Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 9, No. 4, August, 2025

Gradient Descent Decision Tree Algorithm and Nonlinear Programming for Credit Risk Assessment and Credit Strategy

Guoqing Chen [1, 2], Nipaporn Chutiman [1], Sujitta Suraphee [1], Andrei Volodin [3], Piyapatr Busababodhin [1*]

[1] Department of Mathematics, Faculty of Science, Mahasarakham University, Maha Sarakham, 44150, Thailand.

[2] Mathematical Modeling Research Center, Chengdu Jincheng College, Chengdu, 611731, China.

[3] Department of Mathematics and Statistics, University of Regina, Regina, SK S4S 0A2, Canada.

Abstract

This research aimed to develop a scientific and accurate credit risk assessment model for small and medium-sized enterprises (SMEs) to support banks in credit decision-making. An improved decision tree model is proposed by integrating regularization to control complexity and employing an ensemble learning approach to enhance prediction accuracy. Multiple weak classifiers are iteratively refined using gradient descent optimization to form a robust, strong classifier. The model is trained through supervised learning, with the default probability of SMEs as the objective function, enabling a quantitative assessment of credit risk. The findings show that the proposed gradient descent decision tree algorithm achieves an AUC of 0.99 under 70% and 80% training set ratios, outperforming Adaptive Boosting (AUC = 0.97), Random Forest (AUC = 0.91), and traditional decision trees (AUC = 0.82). To further optimize bank loan strategies, this paper constructs a nonlinear multi-objective programming model that maximizes expected loan returns while considering risk constraints. The proposed approach not only improves credit risk prediction but also assists banks in formulating optimal lending strategies. This study advances credit risk modelling by integrating regularization and ensemble learning, offering a novel and practical solution for SME credit assessment.

Keywords:

Credit Loan Default;

Optimal Credit Strategy;

Gradient Descent Decision Tree;

Algorithm;

Nonlinear Programming.

Article History:

Received:	12	January	2025
Revised:	09	June	2025
Accepted:	17	June	2025
Published:	01	August	2025

1- Introduction

Effective assessment and measurement of credit loan default risks is a core task for commercial banks to improve their operational and management levels. It is also the key content required by the Basel II Capital Accord for commercial banks that adopt the internal ratings-based approach. The Basel II Capital Accord, a set of international banking regulations, mandates that banks using the internal ratings-based approach must have robust credit risk measurement systems in place. Credit risk measurement was born with the creation of credit transactions and developed with the development of financial theory, financial markets, statistical theory, and computer technology. Credit risk measurement methods have successively experienced subjective evaluation methods, economic measurement methods, analytical nonparametric models, modern quantitative models of risk measurement, and hybrid models. Since the 21st century, as the scale of credit transactions has continued to expand, the amount of data has also increased. In order to further improve the model's default identification rate and reduce default loss, hybrid models based on a single model have gradually

[*] **CONTACT**: piyapatr.b@msu.ac.th

DOI: http://dx.doi.org/10.28991/ESJ-2025-09-04-05

attracted the attention of scholars. However, for small and micro enterprises, due to their small size, low management efficiency, weak internal control, large operational risks, and other problems, banks have made small, medium, and micro enterprises obtain far less financial financing than large enterprises for credit risk control. The urgency of determining the optimal credit strategy for banks based on the credit risk of enterprises cannot be overstated. It is a key issue that requires immediate attention.

With the development of big data, there are more and more data sources, and multiple data sets can be collected for the same or similar tasks (or research). Although these data from different sources are collected for the same or similar tasks, due to differences in factors such as measurement environments, measurement standards, or statistical calibers, multiple data sets cannot be simply combined and analyzed. How to effectively integrate and analyze such complex data sets from diverse sources is one of the new development directions of statistics and machine learning research. Based on the perspective of the main models for measuring corporate credit risk, some scholars focus on Support Vector Machines (SVMs), and their improved models have been receiving a lot of attention. An innovative multi-criteria optimization classifier, KFP-MCOC, which combines kernel, defuzzification, and penalty factors, was shown to outperform both traditional SVMs and fuzzy SVMs in terms of separateness, efficiency, and generalization through empirical experiments [1-4]. The potential of the KFP-MCOC model to revolutionize credit risk assessment is a topic of intrigue in the field. A novel hybrid integration approach, namely RSB-SVM, based on bagging and stochastic subspaces, employing SVMs as the underlying learning machine [5-8]. A weighted least squares support vector machine (LSSVM) classifier with experimental design, which showed promising classification results in credit risk assessment [9-10].

At the same time, some researchers improved the classical slime mold algorithm and combined it with the underlying SVM model for parameter optimization and constructed the RF-LSMA-SVM model to enhance the classification ability of credit risk rating [11–14]. The RAROC (risk-adjusted return) model, a significant tool in credit rating, calculates the default probability of enterprises and formulates credit strategies accordingly [15-17]. Researchers have compared the three feature selection methods of logistic regression, AIC-Logistic regression and BIC-Logistic regression, and selected the logistic regression model with the best AUC and accuracy rate to construct the personal credit risk assessment index system [18–21]. In research on the application of decision tree models in credit rating, by selecting the top 10 important features and comparing the performance of the tree-based model with that of a multi-layer artificial neural network during the modelling process, the tree-based model was found to be more stable [22, 23]. When comparing the performance of various classification algorithms enabled by Bolasso with other baseline feature selection methods in credit ratings, Bolasso demonstrated impressive feature stability, and the BSRF model outperformed other methods in terms of AUC and accuracy, effectively improving the lender's decision-making process [24, 25].

Small and medium-sized enterprises (SMEs) have a large scale and are full of innovative vitality. They are an important component and core driving force of China's economic and social development, playing a crucial role in promoting market prosperity and economic growth, supporting export earnings, expanding employment, and promoting technological innovation [26]. However, due to factors such as the small scale and low market share of small and medium-sized enterprises, it is difficult to meet the conditions for direct financing. Therefore, in the financing process, they mainly rely on indirect financing, such as commercial bank loans [27]. Therefore, it is particularly important for banks to measure credit risk and rate the creditworthiness of SMEs. The credit risk assessment of SMEs should not be conducted solely from a financial perspective but should also take into consideration the credit and reputation of their actual controllers [28, 29]. There are many empirical studies on credit risk measurement for small and medium-sized enterprises. For example, some researchers have properly and reasonably evaluated the microloans of rural commercial banks based on the Probit model and the Logistic model. The final results show that the debt-to-asset ratio, current ratio, quick ratio, return on total assets, cash-to-debt ratio, corporate credit status, employee age, and corporate size will have different degrees of impact on the risk factors of microloans [30-33]. Some studies have rationally constructed credit risk assessment models through various aspects such as net profit growth rate, effective transaction rate, and profit margin. At the same time, they have rationally quantified internal credit risks of enterprises and effectively provided actual credit loan strategies for banks [34, 35].

Small and medium-sized enterprises are important economic entities in China, but compared with large enterprises, their financial management capabilities and financial data openness are relatively low. Many small and medium-sized enterprises have almost no complete and accurate financial data records and reports, which directly affects the bank's assessment of their credit risk. Therefore, one of the innovations of this study is to use the transaction bill information of upstream and downstream enterprises to measure the credit risk level of small and medium-sized enterprises. In addition, another innovation of this study is that it is based on the traditional decision tree model, improves the traditional decision tree model, adds a regular term function to the original model to restrain the complexity of the decision tree, uses the integrated learning idea to stack multiple decision tree models (weak classifiers), and performs iterative optimization based on the gradient descent algorithm to obtain the final integrated model (strong classifier). Based on the supervised learning model, a credit risk assessment model with default probability as the objective function is established, and the default probability of 120 enterprises (with credit records) is solved to obtain the credit risk of each enterprise. Therefore, this research first establishes a corporate credit rating model based on the improved gradient descent decision tree algorithm and then trains the model using the collected data. Then the bank's specific optimal credit strategy for enterprises is determined. The final result shows the optimal expected return value of the bank, which is a promising outcome of this research. It establishes a nonlinear programming model for credit strategies, such as whether to lend, loan limit interest rate, and term.

The subsequent sections of this paper are organized as follows: Section 2 introduces the data and variables for this study and provides identification and meaning for each variable. In Section 3, the models used in this paper are introduced. Analysis of results is in Section 4, followed by our conclusions and discussion stemming from this research in Section 5 and Section 6.

2- Data and Variables

2-1- Data

This research focuses on small, medium, and micro enterprises in China and collects data on related enterprises from 2016 to 2019. The data was sourced from the Wind database, and the data is based on daily data. Those enterprises had credit record data, including default and reputation ranking. This research does not consider invalid invoice data and directly excludes it.

2-2- Variables

In order to more scientifically evaluate the credit risk of small and medium-sized enterprises, the selection of credit risk evaluation indicators should follow the principles of comprehensiveness, objectivity, independence, and scrupulously consider the impact of various aspects on enterprise credit risk. This article underscores the importance of objectivity in the selection of credit risk evaluation indicators, ensuring a fair and impartial assessment. Seven refined variables, including enterprise scale (x_1), average profit of the enterprise (x_2), average annual profit margin of enterprises (x_3), average annual profit growth rate of enterprises (x_4), average sales growth rate of enterprises (x_5), enterprise service or product quality (x_6), reputation rating (x_7), are selected as measurement variables for risk assessment of small and medium-sized enterprises. The definitions and symbols of relevant variables are shown in Table 1.

