 Data preprocessing and modeling

The preprocessed data, as described in the previous section, was used to develop and evaluate the machine learning models. A 5-fold cross-validation approach was employed. The procedure is as follows:

1.The dataset was partitioned into five equal-sized folds.

2.The model was trained using four of the five folds, and the remaining fold was used for testing. This process was iterated five times, with each fold serving as the test set in turn. This resulted in five distinct data partitions: {1, 2, 3, 4, 5}.

3. The accuracy, precision, recall, and F-measure were calculated for each of the five iterations.

2.4.1.2 Model evaluation and conclusion

Confusion matrices were generated, and the average values of accuracy, precision, recall, and F-measure were calculated for each model: decision tree, support vector machine, logistic regression, and adaptive boosting. The performance of these models was then compared, and conclusions were drawn based on the comparative analysis.

2.4.2 Balanced scenario: Undersampling

To mitigate the class imbalance, the number of non-NPL observations was reduced to 272 through simple random sampling. This resulted in a balanced dataset with 544 observations, evenly split between the NPL (Y=1) and non-NPL (Y=0) classes. Undersampling was chosen for its simplicity and computational efficiency. It is particularly suitable when the majority class has a large number of redundant or less informative samples.

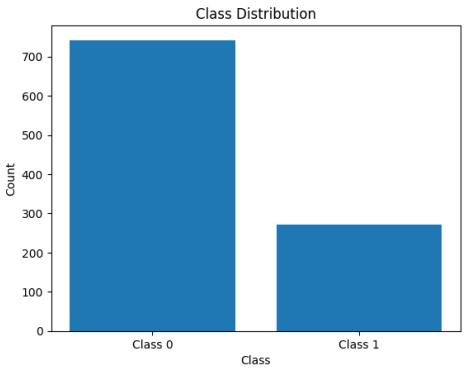


Figure 7 Class distribution

2.4.3 Balanced scenario: Oversampling with SMOTE

In a separate approach to address class imbalance, the synthetic minority over-sampling technique (SMOTE) was employed to increase the number of NPL observations. The SMOTE algorithm generates synthetic samples of the minority class (NPL) to balance the dataset. Unlike simple oversampling by duplication, SMOTE creates new instances by interpolating between existing minority class samples, thus reducing the risk of overfitting. This makes it a preferred oversampling method, especially when the minority class is relatively small. In this scenario, the number of NPL observations was increased to 742, resulting in a balanced dataset with a total of 1484 observations (742 in each class). The experimental procedure for this balanced scenario also follows the same two phases: data preprocessing and modeling and model evaluation and conclusion.

3. Results

In developing the credit risk assessment model, variables that are important for credit consideration are considered. In this research, a total of 13 independent variables were used, and the model development method will be learned through the method of a decision tree, support vector machines, logistic regression, and adaptive boosting models. Also, the model efficiency was tested using the K-fold cross-validation method by setting the value $k = 5$, so that in each round the data will be divided into 2 sets, namely the learning set (Training set), accounting for 80 percent of the total data and testing sets account for 20 percent of the total data. The performance metrics for the models, both for imbalanced and balanced datasets, are presented in Table 4.

Table 4 Evaluation metrics for all experiments

Models	Imbalanced data				Undersampling				SMOTE			
	Acc.	Pre.	Rec.	F-measure	Acc.	Pre.	Rec.	F-measure	Acc.	Pre.	Rec.	F-measure
Decision Tree	0.6980	0.3600	**0.1667**	**0.2278**	0.5556	0.5600	0.5185	0.5385	0.5845	0.5490	**0.9460**	0.6948
SVMs	0.7093	0.3571	0.0909	0.1449	0.5872	0.5918	0.5370	0.5631	0.4680	0.4672	0.4295	0.4476
Logistic Regression	**0.7340**	**1.0000**	0.0181	0.0357	0.5041	0.500	0.0556	0.1000	0.4916	0.4000	0.0268	0.0503
Adaptive Boosting	0.7143	0.4118	0.1273	0.1944	**0.7523**	**0.800**	**0.6667**	**0.7273**	**0.7071**	**0.6582**	0.8658	**0.7478**