Table 1. Definitions and symbols of relevant variables

Symbol	Definition and Description	Unit
j	Enterprise *j*	
k_j	The expected returns of *j* Bank for all customer enterprises with different credit ratings *k*	yuan
Z	The expected return of the bank on the all-customer enterprise *j*	yuan
t_j	The expected return of the bank on the current customer enterprise *j*	yuan
A_j	Possible loan amount for enterprise *j*	yuan
B_j	The actual loan amount of enterprise *j*	yuan
y_j	Indicator for determining whether the bank has lent to existing customer enterprises	
i_j	Enterprise *j* loan interest rate	%
P_j	The repayment probability of enterprise *j*	
q_j	The probability of customer churn in enterprise *j*	
M_j	The customer churn value of enterprise *j*	
d	Annual total credit amount of the bank	yuan
R_j	The default probability of enterprise *j*	%
x_1	Enterprise scale	
x_2	Average profit of the enterprise	yuan
x_3	Average annual profit margin of enterprises	yuan
x_4	Average annual profit growth rate of enterprises	
x_5	Average sales growth rate of enterprises	
x_6	Enterprise service or product quality	
x_7	Reputation rating	

3- Research Methodology

3-1- Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees. It introduces randomness and ensemble strategies to accurately model nonlinear complex problems. It is widely used in credit risk assessment and credit strategy formulation. Its essence is to combine multiple decision trees into a strong classifier or regression model to improve the accuracy and robustness of prediction.

The algorithm process of Random Forest, known for its efficiency, mainly includes two parts: randomness introduction and ensemble prediction. First, several training set subsets are generated by bootstrap sampling technology. Assuming that the original training set is $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, Random Forest generates B training subsets $D_1, D_2, \ldots, D_B$ by sampling D with replacement B times. The number of samples in each subset is the same as that of the original data set, but may contain duplicate samples. Based on these subsets, B decision trees are constructed.

In the node division process of each decision tree, Random Forest introduces a systematic approach to feature selection. In the division of a certain node, all features F are no longer considered, but a feature subset $F' \subset F$ is selected in a systematic manner, and then the optimal split point of each feature in the subset is calculated. Assuming that the sample set of the current node is S, the split point t of feature f_j can be measured by information gain ΔH or Gini index ΔG:

The information gain calculation formula is:

$$\Delta H(S, f_j) = H(S) - \left(\frac{|S_1|}{|S|} H(S_1) + \frac{|S_2|}{|S|} H(S_2)\right) \tag{1}$$

where $H(S) = -\sum_{c \in C} p_c \log_2 (p_c)$, p_c represents the sample proportion of category C, S_1, S_2 are the two child node sample sets after splitting.

The Gini index calculation formula is:

$$\Delta G(S, f_j) = G(S) - \left(\frac{|S_1|}{|S|} G(S_1) + \frac{|S_2|}{|S|} G(S_2)\right) \tag{2}$$

where $G(S) = 1 - \sum_{c \in C} p_c^2$.

The optimal splitting feature and splitting point of the current node are determined by the above criteria, and the decision tree is recursively generated until the stopping condition is met. In the prediction stage, the random forest generates the final prediction by integrating the output results of each decision tree. For classification problems, the voting method, a powerful technique, is used to combine the predictions of individual trees:

$$\hat{y} = \text{mode}\{\hat{y}^{(1)}, \hat{y}^{(2)}, \ldots, \hat{y}^{(B)}\} \tag{3}$$

where $\hat{y}^{(i)}$ is the prediction result of the i-th tree, and mode represents the mode of the voting results. For regression problems, the mean method is used:

$$\hat{y} = \frac{1}{B} \sum_{i=1}^{B} \hat{y}^{(i)} \tag{4}$$

Random forests also have an important feature: assessing feature importance. Feature importance is calculated based on the contribution when the tree splits, and the cumulative value of information gain or Gini index is usually used to measure the importance of the feature f_j:

$$I(f_j) = \sum_{t \in T} \Delta H(S_t, f_j) \tag{5}$$

where T is the set of all nodes containing f_j.

3-2- AdaBoost Integrated Learning Algorithm

Adaptive Boosting (AdaBoost) is a typical ensemble learning algorithm. Its primary function is to construct a strong learner with stronger overall performance by combining multiple weak learners (weak learners) whose performance is slightly higher than random guessing. This combination is usually achieved by weighted voting (classification tasks) or weighted averaging (regression tasks), and the distribution of weights is based on the performance of each weak learner on the training data. The adaptability of AdaBoost is reflected in its mechanism for dynamically adjusting sample weights: for samples misclassified by the previous round of base classifiers, their weights will be improved in the next round of training, thus guiding subsequent weak learners to pay more attention to these difficult-to-classify samples. This gradual adjustment strategy allows the algorithm to iteratively optimize the overall performance of the final model and significantly reduce training errors.

Specifically, for the given training data set $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, where x_i represents the feature vector and $y_i \in \{-1, +1\}$ represents the corresponding label , AdaBoost first initializes the weight of each sample to be equally distributed, that is:

$$w_i^{(1)} = \frac{1}{N}, \forall i = 1,2, \dots, N \tag{6}$$

where $w_i^{(1)}$ represents the weight of the ith sample in the iteration. Next, the algorithm trains a weak learner $h_t(x)$ in each iteration and calculates its classification error ϵ_t:

$$\epsilon_t = \sum_{i=1}^{N} w_i^{(t)} \cdot I(h_t(x_i) \neq y_i) \tag{7}$$

where $I(\cdot)$ is an indicator function, which takes the value of 1 when the condition is met and 0 otherwise. If $\epsilon_t > 0.5$, it means that the performance of the current weak learner is not as good as random guessing, and the training needs to be terminated or the model needs to be re-adjusted. In order to measure the contribution of the weak learner, AdaBoost calculates its weight coefficient α_t based on ϵ_t, as follows:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{8}$$

α_t reflects the classification ability of the weak learner. When ϵ_t is small (the weak learner has better performance), the corresponding weight α_t is large.

Subsequently, AdaBoost increases its attention to the misclassified samples by updating the sample weight distribution. The specific update formula is:

$$w_i^{(t+1)} = w_i^{(t)} \cdot \exp\,(\alpha_t \cdot I(h_t(x_i) \neq y_i)) \tag{9}$$

And normalize all weights:

$$w_i^{(t+1)} = \frac{w_i^{(t+1)}}{\sum_{j=1}^{N} w_j^{(t+1)}} \tag{10}$$

This weight update mechanism ensures that samples misclassified by the current weak classifier have higher importance in the next round of training, thereby guiding the newly trained weak learners to focus on these difficult-to-classify samples.

After T rounds of iterations, AdaBoost weights all weak learners according to their weight coefficients α_t to obtain the final strong classifier $H(x)$:

$$H(x) = \text{sign}\,(\sum_{t=1}^{T} \alpha_t \cdot h_t(x)) \tag{11}$$

This weighted voting mechanism integrates the individual strengths of weak learners to form a strong learner with higher classification accuracy.

3-3- Based on the Improved Decision Tree Algorithm

This paper improves the original decision tree algorithm model and establishes a credit risk model to further improve the prediction accuracy of the gradient-boosting decision tree algorithm. Assuming that the default probability of credit repayment characterizes a company's credit risk, denoted as R, based on the various evaluation indicators considered earlier, the default probability of credit repayment for each company can be expressed as Equation 12.

$$R_j = \varphi(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \tag{12}$$

The optimization function can be expressed as follow:

$$R_j = \varphi(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \delta(f_t) + constant \tag{13}$$

From Equation 13, l is the loss function, $\delta(f_t)$ is the regularizerconstant is the constant term. The objective function construction and the specific iterative calculation process of the algorithm are as follows:

Step 1: Build a single decision tree

Let $f(x_i)$ be a decision tree function, and the number of leaf nodes is T. Then the decision tree feature selection is performed based on the information gain method. For the decision tree, assuming that the current node is denoted as C, the left child node after the split is denoted as L, and the right child node is denoted as R, then the benefit of the split is defined as the difference between the objective function value of the current node and the sum of the objective function values of its left and right child nodes. The sum of the objective function values for the left and right child nodes is given by Equation 14.

$$Gain = f_C - f_L - f_R \tag{14}$$

In the process of generating a decision tree, the features with the greatest benefit are selected as branches.

Step 2: Based on the additive model, build an ensemble learning decision tree

After establishing a single decision tree model, based on the idea of ensemble learning, multiple decision trees (weak classifiers) are integrated. This article uses an additive model to obtain a strong classifier. At this time, the objective function is an additive model composed of K trees as Equation 15.

$$\hat{y}_l = \sum_{k=1}^{K} f_k(x_i), f_k \epsilon F \tag{15}$$

where f is the kth decision tree, and then the Boosting algorithm is used to train and learn the model. Since the learning model is an additive model, the Boosting algorithm can learn the model from front to back, learning only one basis function and its coefficients (structure) at each step, gradually approaching the optimization objective function, and thus simplifying the complexity of the operation. The method starts with a constant prediction and learns, resulting in Equations 16 to 19.

$$\hat{y}_i^0 = 0 \tag{16}$$

$$\hat{y}_i^1 = f_1(x_i) = \hat{y}_i^0 + f_1(x_i) \tag{17}$$

$$\hat{y}_i^2 = f_1(x_i) + f_2(x_i) = \hat{y}_i^1 + f_2(x_i) \tag{18}$$

$$\hat{y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i) \tag{19}$$

At step t, its objective function can be written as Equation 20.

$$R_j^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) \tag{20}$$

On the other hand, from the second-order expansion of Taylor's formula at point x, Equation 21 is obtained.

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \tag{21}$$

Then the above formula is transformed into Equation 22.

$$R_j^t = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^t) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] \tag{22}$$

The loss function l selected in this article is the square loss function, and the objective function is Equation 23.

$$R_j^t = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{t-1} + f_t(x_i)))^2 = \sum_{i=1}^{n} [2(\hat{y}_i^{t-1} - y_i) f_t(x_i) + f_t(x_i)^2] \tag{23}$$

In Equation 23, $(\hat{y}_i^{t-1} - y_i)$ is the residual. When using the square loss function, the integrated decision tree continuously fits the model by fitting the residual in the previous step model.

Step 3: Based on regular optimization, improved gradient descent decision tree model

Let $f(x_i)$ be a decision tree function, and the number of leaf nodes is T. In order to avoid overfitting problems in the decision tree, this article adds regular terms to the original decision tree model. The complexity of the decision tree is expressed by the regular term as in Equation 24.

$$\delta(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{24}$$

Then at step t, its objective function can be written as Equation 25.

$$R_j^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \sum_{i=i}^{t} \delta(f_i) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{t-1} + f_t(x_i)\right) + \delta(f_t) + constant \tag{25}$$

For a single decision tree, define the set $I_j = \{i|q(x_i) = j\}$ is the set of all training samples divided into leaf nodes. Then Equation 22 can be reorganized into the sum of T independent quadratic functions according to the leaf nodes of the tree, Equation 26 is obtained.

$$R_j^t = \sum_{i=1}^{n} \left[g_i w_q(x_i) + \frac{1}{2} h_i w_{q(x_j)}^2\right] + \gamma T + \frac{1}{2}\gamma \sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T} \left[\left(\sum_{i\epsilon l} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\epsilon l_j} h_i + \lambda\right) w_j^2\right] + \gamma T \tag{26}$$

Definition $G_j = \sum_{i\epsilon I_j} g_i, H_j = \sum_{i\epsilon I_j} h_t$,Then Equation 26 can be expressed as Equation 27.

$$R_j^t = \sum_{j=1}^{T} \left[G_i w_j + \frac{1}{2}(H_i + \lambda) w_j^2\right] + \gamma T \tag{27}$$

Find the first derivative of Equation 27 and set the first derivative equal to 0, leading to Equation 28.

$$w_j^* = -\frac{G}{H_j + \lambda} \tag{28}$$

Then the value of the objective function at this time is Equation 19.

$$R_j^t = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{29}$$

From this, the income for each decision tree split is Equation 30.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{30}$$

At each iteration, the profit and loss are calculated through Equation 30, a new decision tree is generated based on the principle of maximum benefit, and the predicted value corresponding to each leaf node is calculated through Equation 29, and the newly generated decision tree $f_t(x)$ is added to the model again, as in Equation 31.

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \tag{31}$$

Through the above process, iteration is continued, and each iteration generates a weak classifier. Each weak classifier is trained on the basis of the residuals of the previous classifier. In this way, iteration is continued until the target accuracy is reached. The model's iteration flow chart is as shown in Figure 1.

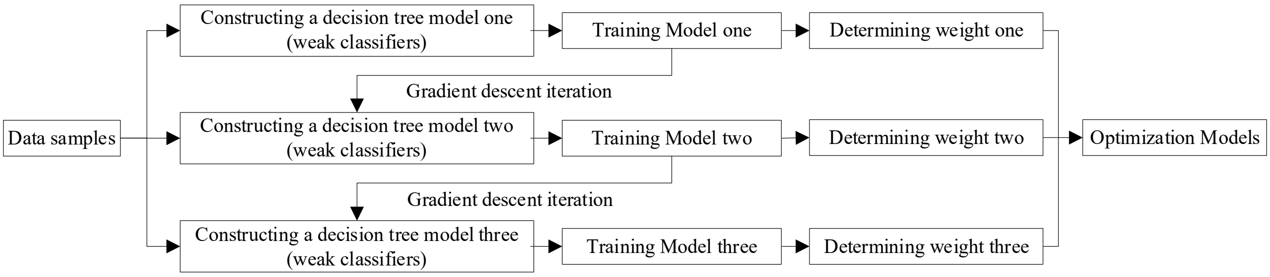

Data samples
Constructing a decision tree model one (weak classifiers)
Training Model one
Determining weight one
Gradient descent iteration
Constructing a decision tree model two (weak classifiers)
Training Model two
Determining weight two
Optimization Models
Gradient descent iteration
Constructing a decision tree model three (weak classifiers)
Training Model three
Determining weight three

Figure 1. Model iteration flowchart

The Gradient Descent Decision Tree (GDDT) algorithm, when compared to traditional decision tree algorithms like C4.5 and ID3 in the context of credit risk assessment, exhibits notable strengths and weaknesses. One of its key advantages is its ability to handle complex, nonlinear relationships between borrower characteristics and credit risk. Unlike ID3 and C4.5, which rely on heuristic criteria such as information gain or gain ratio for splitting, GDDT employs gradient-based optimization, allowing it to capture intricate dependencies within financial data. Moreover, GDDT's robustness to high-dimensional and noisy datasets is a significant advantage, particularly in credit risk assessment, where data can include extensive financial indicators and transaction histories. This robustness provides a sense of reassurance to data scientists and financial analysts, as it ensures the algorithm's applicability to their work. The built-in regularization mechanisms in GDDT, such as L1 and L2 penalties, help mitigate overfitting—an issue that ID3 and C4.5 often face, as they require post-hoc pruning to control model complexity [36-39].

Furthermore, GDDT is well-suited for handling imbalanced datasets, a common challenge in credit scoring where defaults are far less frequent than non-defaults; by adjusting sample weights, it ensures that minority classes receive adequate attention, unlike traditional decision trees that may be biased toward majority classes. Lastly, GDDT seamlessly integrates into ensemble learning frameworks such as Gradient Boosting Machines (GBM), XGBoost, and LightGBM, significantly enhancing predictive accuracy over standalone decision trees. However, GDDT also has notable limitations. It is computationally more expensive, requiring multiple iterations over data, whereas ID3 and C4.5 operate in a single-pass manner, making them faster and more efficient for smaller datasets. Additionally, GDDT is hyperparametersensitive, necessitating careful tuning of learning rates, tree depth, and regularization parameters, whereas traditional decision trees have fewer tuning requirements. Another major drawback is its lack of interpretability; while ID3 and C4.5 produce clear, transparent decision paths, GDDT generates complex tree structures that often require advanced techniques such as SHAP values for explanation—an important consideration in financial applications where regulatory compliance demands explainability. Lastly, GDDT can overfit small or noisy datasets if not properly regularized, whereas C4.5, with its built-in pruning, can sometimes generalize better in such scenarios. Ultimately, while GDDT offers superior predictive performance in credit risk assessment, traditional decision trees may still be preferable in situations requiring speed, transparency, and ease of implementation.

Handling feature selection and dimensionality reduction, particularly when integrating financial and non-financial variables into the risk model, requires addressing scaling differences, categorical encoding, and feature interactions. The role of scaling methods is crucial in ensuring comparability between financial and non-financial data. Continuous financial features are standardized using Min-Max scaling or Z-score normalization, preventing skewed influences from variables with larger magnitudes. Meanwhile, non-financial qualitative factors are processed using target encoding or embedding representations, enhancing their predictive power while preserving their categorical nature. Since traditional decision trees struggle to capture complex feature interactions, Gradient Descent Decision Tree (GDDT) leverages gradient-based optimization to automatically learn intricate relationships between financial and behavioral risk indicators.

3-4- Credit Strategy Model-Based on Nonlinear Programming Algorithm

Step 1: Loan lines are divided based on the company's operating income

This research examines the bank's loan limit for each loanable enterprise, which ranges from 0.1 to 1 million yuan. To effectively manage bank lending risks and ensure greater bank loan profits, it is imperative for banks to provide different loan amounts to enterprises of different sizes and strengths. This approach, which tailors loans to the specific needs of each enterprise, is a key strategy for banks. Most banks estimate the future capital needs of each company based on the company's operating income, capital turnover rate, and other indicators, and then grant corresponding loan amounts, usually 10% to 20% of the company's operating income. This article focuses on the operating income and scale classification of enterprises in various industry categories, using the operating income of enterprises in most industry categories as a benchmark to divide the loan quotas of enterprises of different sizes. Our research is centered on small, medium, and micro enterprises, with large-scale enterprises not considered. The size classification of individuals is retained based on the data characteristics, as shown in Table 2.

Table 2. Loan Limits for Enterprises of Different Sizes

The size of the enterprise	The maximum amount of the loan/yuan
Big	—
middle	1 million
Small	1 million
Micro	0.6 million
Individual	0.4 million

Step 2: Divide customer churn value based on business size

In this study, bank loan interest rates influence the customer churn rate. For a bank, the more customers it loses, the more future loan income it will lose from these customers, and the more it will affect the bank's reputation. However, it's important to note that the potential for introduction and development is significant, and managing customer churn effectively can lead to new opportunities. The cost of attracting new customers to the bank is also the loss caused by the loss of customers. In order to better measure the impact of lost customers on bank loan income, this article introduces the concept of customer loss value, based on the maximum loan income that the size and strength of the lost customer can bring to the bank, whichever amount of 10% to 25% can approximate its customer churn value. Enterprises of different sizes have different customer churn values, as shown in Table 3.