Remark: Acc. - Accuracy, Pre. – Precision, and Rec. - Recall

3.1. Imbalanced data

The evaluation metrics of imbalanced data in Table 4 and Figure 8 reveal varied performance across four models: decision tree, SVMs, logistic regression, and adaptive boosting. Logistic regression achieves the highest accuracy (0.7340) and perfect precision (1.0), but suffers from extremely low recall (0.0181), indicating it rarely identifies positive cases. The decision tree model shows a more balanced performance, with the highest recall (0.1667) and F-measure (0.2278) among all models, suggesting it is better at identifying positive cases without sacrificing too much precision. SVMs and adaptive boosting fall between these extremes, with adaptive boosting showing slightly better overall performance. These results highlight a trade-off between precision and recall across the models, with no single model excelling in all metrics. The choice of model would depend on the specific requirements of the classification task, such as whether false positives or false negatives are more costly in the given context. In the non-performing loans (NPLs), it is recommended to prioritize the analysis of false negatives. Consequently, for imbalanced data scenarios, we recommend utilizing the decision tree model.

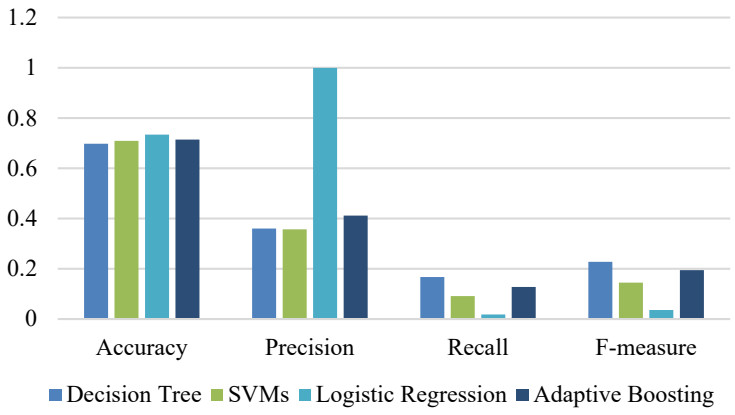


Figure 8 Model performance among imbalanced data scenarios

In creating the decision tree model, it was found that the variables used to classify the current installment payment status that is likely to be an NPL borrowers include age (years) at the time of loan application (Age), number of family members under responsibility, expenses at the time of loan application (Member), number of employees (Num_Emp), expenses per year (Expenses_Y), Average annual growth rate of Income in the past 3 years at the time of loan application (Growth_rate_3Y), Net profit rate of the past year at the time of loan application (Net-income_LY), total debt borrowed from financial institutions in the system for use in business at the time of loan application (IN_debt_Bus), total loan amount received at the time of loan application (Loan_Amount), and period for repaying the loan at the time of loan application (period). Utilizing the decision tree model, the resulting decision tree diagram is depicted in Figure 9. The confusion matrix for the decision tree model is presented in Table 5.

Table 5 The confusion matrix of the decision tree model

Prediction Value	Actual Value		Total
	NPL Borrower	Normal Borrower	
NPL Borrower	9	16	25
Normal Borrower	45	123	177
Total	54	148	202