Table 3. Customer churn value for businesses of different sizes

The size of the enterprise	The maximum amount of the loan/yuan
Big	—
middle	10000
Small	5000
Micro	2000
Individual	1000

Step 3: Do not grant loans to companies with a high probability of default

In the credit risk assessment model, the default probability of each enterprise has been obtained. When the default probability of the enterprise is too high, the enterprise is very likely to fail to repay, causing the bank to suffer losses. Therefore, this research uses the bank's expected return t_j for its current customer company j to measure the company's credit risk. Let A_j be the loan amount of the enterprise, i_j be the enterprise loan interest rate, and P_j be the repayment probability of the enterprise. Equation 32 is obtained.

$$t_j = A_j \cdot i_j \cdot P_j - A_j \cdot 0.22 \cdot (1 - P_j) \tag{32}$$

Suppose $t_j \leq 0$, The company is considered to have a high risk of default and the bank will not lend, and $y_j = 0$.Suppose $t_j > 0$, It is believed that the risk of default of the company is low, and the bank lends, and $y_j = 1$.

This formula represents the expected repayment amount of Enterprise J's loan after one year and the expected income that the bank can obtain. If the enterprise does not default and repays normally, the bank can get the profit of the loan amount, which is the income; if the enterprise defaults and cannot repay normally, the bank will suffer a greater loss. Considering that corporate default does not mean complete non-repayment of principal and interest, and small and micro enterprises have a higher probability of default, if calculated based on the total loss of principal and interest, almost all small and micro enterprises will have negative income expectations and cannot be issued loans. Therefore, this article refers to people's bank of China standard. Based on previous research, the one-year average loan default loss rate of 0.22 was selected for calculation. This rate was chosen to avoid defaulting to a complete loss of loans, which would lead to too many companies with high default risks and being unable to issue loans. Therefore, the final loan amount of the enterprise is expressed as B_j, then Equation 33 is obtained.

$$B_j = A_j \cdot y_j \tag{33}$$

Only then will the bank lend to the enterprise and set a specific loan amount and interest rate.

Step 4: Bank's credit strategy model

The development of the bank's credit strategy model will begin. Banks are not allowed to lend to enterprises with a credit rating of D, and are not allowed to lend to enterprises with a high probability of default. Establishing a credit disbursement strategy model based on nonlinear programming algorithm. The bank's credit strategy is shown in Figure 2.

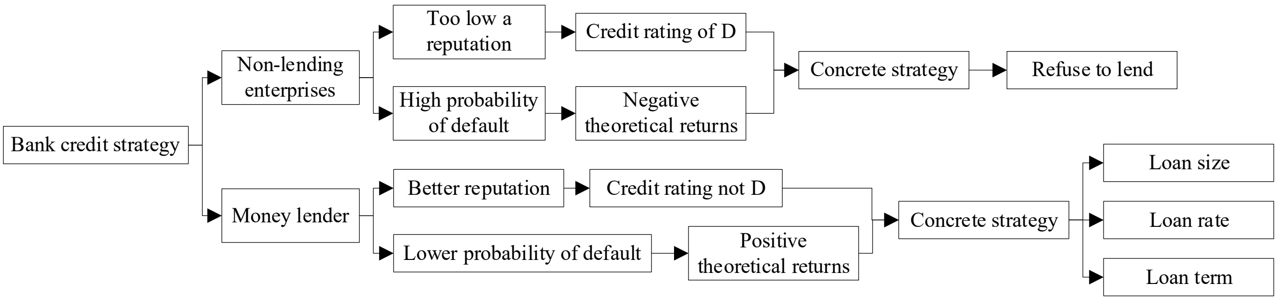


Figure 2. The bank's credit strategy

For a bank, lending to different enterprises aims to maximize loan returns while minimizing customer churn and default risk. This way, the bank can obtain relatively stable and generous loan returns, as well as a stable customer base, which is conducive to the development and profitability of the bank's business. Therefore, this paper establishes a multi-objective programming model with the goals of maximizing loan returns, minimizing customer churn rates, and minimizing customer default risks. Based on the known conditions and assumptions of the problem, the planning model needs to meet the following constraints:

(1) Banks have a specific limit on the loan amount for different enterprises, and the total loan amount for all enterprises does not exceed the bank's loan limit.

(2) Banks have a specific range of loan interest rates for different enterprises.

(3) Banks can decide whether to lend to enterprises with different default risks.

(4) Customer churn in banks can cause certain losses to expected returns, and the value of customer churn varies among enterprises of different scales.

The final objective function is Equation 34.

$$maxZ = \sum_{j=1}^{120} y_j\left[A_j \cdot i_j \cdot \left(1 - R_j\right) - A_j \cdot 0.22 \cdot R_j\right]\left(1 - q_j\right) - M_j \cdot q_j \tag{34}$$

The constraints that need to be met are Equations 35 to 43:

$$\sum_{j=1}^{120} A_j \leq d \tag{35}$$

$$10 \leq A_j \leq E_j \tag{36}$$

$$0.04 \leq i_{kj} \leq 0.15 \tag{37}$$

$$t_j = A_j \cdot i_j \cdot P_j - A_j \cdot 0.22 \cdot \left(1 - P_j\right) \tag{38}$$

$$y_j = \begin{Bmatrix} 0, t_j \leq 0 \\ 1, t_j > 0 \end{Bmatrix} \tag{39}$$

$$q = f(i) \tag{40}$$

$$M = \eta(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \tag{41}$$

$$R = \phi(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \tag{42}$$

$$P + R = 1 \tag{43}$$

where; M, P, R, E representing customer churn value, enterprise repayment probability, enterprise default probability, and maximum loan limit respectively.

Equation 34 ensures that even if the loan interest rate is high, if the default risk is too great, it will still be subject to punitive adjustments, preventing banks from adopting overly aggressive credit strategies. Equation 35 ensures that the bank's total loan amount does not exceed its available funds, preventing liquidity risk. Equation 36 limits the size of loans and prevents banks from granting excessive credit to a single enterprise. Equation 37 limits the range of loan interest rates, ensuring that the loan business is profitable but does not cause borrowers to default due to excessively high interest rates. Equation 38 measures the net return on lending while taking default risk into account, allowing banks to screen out truly profitable loans. Equation 39 allows only profitable loans to be approved, ensuring that the bank does not increase its bad debt ratio due to unprofitable loans. Equation 43, the sum of the loan repayment probability and the default probability must be 1 to ensure the rationality of the credit risk assessment and can be adjusted according to the market environment. Equation 41 and Equation 42 can dynamically adjust credit risk forecasts based on macroeconomic indicators, corporate financial conditions, market fluctuations and other factors to improve the adaptability of decision-making. The nonlinear programming model finds the best balance between maximizing loan returns and minimizing credit risk through constrained optimization. At the same time, its dynamic adjustment framework ensures real-time optimization based on market conditions and financial risks, making it a robust credit risk assessment tool and improving the bank's long-term profitability and risk resistance.

4- Analysis of Results

4-1-Simulation

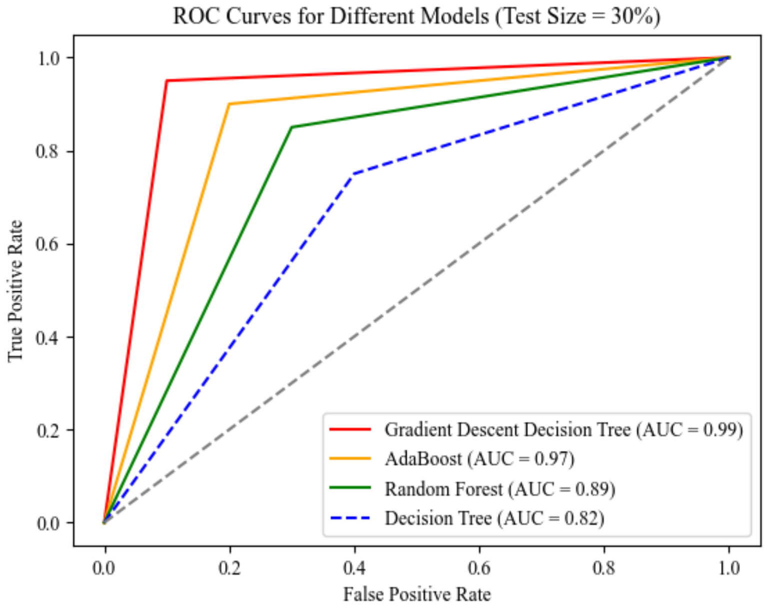
This study, conducted with meticulous attention to detail, evaluated various models to compare their performance in handling imbalanced datasets, a common challenge in credit risk assessment and other classification problems. Imbalanced data often lead to biased model predictions, favoring the majority class, necessitating robust techniques and performance evaluation metrics to ensure fair and reliable model performance. To conduct a thorough analysis, the data was divided into training and test sets using three different ratios—70%, 80%, and 90% for training, with the remaining portions used for testing. Careful attention was given to ensuring the consistency of each test set to maintain the validity and comparability of results across different experimental conditions. The first phase of the evaluation involved using 70% of the data for training. Four machine learning algorithms were employed: the gradient descent decision tree algorithm, the AdaBoost algorithm, the Random Forest algorithm, and the traditional decision tree algorithm. In particular, the reason for model the AdaBoost algorithm was chosen for comparison is that its effectiveness in enhancing weak learners, particularly decision trees, and its widespread application in imbalanced classification problems, such as credit risk assessment. AdaBoost works by iteratively adjusting the sample weights, focusing more on misclassified instances, which is highly relevant in credit risk scenarios where default cases are often underrepresented [40-42].

This makes it a strong benchmark for evaluating the Gradient Descent Decision Tree (GDDT) algorithm, which also aims to improve classification accuracy through iterative optimization. Moreover, when comparing model interpretability, AdaBoost, like other ensemble methods, constructs a series of weak classifiers, making it less transparent than a single decision tree model, although still more interpretable than deep learning approaches. The results suggest that AdaBoost performs well in cases with moderate data complexity, but GDDT offers better adaptability to feature interactions and dataset variability, making it a more robust choice for credit risk assessment tasks requiring precision, scalability, and stability.

These models were chosen for their varying classification and ensemble learning approaches, each offering unique strengths in managing complex datasets. During this phase, each model's training time and predictive performance were meticulously recorded. Tracking training time provided insights into computational efficiency, a crucial factor in real-world applications where resource constraints may impact model selection. The performance was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC), a widely accepted metric for assessing a model's ability to distinguish between classes. A higher AUC value indicates superior classification performance, demonstrating the model's effectiveness in balancing sensitivity and specificity. In the next stage, the training set was increased to 80% of the total data, and the same four algorithms were trained and evaluated. The use of a more extensive training set aimed to improve model generalization by providing more data points for learning patterns. The steps from the first phase were repeated, with training time and AUC values again recorded to measure changes in model performance. Analyzing the impact of additional training data on performance highlights each algorithm's scalability and learning capacity. For some models, increased training data may significantly improve predictive accuracy.

In contrast, for others, diminishing returns may be observed if the model complexity or structure limits its learning potential. Finally, the training set was expanded to 90% of the total data, with the same process repeated for all four models. This final phase provided insights into the performance trade-offs when using an extensive training set, where the test set size reduction may influence the evaluation metrics' robustness. The consistency of test set results remained a priority to ensure reliable comparisons across all experiments. A comprehensive understanding of how different data proportions impact model performance was obtained by calculating the AUC values for each model using the 70%, 80%, and 90% training set splits.

Figure 3 illustrates the ROC curves for each model across the different training set ratios, with the area under each curve serving as the primary performance indicator. The ROC curve visually represents a model's actual positive rate (sensitivity) against its false positive rate (1-specificity) at various threshold settings. The closer the curve approaches the top-left corner, the better the model's classification capability. A model with an AUC value close to 1.0 is considered highly effective, while an AUC near 0.5 indicates poor performance, comparable to random guessing.



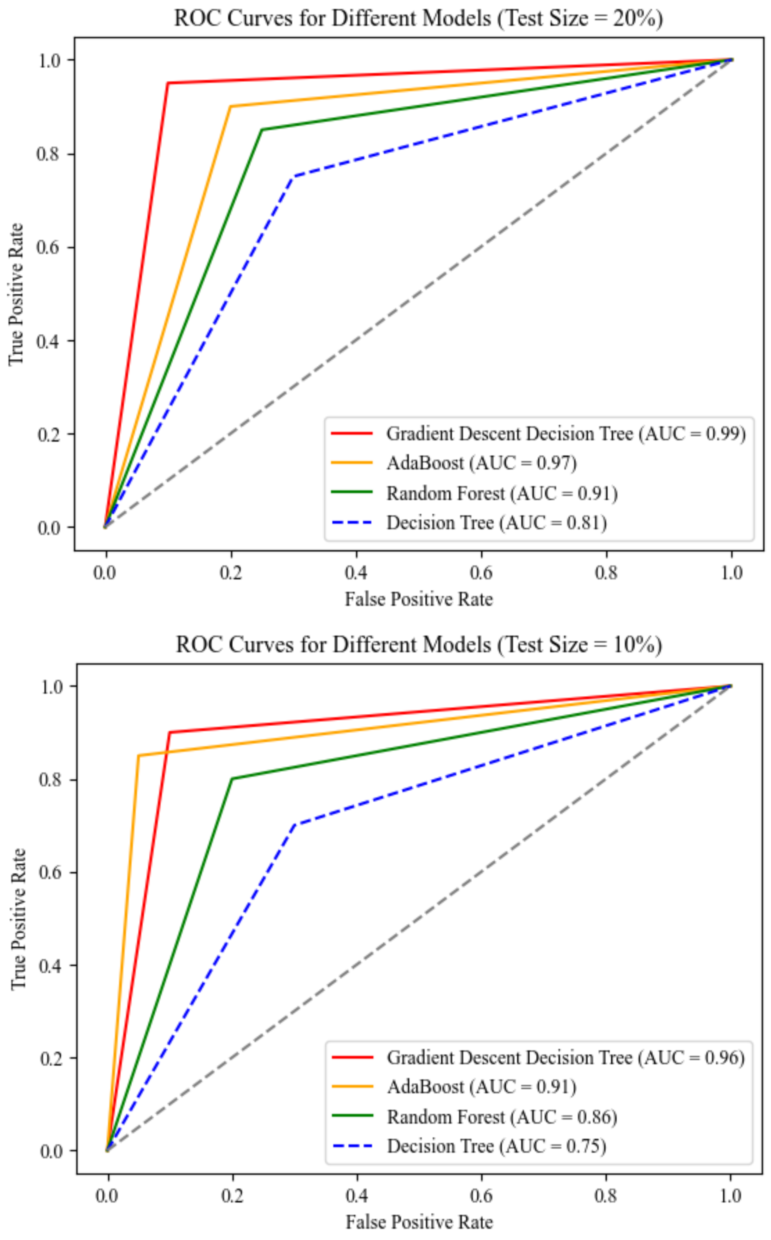


Figure 3. Comparison of results of different models with different proportions of training samples

The ROC curve represents the relationship between the true positive rate (TPR) and the false positive rate (FPR) of the classification model under all possible classification thresholds. The calculation formulas for the true positive rate and false positive rate are:

$$TPR = \frac{TP}{TP+FN} \tag{43}$$

$$FPR = \frac{FP}{FP+TN} \tag{44}$$

Among them, TP represents true positive examples, FN represents false negative examples, FP represents false positive examples, and TN represents true negative examples. The AUC value is the area under the ROC curve, and its value range is between 0 and 1. The larger the AUC value, the better the model performance.

For model training under different training set ratios, the number of samples in the training set N_{train} is first calculated. and the number of samples in the test set N_{test}, where N_{train} is 70%, 80%, and 90% of the number of samples in the training set, respectively, and $N_{\text{test}} = N_{\text{total}} - N_{\text{train}}$, where N_{total} is the total number of samples in the data set.

Through these training and evaluation steps, The AUC values of each model under different training set ratios are recorded. Assuming that the training set ratio is $p \in \{70\%, 80\%, 90\%\}$, the corresponding AUC value can be expressed as:

$$\text{AUC}_{\text{model}}(p) = \int_0^1 TPR(p)dFPR(p) \tag{45}$$

Finally, the experimental results show that the AUC value of the gradient descent decision tree algorithm reaches 0.99 under 70% and 80% training set ratio, which is higher than the AdaBoost algorithm (the maximum value of AUC is 0.97), Random Forest (the maximum value of AUC is 0.91) and the traditional decision tree algorithm (the maximum value of AUC is 0.82). The study introduces an innovative Gradient Descent Decision Tree Algorithm (GDDT), which integrates the principles of gradient-based optimization with decision tree structures to improve predictive performance. Unlike traditional decision trees, which rely on greedy heuristics such as Gini impurity or entropy for splitting criteria, the GDDT algorithm continuously optimizes the tree structure using gradient descent. This enables it to make more precise adjustments during training, reducing both bias and variance and leading to improved generalization capabilities. The Adaptive Boosting (AdaBoost) algorithm is an ensemble learning technique that improves classification performance by iteratively combining weak learners. While AdaBoost is effective in reducing bias and variance, it still relies on weighted averaging rather than a gradient-based optimization approach. The fact that the GDDT achieves an AUC of 0.99 compared to AdaBoost's 0.97 suggests that gradient-based optimization enables more precise decision boundary refinements, reducing misclassification rates for credit risk assessment. Random Forest is another ensemble method that constructs multiple decision trees and aggregates their predictions. It typically provides higher accuracy and generalization than a single decision tree. However, its reliance on bagging (bootstrap aggregation) rather than an iterative optimization process may limit its performance. The substantial gap in AUC values (0.99 vs. 0.91) highlights that the GDDT algorithm learns more effectively from training data, potentially due to its ability to minimize loss functions through gradient-based updates. A standard decision tree splits the data based on greedy heuristics at each node, making local decisions without considering the broader optimization landscape. As a result, it suffers from overfitting and limited generalization. The fact that the traditional decision tree has an AUC of 0.82, much lower than 0.99, reinforces the advantage of using gradient-based optimization in constructing decision trees, allowing the model to optimize splits in a globally optimal manner rather than relying on locally optimal criteria.

When comparing the performance of the gradient descent decision tree algorithm (GDDT) with AdaBoost, random forest, and traditional decision tree in credit risk assessment, in addition to AUC (a comprehensive indicator for measuring classification ability), Accuracy is also an important evaluation criterion. This study selected training sets and test sets with generally good AUC values. In the case of 80% training set and 20% test set, Accuracy was calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{46}$$

Accuracy applies to cases where data is balanced. Suppose the distribution of samples with high and low credit risks is seriously unbalanced. In that case, Accuracy may lead to misleading results (i.e., as long as the model predicts that all samples are low risk, Accuracy may still be high). Therefore, in credit risk assessment, Accuracy needs to be used in conjunction with Precision, Recall, and F1-score to ensure that the model can correctly classify the majority class and effectively identify the minority class (high-risk users). The calculation results are shown in Table 4.