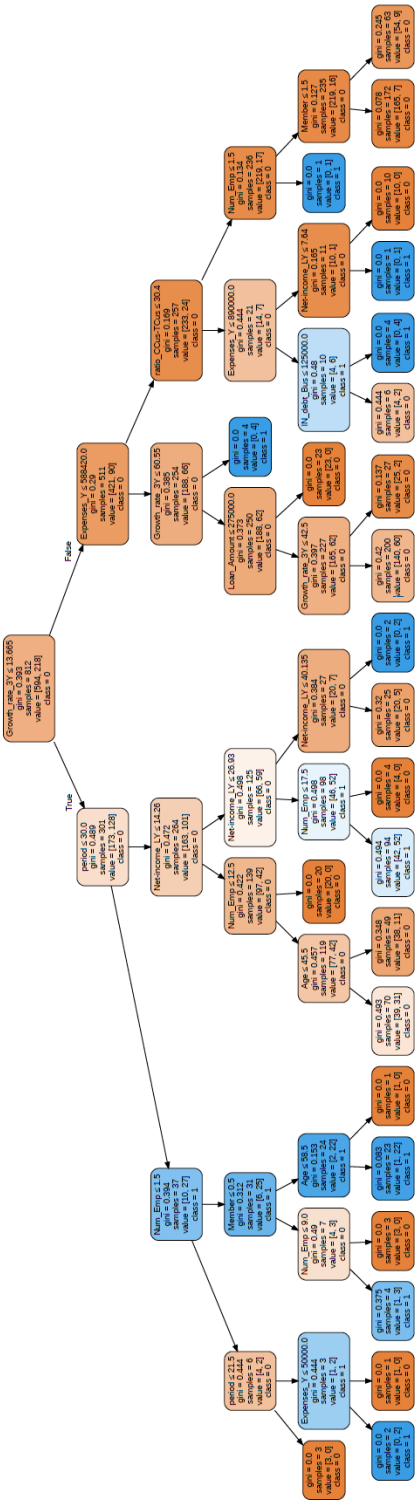


Figure 9 Decision tree diagram

3.2. Observations Reduction

After the observation reduction process, the evaluation metrics of all models are shown in Table 4 and Figure 10. Among all criteria, the adaptive boosting model demonstrates the highest performance, achieving an accuracy of 75.23%, a precision of 80.00%, a recall of 66.67%, and an F-measure of 72.73%. In contrast, the Logistic Regression model exhibits the lowest performance, with an accuracy of 50.41%, a precision of 50.00%, a recall of 5.56%, and an F-measure of 10.00%. Therefore, we recommend utilizing the adaptive boosting model to predict the installment status of new borrowers, thereby aiding financial institutions in making informed lending decisions before they approve any new credit. The confusion matrix for the adaptive boosting model is presented in Table 6.

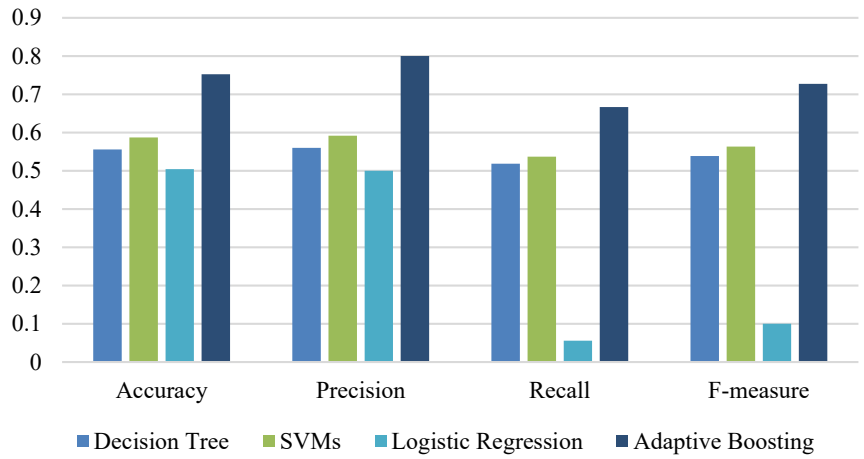


Figure 10 Model performance among observation reduction scenarios

Table 6 The confusion matrix of the adaptive boosting model of observation reduction scenarios

Prediction Value	Actual Value		Total
	NPL Borrower	Normal Borrower	
NPL Borrower	36	9	45
Normal Borrower	18	47	65
Total	54	56	110

3.3. Observations increasing using SMOTE

In terms of accuracy, the evaluation metrics of all models are shown in Table 4 and Figure 11, precision, and F-measure the adaptive boosting model performs the best by achieving 70.71% accuracy, 65.82% precision, and 74.78% F-measure while the SVMs model achieves the lowest accuracy at 46.80%, the logistic regression model achieves the lowest precision, and F-measure at 40.00%, and 5.03%, respectively.

In terms of recall, the decision tree model performs the best by achieving 94.60% recall, while the logistic regression performs surprisingly poorly by achieving only 2.68% recall.

Therefore, the results of balancing data by increasing data using SMOTE suggest that for SME loan default prediction, using a balanced dataset with an adaptive boosting model could provide the most reliable and accurate assessments, effectively balancing the identification of both NPL and non-NPL cases. The confusion matrix for the adaptive boosting model is presented in Table 7.