Table 4. Performance comparison of different models

Model	Accuracy	Precision	Recall	F1-score
GDDT	**98.5%**	**97.9%**	**96.2%**	**97.0%**
AdaBoost	96.7%	96.1%	92.4%	94.2%
Random Forest	93.2%	91.5%	88.7%	90.1%
Traditional Decision Tree	86.4%	82.3%	79.1%	80.6%

Table 4 presents a comparative evaluation of machine learning models based on accuracy, precision, recall, and F1-score. The results highlight the effectiveness of ensemble-based boosting methods, which outperform traditional approaches in classification tasks.

Among the models, the Gradient-Boosted Decision Tree (GDDT) achieves the highest accuracy (98.5%), with a well-balanced precision (97.9%), recall (96.2%), and F1-score (97.0%). This indicates that GDDT effectively minimizes false positives and false negatives, making it highly reliable for classification. The strong balance between precision and recall ensures superior generalization performance.

AdaBoost, another boosting-based model, follows with 96.7% accuracy. However, its recall (92.4%) is lower than GDDT's, suggesting it may miss some true positives. Nonetheless, its precision (96.1%) ensures that positive classifications are highly accurate, making it effective in scenarios where false positives must be minimized.

The Random Forest model, which uses multiple decision trees, demonstrates a drop in performance compared to boosting methods, with 93.2% accuracy and an F1-score of 90.1%. Its recall (88.7%) is lower than both GDDT and AdaBoost, indicating that it struggles to capture all positive instances. Though more interpretable and computationally efficient, it offers inferior predictive accuracy.

The Traditional Decision Tree performs the weakest, with 86.4% accuracy, 82.3% precision, and 79.1% recall, resulting in the lowest F1-score (80.6%). These results confirm that while simple and interpretable, single decision trees are prone to overfitting and lack robustness.

Overall, boosting techniques (GDDT, AdaBoost) outperform both Random Forest and Traditional Decision Trees, achieving higher classification accuracy with a better precision-recall balance. These findings highlight the importance of ensemble learning in improving model generalization and predictive performance.

These results show that the gradient descent decision tree algorithm has significant advantages in this research. Next, this study will calculate the default probability of 120 companies using the gradient descent decision tree algorithm.

4-2- Analysis of the Default Probability and Credit Risk

The improved decision tree model in this study is used to evaluate and calculate the probability of default for 120 enterprises, with the detailed results presented in Table 5. The findings reveal a significant degree of variability in default probabilities across different enterprises, underscoring the diverse levels of credit risk within the sample group. This variability reflects a range of factors, including financial health, operational stability, and external market conditions that each enterprise faces. By thoroughly analyzing these variations, banks can develop a deeper understanding of the unique risk profiles of their potential borrowers, enabling more data-driven and effective credit decision-making.

Table 5. Default probability of each enterprise

Number	Probability	Number	Probability	Number	Probability	Number	Probability
C1	0.3663	C31	0.3211	C61	0.4442	C91	0.4697
C2	0.4289	C32	0.3799	C62	0.4918	C92	0.4680
C3	0.3471	C33	0.3338	C63	0.3662	C93	0.4617
C4	0.3662	C34	0.3277	C64	0.4381	C94	0.4571
C5	0.3662	C35	0.4568	C65	0.4613	C95	0.4466
C6	0.3576	C36	0.5337	C66	0.3545	C96	0.6643
C7	0.3689	C37	0.3691	C67	0.4458	C97	0.4600
C8	0.3031	C38	0.4047	C68	0.3870	C98	0.5268
C9	0.3713	C39	0.3783	C69	0.3712	C99	0.6059
C10	0.3403	C40	0.3489	C70	0.4707	C100	0.5468
C11	0.3662	C41	0.4438	C71	0.3442	C101	0.5397
C12	0.3954	C42	0.3601	C72	0.3755	C102	0.5400
C13	0.3642	C43	0.3725	C73	0.4747	C103	0.5224
C14	0.3456	C44	0.3282	C74	0.4200	C104	0.4715
C15	0.4526	C45	0.5395	C75	0.6639	C105	0.4576
C16	0.3611	C46	0.4274	C76	0.4556	C106	0.4696
C17	0.3576	C47	0.3495	C77	0.4609	C107	0.5597
C18	0.3954	C48	0.4608	C78	0.4847	C108	0.5436
C19	0.3662	C49	0.4772	C79	0.4678	C109	0.5408
C20	0.3614	C50	0.4235	C80	0.4654	C110	0.4802
C21	0.4556	C51	0.4425	C81	0.4577\	C111	0.5577
C22	0.4718	C52	0.5395	C82	0.6809	C112	0.7475
C23	0.3576	C53	0.3859	C83	0.4481	C113	0.5646
C24	0.3410	C54	0.5100	C84	0.4834	C114	0.7170
C25	0.3568	C55	0.4844	C85	0.4999	C115	0.6872
C26	0.4526	C56	0.4738	C86	0.4163	C116	0.5535
C27	0.4705	C57	0.5469	C87	0.5482	C117	0.5366
C28	0.3835	C58	0.4464	C88	0.4538	C118	0.5497
C29	0.5332	C59	0.3282	C89	0.4690	C119	0.8345
C30	0.3295	C60	0.4287	C90	0.5133	C120	0.5923

The results from Table 5 illustrate that default probabilities can range from relatively low to critically high, highlighting the uneven risk distribution within the corporate landscape. For instance, Enterprise C1 has a default probability of 0.3663, indicating a moderate level of credit risk. This suggests that while the enterprise may not be completely risk-free, it possesses a relatively stable financial position with a reasonable likelihood of fulfilling its loan repayment obligations. Given this moderate risk level, the bank can consider approving loan requests for this enterprise, but with an appropriate level of due diligence. Lending to Enterprise C1 would likely involve offering standard loan terms while implementing general risk control measures, such as monitoring financial performance periodically and assessing any potential external risks that may affect repayment capability.

Conversely, Enterprise C120 presents a far higher default probability of 0.5923, signaling considerable financial instability or adverse credit conditions that significantly increase the likelihood of repayment failure. A default probability at this level necessitates a much more cautious lending approach. For this enterprise, the bank must conduct a comprehensive and rigorous review of its financial health before making any lending decisions. This process should include an in-depth analysis of financial statements, cash flow projections, and external market influences that may impact the company's ability to meet its debt obligations. Additionally, the bank should explore risk mitigation strategies to safeguard against potential losses. These measures may include requiring additional collateral, setting higher interest rates to compensate for the increased risk, or reducing the loan amount to limit exposure. In extreme cases, where the financial instability appears insurmountable, the bank may ultimately decide to reject the loan application altogether, as the risk of default may be too substantial to justify extending credit.

Beyond individual credit assessments, the insights derived from default probability analysis play a crucial role in shaping the bank's overall lending strategy and resource allocation. By prioritizing loans for enterprises that fall within the low- to moderate-risk category, the bank can optimize its loan portfolio for both profitability and long-term stability. This risk-based allocation strategy allows the bank to maintain a balanced portfolio, ensuring that its credit exposure remains within acceptable limits while still fostering growth in the lending business. At the same time, a targeted risk management approach for higher-risk enterprises helps the bank mitigate the likelihood of significant credit losses, thereby preserving financial health and operational resilience.

This approach aligns with the broader principles of sound credit risk management, which seek to strike a delicate balance between risk and return while ensuring the sustainability and profitability of lending activities. By leveraging advanced decision tree models and data-driven risk assessment techniques, banks can refine their credit evaluation processes, enhance their ability to identify and mitigate potential defaults, and ultimately achieve greater financial stability and growth in an increasingly complex and dynamic economic landscape

The probability of default (PD) is a fundamental indicator used to measure a company's credit risk. It represents the likelihood that a company will be unable to meet its financial obligations, specifically failing to repay a loan within the agreed-upon time frame, typically one year. A default event may manifest in various forms, including delayed payments, partial payments, or complete non-payment of the outstanding loan amount. These different default scenarios impose varying degrees of financial losses on the lending institution. Therefore, the probability of default is a comprehensive measure that encapsulates the potential financial harm a bank may suffer due to a borrower's failure to fulfill its repayment obligations. This indicator not only quantifies the overall risk exposure associated with lending to a particular enterprise but also serves as a critical input for developing effective credit risk management strategies.

In this study, the probability of default is employed as a key metric for assessing the credit risk of individual enterprises. By evaluating this probability, banks can categorize enterprises into different risk tiers, enabling a more granular and targeted approach to lending decisions. A higher probability of default indicates that a company presents a greater risk of non-repayment, which increases the potential for significant credit losses. Such companies are generally characterized by weaker financial health, unstable cash flows, or poor credit histories. Consequently, banks should adopt a more cautious and restrictive lending approach when dealing with high-PD enterprises. This may include imposing higher interest rates, requiring additional collateral, or limiting loan amounts to mitigate potential losses. In extreme cases, banks may choose to deny credit altogether to enterprises deemed too risky.

On the other hand, enterprises with a lower probability of default are considered to have a more favorable credit profile. These companies typically demonstrate strong financial performance, consistent repayment behavior, and reliable cash flow management. A lower PD suggests a reduced likelihood of credit losses, allowing banks to offer more favorable lending terms to these borrowers. For instance, banks may provide lower interest rates, more flexible repayment schedules, or higher credit limits to enterprises with strong creditworthiness. By extending concessions to low-risk borrowers, banks not only foster long-term business relationships but also enhance customer loyalty and market competitiveness.

The ability to accurately estimate and interpret the probability of default is a cornerstone of effective credit risk management. By incorporating PD analysis into their decision-making processes, banks can optimize loan allocation, ensuring that capital is deployed where it is most likely to generate sustainable returns with minimal risk. This scientific, data-driven approach reduces the reliance on subjective judgment, improving the objectivity and consistency of credit evaluations. Moreover, by continuously monitoring changes in PD, banks can proactively adjust their lending strategies to reflect evolving economic conditions or shifts in a borrower's financial health, thereby maintaining robust risk controls.

Assessing the probability of default also contributes to enhanced portfolio risk management. Banks can aggregate individual PDs to evaluate the overall risk profile of their loan portfolio, identifying concentrations of high-risk exposure that may threaten financial stability. By diversifying lending activities and balancing risk across sectors and borrower types, banks can mitigate systemic risks and ensure a more resilient credit portfolio. Additionally, integrating PD analysis with other risk metrics, such as loss given default (LGD) and exposure at default (EAD), provides a comprehensive framework for estimating expected credit losses (ECL), a key component of modern regulatory compliance and financial reporting standards, such as IFRS 9 and Basel III.

Another critical benefit of leveraging PD analysis is its role in strategic decision-making and resource allocation. Banks can use PD-driven insights to prioritize lending opportunities that align with their risk appetite and business objectives. For example, sectors or industries demonstrating low default probabilities may be targeted for growth, while high-risk sectors are approached with caution. Furthermore, PD models allow banks to evaluate the impact of macroeconomic variables on default risk, such as interest rate fluctuations, inflation trends, and changes in regulatory policies. By integrating forward-looking economic indicators, banks enhance their predictive capabilities, positioning themselves to adapt swiftly to emerging risks and opportunities.

In summary, the probability of default is a vital tool for assessing and managing credit risk in banking. It provides a quantitative basis for differentiating between high-risk and low-risk enterprises, guiding more informed lending decisions that enhance the safety and profitability of the loan portfolio. A robust understanding of PD empowers banks to align their credit strategies with sound risk management principles, foster sustainable growth, and maintain financial stability in an increasingly complex and competitive lending environment.

4-3- Analysis of the Bank's Optimal Credit Strategy

The optimal credit strategy for the bank for different enterprises is shown in Table 6. In this context, the annual total credit of the bank is 100 million yuan. Based on the data in the table, the bank decides to provide loans to multiple enterprises and stipulates specific loan amounts and interest rates, where the loan amount and interest rate for each enterprise differ based on its creditworthiness and risk assessment. Enterprise C1 was granted a loan of 21.04 million yuan at an annual interest rate of 5.00%, while enterprise C2 was granted a loan of 96.49 million yuan at an annual interest rate of 6.40%. At the same time, the bank also decided to refuse to provide loans to some enterprises, especially those with a credit rating of D or a negative theoretical income. For other eligible enterprises, the bank provides loans based on their risk assessment and credit rating.

Table 6. Banks' optimal credit strategies for various enterprises (part)

Enterprise Number	Whether to lend	Loan amount/ Ten thousand yuan	Interest rate
C1	YES	21.04	5.00%
C2	YES	96.49	6.40%
C3	YES	96.77	4.05%
C4	YES	96.74	6.67%
C5	YES	56.37	10.27%
C81	YES	36.85	6.58%
C82	N0	-	-
C83	YES	36.68	8.20%
C84	YES	36.73	7.11%
C85	YES	36.57	9.25%
C106	YES	36.05	9.50%
C107	NO	-	-
C108	NO	-	-
C109	NO	-	-
C110	YES	36.05	9.50%

The bank's loan cycle is typically set at one year, with both the loan amount and interest rate tailored to the specific circumstances and creditworthiness of individual enterprises. This flexible, enterprise-specific approach allows the bank to better align its credit offerings with the unique financial needs and risk profiles of its customers. The primary objective behind this loan strategy is to maximize the bank's overall loan income while simultaneously minimizing customer churn and the risk of default. By focusing on optimizing loan structures and terms for a diverse range of enterprises, the bank can foster sustainable growth in its loan portfolio while adhering to sound risk management principles.

This strategy represents a forward-thinking approach that positions the bank for long-term profitability in today's dynamic economic environment. By carefully balancing risk and return, the bank can enhance its competitive advantage in the lending market. This paper, therefore, asserts that implementing a refined credit strategy tailored to individual enterprise needs is the most effective path for banks to navigate the current financial landscape. Not only does it strengthen the institution's ability to generate income, but it also ensures that potential credit risks remain within manageable bounds. The cornerstone of this strategy lies in precision-driven credit assessment and robust risk management frameworks. By employing advanced data analytics, predictive modeling, and comprehensive credit scoring systems, banks can gain deeper insights into borrower behavior and default probabilities. This enables more accurate predictions of creditworthiness and enhances decision-making processes. In doing so, banks can allocate capital more efficiently, directing funds toward enterprises with strong growth potential and reliable repayment capacity. Moreover, the adoption of a dynamic risk-adjusted pricing model allows for flexibility in setting interest rates that appropriately reflect the risk level associated with each borrower. Enterprises with lower risk profiles can benefit from more competitive interest rates, thereby strengthening customer relationships and reducing the likelihood of attrition to competing financial institutions. Conversely, higher-risk enterprises can still access needed capital, but at rates that sufficiently compensate the bank for the additional risk exposure.

In implementing this credit strategy, banks achieve a dual benefit: enhanced profitability through optimized interest rate spreads and improved risk mitigation through diligent credit monitoring. Additionally, by maintaining a customer-centric approach, banks can build lasting partnerships with enterprises by offering not only credit solutions but also financial advisory services that help businesses manage debt more effectively and foster sustainable growth.

Ultimately, this credit strategy underscores the importance of striking the optimal balance between risk and return. By leveraging innovative technologies, advanced credit risk models, and prudent financial management practices, banks can secure the long-term safety of their credit assets while driving sustainable business growth. This comprehensive and adaptive approach positions banks to thrive in an evolving economic environment, reinforcing their role as key enablers of enterprise success and economic development.

5- Conclusion

This study presents a scientifically rigorous and highly accurate credit risk assessment model tailored for small and medium-sized enterprises (SMEs), addressing a critical challenge in modern banking: balancing risk assessment accuracy with optimal credit decision-making. By integrating regularization techniques to control model complexity and employing ensemble learning to enhance predictive performance, the research introduces an improved decision tree model that significantly outperforms conventional methods. The proposed Gradient Descent Decision Tree (GDDT) algorithm, trained through supervised learning with SME default probability as the objective function, iteratively refines weak classifiers using gradient descent optimization to construct a robust predictive model. Empirical results demonstrate the model's superior performance, achieving an AUC of 0.99 under 70% and 80% training set ratios—exceeding the predictive accuracy of AdaBoost (AUC = 0.97), Random Forest (AUC = 0.91), and traditional decision trees (AUC = 0.82). These findings validate the effectiveness of the proposed approach in enhancing the precision of credit risk evaluations. The high AUC of 0.99 may be due to the precise category distribution of the training data, strong feature correlation, and good model optimization. However, replicating the same performance on different datasets depends on the data's similarity, the features' consistency, and the model's generalization ability.

Beyond risk prediction, this research extends its impact by introducing a nonlinear multi-objective programming model designed to optimize bank lending strategies. This model simultaneously maximizes expected loan returns while incorporating risk constraints, ensuring a balanced approach to credit allocation. By integrating advanced machine learning techniques with quantitative risk assessment and optimization strategies, this study provides a practical, data-driven framework for financial institutions seeking to enhance their credit risk management. The combination of regularization, ensemble learning, and nonlinear programming represents a significant advancement in SMEs' credit assessment, offering a scalable and effective solution for modern banking systems. These findings not only contribute to the academic field of credit risk modeling but also provide actionable insights for financial institutions aiming to refine their lending strategies and improve overall financial stability.

6- Declarations

6-1- Author Contributions

Conceptualization, P.B.; methodology, P.B.; software, G.C.; validation, G.C. and N.C.; formal analysis, P.B., G.C., and S.S.; investigation, P.B.; resources, G.C.; data curator, G.C.; writing—original draft preparation, G.C.; writing—review and editing, P.B. and A.V.; visualization, P.B. and G.C.; supervision, P.B. and N.C.; project administration, P.B. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

Observational data were provided by Wind Economic Database (EDB) (EBD accessed on 10 February 2024) at: *https://www.wind.com.cn/mobile/EDB/en.html*.

6-3- Funding

This research project was financially supported by Mahasarakham University.

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Djebali, N., & Zaghdoudi, K. (2020). Threshold effects of liquidity risk and credit risk on bank stability in the MENA region. Journal of Policy Modeling, 42(5), 1049–1063. doi:10.1016/j.jpolmod.2020.01.013.

[2] Dodd, O., Kalimipalli, M., & Chan, W. (2021). Evaluating corporate credit risks in emerging markets. International Review of Financial Analysis, 73, 101–13. doi:10.1016/j.irfa.2020.101610.

[3] He, S. S., Hou, W. H., Chen, Z. Y., Liu, H., Wang, J. Q., & Cheng, P. F. (2025). Early warning model based on support vector machine ensemble algorithm. Journal of the Operational Research Society, 76(3), 411-425. doi:10.1080/01605682.2024.2360111.