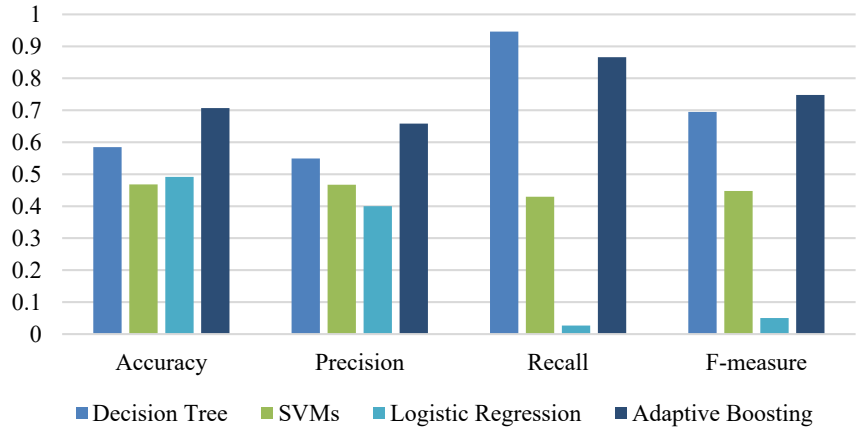


Figure 11 Model performance among increased data scenarios

Table 7 The confusion matrix of the adaptive boosting model of increasing data

Prediction Value	Actual Value		Total
	NPL Borrower	Normal Borrower	
NPL Borrower	129	67	196
Normal Borrower	20	81	101
Total	149	148	297

Based on the data analysis results in Table 4, considering the accuracy metric, it is observed that balancing the data does not increase the Accuracy, with the exception of the adaptive boosting analysis. The highest accuracy is achieved with the undersampling method. Regarding precision, data balancing does not lead to an increase in Precision either. However, for the adaptive boosting method, Precision can be improved when the data is balanced. For the recall metric, it is found that balancing the data increases the recall value. The adaptive boosting method shows a particularly significant increase. Similarly, for the F-measure, data balancing leads to an increase, especially with the adaptive boosting method. Therefore, balancing the data can improve accuracy, precision, recall, and F-measure using the adaptive boosting technique. Consequently, adaptive boosting analysis should be employed for predicting the status of loan applicants. For imbalanced data, decision tree analysis can be used, as the decision tree model shows a more balanced performance.

4.4. Discussion

Credit scoring models have become invaluable tools for financial institutions to evaluate the creditworthiness of potential borrowers and make informed lending decisions. In this study, four classification techniques—decision tree, support vector machines (SVM), logistic regression, and adaptive boosting—were developed to classify installment payment status. The study considered two main scenarios: imbalanced data (reflecting the original distribution) and balanced data (through undersampling and SMOTE).

Initially, the original dataset was highly imbalanced, with a larger proportion of non-NPL (non-performing loan) debtors. Under this scenario, the decision tree method slightly outperformed the

other models in terms of F-measure (0.2278), followed by adaptive boosting (0.1944). Despite moderate accuracy levels (0.6980 and 0.7143, respectively), all models exhibited low F-measure scores, indicating difficulty in identifying NPL cases due to class imbalance.

To address this issue, undersampling was applied to reduce the number of non-NPL instances, thereby balancing the dataset. The analysis revealed that adaptive boosting achieved the best performance on the balanced data, particularly under the undersampling scenario, with an F-measure of 0.7273—substantially higher than under the imbalanced case.

Subsequently, SMOTE (synthetic minority oversampling technique) was employed to synthetically balance the dataset. The results further validated that Adaptive Boosting remained the most effective model, reaching an F-measure of 0.7478, the highest among all experiments. This indicates its robustness and suitability in handling class imbalance effectively.

To statistically confirm these findings, the Friedman non-parametric test was conducted to compare the models across different scenarios. Table 4 reports performance metrics for all models, while Table 8 summarizes their rankings based on F-measure. The Friedman test showed a statistically significant difference in model performance (p-value = 0.0719 < 0.1), with Adaptive Boosting achieving the best mean rank (1.33), followed by decision tree, SVM, and logistic regression. These results confirm that adaptive boosting consistently delivers superior performance across all conditions.

Furthermore, among the three data-handling scenarios—imbalanced, undersampling, and SMOTE—SMOTE yielded the highest F-measure values across most models, especially when combined with adaptive boosting and decision tree techniques. This suggests that SMOTE is the most effective technique for addressing class imbalance in this context.

In conclusion, the adaptive boosting model, particularly when applied to data preprocessed using SMOTE, is recommended for predicting installment status. Its ability to balance precision and recall ensures more accurate identification of both NPL and non-NPL cases. This makes it a valuable tool for financial institutions aiming to enhance their credit risk assessment and lending decisions.

Table 8 Friedman test rankings of classification models based on F-measure

Model	Mean Rank	Rank
Adaptive Boosting	1.33	1
Decision Tree	2	2
SVM	2.67	3
Logistic Regression	4	4

5. Conclusions

This study compared the performance of various machine learning techniques for predicting SME loan defaults. While logistic regression achieved high accuracy in identifying creditworthy borrowers, its tendency to misclassify defaults, and high false negative rate, limited its effectiveness. Conversely, decision trees and adaptive boosting offered a balance between overall accuracy and correctly identifying defaulting loans, a high recall and F-measures score. Considering the inherent data imbalance and the critical nature of accurately predicting defaults. The results suggest that for SME loan default prediction, using a balanced dataset with an adaptive boosting model could provide the most reliable and accurate assessments, effectively balancing the identification of both NPL and non-NPL cases. These findings suggest that machine learning holds promise for enhancing financial institutions' ability to assess SME loan applications more objectively and mitigate credit risk.

Future research could explore the generalizability of these findings across diverse SME loan portfolios and investigate the potential for incorporating additional data sources to further improve model performance.

Acknowledgments

The authors would like to express my sincere thanks to the Department of Industrial Promotion, Ministry of Industry for support data and this research was funded by Suansunandha Rajabhat University.

References

Abdulsaleh AM, Worthington AC. Small and medium-sized enterprises financing: A review of literature. Int J Bus Manag. 2013; 8(14), 36; http://www.doi.org/10.5539/ijbm.v8n14p36

Akwaa-Sekyi EK, Bosompra P. Determinants of business loan default in Ghana. J Sci Res. 2015; 1(1): 10-26.

Alfaro R, Gallardo N. The determinants of household debt default. Rev Anal Econom. 2012; 27(1): 55-70.

Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.

Bhowmik T. Data mining. New Delhi: Dorling Kindersley India; 2011.

Bhowmik R. Detecting auto insurance fraud by data mining techniques. CIS J. 2011; 2(4): 156-162.

Botchey FE, Qin Z, Hughes-Lartey K. Mobile money fraud prediction—a cross-case analysis on the efficiency of support vector machines, gradient boosted decision trees, and naïve Bayes algorithms. Inf. 2020; 11(8): 383, https://doi.org/10.3390/info11080383

Breiman L. Random forests. Mach Learn. 2001; 45(1): 5-32.

Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and regression trees. New York: Chapman and Hall/CRC; 1984.

Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995; 20(3): 273-297.

Dahiya S, Handa SS, Singh Netra PS. Credit scoring using ensemble of various classifiers on reduced feature set. Industrija. 2015; 43(4): 163-172.

Drucker H, Burges CJC, Kaufman L, Smola AJ, Vapnik VN. Support vector regression. Adv Neural Inf Process Syst. 1996; 9: 155-161.

Fantazzini D, Figini S. Random survival forests models for SME credit risk measurement. Methodol Comput Appl Prob. 2009; 1: 29-45.

Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001; 29(5): 1189-1232.

Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997; 55(1): 119-139.

Goh RY, Lee LS. Credit scoring: A review on support vector machines and metaheuristic approaches. Adv Oper Res. 2019; 2019: 1-30.

Gunnarsson BR, Broucke SV, Baesens B, Óskarsdóttir M, Lemahieu W. Deep learning for credit scoring: Do or don't?. Eur J Oper Res. 2021; 295(1): 292-305.

Harrell Jr FE. Regression modeling strategies. New York: Springer; 2001.

Harrell FE. Binary logistic regression. In: Harrell FE, editor. Regression modeling strategies. New York: Springer Series in Statistics; 2001. p. 215-267.

Li XL, Zhong Y. An overview of personal credit scoring: techniques and future work. Int J Intell Sci. 2012; 2: 181-189.

Kithinji AM. Credit risk management and profitability of commercial banks in Kenya. Nairobi: School of Business, University of Nairobi; 2010.

Li Y, Chen W. A comparative performance assessment of ensemble learning for credit scoring. Math. 2020; 8(10): 1756; https://doi.org/10.3390/math8101756

Liu X, Fu H, Lin W. A modified support vector machine model for credit scoring. Int J Comput Intell Syst. 2010; 3(6): 797-804.