[4] Kaur, S., Singh, A., & Aggarwal, A. (2024). Mean-Variance optimal portfolio selection integrated with Support Vector and Fuzzy Support Vector Machines. Journal of Fuzzy Extension and Applications, 5(3), 434–468. doi:10.22105/jfea.2024.453926.1453.

[5] Doumpos, M., & Zopounidis, C. (2007). Model combination for credit risk assessment: A stacked generalization approach. Annals of Operations Research, 151(1), 289–306. doi:10.1007/s10479-006-0120-x.

[6] Tandel, V., Gandhi, S., & Tabarrok, A. (2023). Building networks: Investigating the quid pro quo between local politicians & developers. Journal of Development Economics, 164, 103138. doi:10.1016/j.jdeveco.2023.103138.

[7] Quan, J., & Sun, X. (2024). Credit risk assessment using the factorization machine model with feature interactions. Humanities and Social Sciences Communications, 11(1), 1–10. doi:10.1057/s41599-024-02700-7.

[8] Yu, T., Huang, W., Tang, X., & Zheng, D. (2025). A hybrid unsupervised machine learning model with spectral clustering and semi-supervised support vector machine for credit risk assessment. PLoS ONE, 20(1), 316557. doi:10.1371/journal.pone.0316557.

[9] Carling, K., Jacobson, T., Lindé, J., & Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. Journal of Banking and Finance, 31(3), 845–868. doi:10.1016/j.jbankfin.2006.06.012.

[10] Zhu, S. (2025). Research on Credit Rating of Commercial Banks Based on Support Vector Machine—Data from China's Listed Commercial Banks as Evidence. Finance, 15(02), 423–429. doi:10.12677/fin.2025.152044.

[11] Jarrow, R. A., & Turnbull, S. M. (1995). Pricing Derivatives on Financial Securities Subject to Credit Risk. The Journal of Finance, 50(1), 53–85. doi:10.1111/j.1540-6261.1995.tb05167.x.

[12] Giudici, P., Hadji-Misheva, B., & Spelta, A. (2020). Network based credit risk models. Quality Engineering, 32(2), 199–211. doi:10.1080/08982112.2019.1655159.

[13] Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, 249(2), 417–426. doi:10.1016/j.ejor.2015.05.050.

[14] Bulut, C., & Arslan, E. (2024). Comparison of the impact of dimensionality reduction and data splitting on classification performance in credit risk assessment. Artificial Intelligence Review, 57(9), 252 10 1007 10462–024–10904–1. doi:10.1007/s10462-024-10904-1.

[15] Kubo, H., & Sakai, Y. (2011). On long-term credit risk assessment and rating: Towards a new set of models. Journal of Risk Research, 14(9), 1127–1141. doi:10.1080/13669877.2011.571793.

[16] Shen, C., & Wu, J. (2025). Research on credit risk of listed companies: a hybrid model based on TCN and DilateFormer. Scientific Reports, 15(1), 2599. doi:10.1038/s41598-025-86371-7.

[17] Tarigan, P. A., & Manurung, A. H. (2024). Determination of Bank RAROC using Internal and External Factors of the Bank and Size as a Moderating Variable. Ekonomis: Journal of Economics and Business, 8(2), 1611. doi:10.33087/ekonomis.v8i2.2063.

[18] Lopez, J. A., & Saidenberg, M. R. (2000). Evaluating credit risk models. Journal of Banking and Finance, 24(1–2), 151–165. doi:10.1016/S0378-4266(99)00055-2.

[19] Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert Systems with Applications, 41(4), 2052–2064. doi:10.1016/j.eswa.2013.09.004.

[20] Luo, S., Kong, X., & Nie, T. (2016). Spline based survival model for credit risk modeling. European Journal of Operational Research, 253(3), 869–879. doi:10.1016/j.ejor.2016.02.050.

[21] Liu, J., Liu, J., Wu, C., & Wang, S. (2024). Enhancing credit risk prediction based on ensemble tree-based feature transformation and logistic regression. Journal of Forecasting, 43(2), 429–455. doi:10.1002/for.3040.

[22] Stefania, G. C., Claudia, G. D., Guillermo, A. M. L., Gustavo, G., Dario, R. H. J., Alfonso, M. T. F., & Tatiana, A. N. S. (2023). Credit Risk Scoring Model Based on the Discriminant Analysis Technique. Procedia Computer Science, 220, 928–933. doi:10.1016/j.procs.2023.03.127.

[23] Wang, Z., Zhang, X., Zhang, Z. K., & Sheng, D. (2022). Credit portfolio optimization: A multi-objective genetic algorithm approach. Borsa Istanbul Review, 22(1), 69–76. doi:10.1016/j.bir.2021.01.004.

[24] Zhou, X., Jiang, W., & Shi, Y. (2010). Credit risk evaluation by using nearest subspace method. Procedia Computer Science, 1(1), 2449–2455. doi:10.1016/j.procs.2010.04.276.

[25] Kotelnikova, Y., Clark, L. A., Vernon, P. A., & Hayden, E. P. (2014). Development and Validation of the Schedule for Nonadaptive and Adaptive Personality Brief Self-Description Rating Form (SNAP-BSRF). Assessment, 22(1), 3–16. doi:10.1177/1073191114534959.

[26] Kallmuenzer, A., Mikhaylov, A., Chelaru, M., & Czakon, W. (2025). Adoption and performance outcome of digitalization in small and medium-sized enterprises. Review of Managerial Science, 19(7), 2011–2038. doi:10.1007/s11846-024-00744-2.

[27] Liang, Y., Zhou, B., & Zhao, S. (2024). Risking or de-risking? The effect of banking competition on large state-owned banks and small and medium-sized enterprise lending: Evidence from China. International Review of Financial Analysis, 94, 103258. doi:10.1016/j.irfa.2024.103258.

[28] Judijanto, L., Hairuddin, S. H., Subhan, S., & Sipayung, B. (2024). Analysis of the Effect of Risk Management and Compliance Practices on Financial Performance and Corporate Reputation in the Financial Industry in Indonesia. The Es Accounting and Finance, 2(03), 177–191. doi:10.58812/esaf.v2i03.293.

[29] Carè, R., Fatima, R., & Lèvy, N. (2024). Assessing the evolution of banking reputation literature: a bibliometric analysis. International Journal of Bank Marketing, 42(5), 1059–1091. doi:10.1108/IJBM-07-2023-0417.

[30] Kim, S., & Kim, E. (2024). Population decline in small and medium-sized cities and spatial economic patterns: spatial probit model of South Korea. Applied Economics Letters, 1–6. doi:10.1080/13504851.2024.2363303.

[31] De la Llave, M., & López, F. A. (2024). Searching for correct specification in spatial probit models. Classical approaches versus Gradient Boosting algorithm. Spatial Statistics, 61, 100815. doi:10.1016/j.spasta.2024.100815.

[32] Campanella, F., Ferri, L., Serino, L., & Zampella, A. (2025). Exploring the link between sustainable performance and credit access: the moderating role of intellectual capital. Journal of Intellectual Capital, 26(1), 205–228. doi:10.1108/JIC-06-2024-0191.

[33] Hou, L., Lu, K., & Bi, G. (2024). Predicting the credit risk of small and medium-sized enterprises in supply chain finance using machine learning algorithms. Managerial and Decision Economics, 45(4), 2393–2414. doi:10.1002/mde.4130.

[34] Li, S. (2024). Leveraging Big Data for SME Credit Risk Assessment: A Novel BP-KMV and GARCH Integration. Journal of the Knowledge Economy, 1–29. doi:10.1007/s13132-024-01995-w.

[35] Hou, L., Bi, G., & Guo, Q. (2025). An improved sparrow search algorithm optimized LightGBM approach for credit risk prediction of SMEs in supply chain finance. Journal of Computational and Applied Mathematics, 454, 116197. doi:10.1016/j.cam.2024.116197.

[36] Fakir, Y., Azalmad, M., & Elaychi, R. (2020). Study of The ID3 and C4.5 Learning Algorithms. Journal of Medical Informatics and Decision Making, 1(2), 29–43. doi:10.14302/issn.2641-5526.jmid-20-3302.

[37] Chanmee, S., & Kesorn, K. (2023). Semantic decision Trees: A new learning system for the ID3-Based algorithm using a knowledge base. Advanced Engineering Informatics, 58, 102156. doi:10.1016/j.aei.2023.102156.

[38] Mei, L. (2025). Application and Optimization of Decision Tree ID3 Algorithm in Innovation and Entrepreneurship Analysis Model. International Journal of High Speed Electronics and Systems, 34(2). doi:10.1142/S0129156424400913.

[39] Cirgon, B. Z. (2025). Comparison of C4.5 and C5.0 Methods for Classification of Bad Credit and Good Credit. Formosa Journal of Science and Technology, 4(2), 789–808. doi:10.55927/fjst.v4i2.16.

[40] Ileberi, E., Sun, Y., & Wang, Z. (2021). Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost. IEEE Access, 9, 165286–165294. doi:10.1109/ACCESS.2021.3134330.

[41] Wang, S., Liu, W., Yang, S., & Huang, H. (2024). An optimized AdaBoost algorithm with atherosclerosis diagnostic applications: adaptive weight-adjustable boosting. Journal of Supercomputing, 80(9), 13187–13216. doi:10.1007/s11227-024-05951-y.

[42] Asbai, N., Bounazou, H., & Zitouni, S. (2025). A novel approach to deriving adaboost classifier weights using squared loss function for overlapping speech detection. Multimedia Tools and Applications, 1-28. doi:10.1007/s11042-025-20718-0.