Loh WY. Classification and regression trees. Wiley Interdiscip Rev Data Mining Knowl Discov. 2011; 1(1): 14-23.

Lu Y, Yang L, Shi B, Li J, Abedin MZ. A novel framework of credit risk feature selection for SMEs during industry 4.0. Ann Oper Res. 2022: 1-28.

McCann F, McIndoe-Calder T. Determinants of SME loan default: The importance of borrower-level heterogeneity. Research Technical Papers from Central Bank and Financial Services Authority of Ireland. 2012; 06/RT/12.

Noriega JP, Rivera LA, Herrera JA. Machine learning for credit risk prediction. A systematic literature review. Data. 2023; 8: 169.

Ptak-Chmielewska A, Matuszyk A. Application of the random survival forests method in bankruptcy prediction for small and medium enterprises. Argum Oeconom. 2020; 1(44): 127-142.

Quinlan JR. Induction of decision trees. Mach Learn. 1986; 1(1): 81-106.

Si Z, Niu H, Wang W. Credit risk assessment by a comparison application of two boosting algorithms. In: Tallón-Ballesteros AJ, editor. Fuzzy Systems and Data Mining VIII. IOS Press; 2022. p. 34-40.

Simmachan T, Manopa W, Neamhom P, Poothong A, Phaphan W. Detecting fraudulent claims in automobile insurance policies by data mining techniques. Thail Stat. 2023; 21(3): 552-568.

Srisuradetchai P, Panichkitkosolkul W. Using ensemble machine learning methods to forecast particulate matter (PM) in Bangkok, Thailand. In: International Conference on Multi-disciplinary Trends in Artificial Intelligence; 2022. p. 204-215.

So YS, Hong SK. Random effects logistic regression model for default prediction of technology credit guarantee fund. Eur J Oper Res. 2007; 183(1): 472-478.

Tiwari KK, Somani R, Mohammad I. Determinants of loan delinquency in personal loan. Int J Manag. 2020; 11(11): 2566-2575.

Vakili M, Ghamsari M, Rezaei M. Performance analysis and comparison of machine and deep learning algorithms for IoT data classification. ArXiv. 2020; 2001.09636.

Vapnik VN. Statistical learning theory. New York: John Wiley & Sons; 1998.

Vapnik VN. The nature of statistical learning theory. New York: Springer-Verlag; 1998.

Wang G, Hao J, Ma J, Jiang H. A comparative assessment of ensemble learning for credit scoring. Expert Syst Appl. 2011; 38(1): 223-230.

Walusala WS, Rimiru DR, Otieno DC. A hybrid machine learning approach for credit scoring using PCA and logistic regression. Int J Comput. 2017; 27(1): 84-102.

Xie R, Liu R, Liu XB, Zhu JM. Evaluation of SMEs' credit decision based on support vector machine-logistics regression. J Math. 2021:1-10.

Zan H, Hsinchun C, Chia-Jung H, Wun-Hwa C, Soushan W. Credit rating analysis with support vector machines and neural networks: a market comparative study. Decis Support Syst. 2004; 37(4): 543-558.

Zan HW, Wang P, Zhang YQ. Bankruptcy prediction using support vector machine and boosting in an unbalanced dataset. In: Int Conf Mach Learn Cybernetics. IEEE; 2004. p. 3209-3213.

Zhang GP. Support vector machines in theory and financial time series forecasting. Eur J Oper Res. 2001; 129(1): 49-68.

Zhang L, Hu H, Zhang D. A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance. Financ Innov. 2015;1(1):1-21.

Zhang T. An introduction to support vector machines and other kernel-based learning methods. AI Mag. 2001; 22(2): 103-103.

Zhang Y, Zhou W, Luo Y. A novel hybrid credit risk evaluation model based on SVM and assembly strategy. J Comput Theor Nanosci. 2015; 12(8): 1835-1841.

Zhang Z, Niu K, Liu Y. A deep learning based online credit scoring model for P2P lending. IEEE Access. 2020; 8: 177307-177317.

Zhu Y, Li Z, Xie C, Wang GJ, Nguyen TV. Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach. Int J Prod Econ. 2019; 211: 22-33.