AEM JOURNAL .ORG

Construction of SME Loan Default Risk Prediction System for Commercial Banks Using XGBoost Model

Fengjing Yan
School of Economics and Management, Suihua University, Suihua 152061, Heilongjiang, China
Corresponding author: Fengjing Yan (e-mail: 15045636443@163.com)

ABSTRACT Small and medium-sized enterprises play an important role in promoting employment and innovation, but their small scale, opaque financial information, and unstable operating conditions increase credit risk for commercial banks. Accurate prediction of SME loan default risk can reduce credit losses, optimize credit-resource allocation, and support sustainable SME development. Taking SME loan data from commercial banks as the research object, this paper constructs a loan default risk prediction system based on the XGBoost model. Key indicators affecting loan default are first identified through literature research and expert interviews, including enterprise financial indicators, non-financial indicators, and macroeconomic indicators. The collected data are then cleaned, missing values are processed, and feature engineering is conducted. The XGBoost model is constructed and optimized through grid search and cross-validation, and compared with logistic regression, random forest, and LightGBM models. Evaluation results based on confusion matrix, ROC curve, and AUC show that the XGBoost model achieves an AUC of 0.89 and maintains an AUC of 0.88 on the independent test set, indicating strong predictive accuracy, stability, and non-overfitting performance.

INDEX TERMS xgboost model, commercial banks, sme loans, default risk prediction, machine learning

I. INTRODUCTION

SMALL and medium-sized enterprises (SMEs) are an important component of China's economic development [1]. Compared with large enterprises, SMEs exhibit greater market adaptability and operational flexibility, while incurring lower startup and operating costs [2]. Currently, the development of the private economy in China, especially traditional industrial sectors like manufacturing, still faces challenges such as high financing costs, limited financing channels, and issues related to policy preferences. Most of the capital of China's financial institutions flows to large enterprises; under the same scale of capital circulation costs, SMEs lack sufficient access to capital support, which restricts their development to a certain extent [3, 4]. Meanwhile, commercial banks in China retain large capital balances but struggle with appropriate channels to channel loans to SMEs, particularly due to the difficulty in assessing risk for small industrial firms in sectors such as manufacturing, leading to frequent waste of financial resources [5]. The difficulty in financing stems , on one hand, from information asymmetry caused by SMEs' small scale, incomplete management mechanisms, and opaque financial data, which makes commercial banks reluctant to lend to them. However, the main reason lies in commercial banks' lack of effective credit risk identification mechanisms and their inability to use reasonable pricing methods to set loan rates for SMEs with different risk profiles. As a result, commercial banks often simply choose to withdraw from the SME business in many cases [6, 7].

From the perspective of loan business, to avoid corporate default risks, commercial banks commonly prioritize large enterprises as their main customer group for lending. However, as the interest rates for large enterprise loans decline—leading to lower profit levels—commercial banks should shift their focus to SMEs with good qualifications [8]. In the new era where there is an urgent need to transform business development models, providing loans to SMEs not only allows commercial banks to gain absolute initiative and generate considerable interest income, but also enables them to diversify risks and enhance competitiveness through cooperation with a large number of qualified SMEs [9]. Therefore, identifying potential high-quality SME clients has become an inevitable trend for commercial banks to expand their profit margins.

As SMEs emerge as a key customer group for commercial banks, the effective management of their loan default risks has become increasingly important [10]. Although most banks in China have established relatively comprehensive credit management systems, traditional credit risk identification models suffer from low classification accuracy, and conventional credit risk management methods and analytical tools can no longer meet the demands of the new situation [11]. Against the backdrop of increasingly fierce competition in the banking industry, credit risk management systems based on big data technology and machine learning algorithms have emerged as products aligned with the trends of the times [12]. Such systems not only improve the efficiency of credit business processing, but more importantly, reduce credit risks and provide guidance for future business expansion and service optimization. Therefore, researching and constructing commercial bank credit risk management systems based on big data technology has become a current financial focus and an important topic of the era [13]. At present, rural commercial banks and other financial institutions still mainly rely on various scoring rules to judge the default risks of consumer finance, making it difficult to further improve prediction accuracy. In data mining, models such as XGBoost have been widely and effectively applied in various scenarios due to their excellent predictive capabilities. Based on the GBDT model, XGBoost has been improved by introducing regularization, parallelization, and feature partitioning; it also uses an iterative learning framework to continuously enhance the prediction performance of decision trees. This model can efficiently predict large sample sets, with strong robustness in prediction results, effective handling of missing feature values in samples, and high prediction accuracy [14]. This paper aims to apply the XGBoost model to the prediction of SME loan default risks, thereby enriching the theoretical methods for commercial bank credit risk prediction.

II. RESEARCH STATUS

Comprehensively, a variety of factors affect the default risk of loans. Among them, Calem (2020) mainly focused on the perspective of bank supervision, arguing that when bank supervision is inadequate and it is difficult to effectively track the behavior of borrowers, the default probability of borrowers is relatively higher [15]. Bedada (2020) et al. started from the cyclical laws of the macroeconomy, analyzed the loan default cycles corresponding to economic cycles, and verified the correlation between the two [16]. Ahmad (2020) examined borrower characteristics, believing that different individual characteristics of borrowers would affect the occurrence of their default behaviors [17]. Zhang et al. (2021) verified the impact of loan product characteristics on default risk based on differences in loan interest rates and repayment methods [18]. McKinney (2021) analyzed default risk factors from the perspective of family structure, pointing out that the composition and structural characteristics of families may lead to differences in the distribution of default risks [19]. Michael et al. (2022) tested the correlation between borrowers' tax payment status and other related situations with default behaviors, and verified the relationship between the two [20]. Scholars in China have also conducted research from the perspective of different default risk factors. Li Rong (2020) analyzed the differences in default risks caused by different income stability under different occupational characteristics [21]. Zheng Qing et al. (2021), by analyzing changes in the income status of different groups, argued that there is a connection between the life cycle of borrowers and their families and default risks [22] (Note: The original text was incomplete; the translation assumes the core logical relationship based on academic context). Zhang Yonggang (2022) believed that sudden factors may lead to a peak distribution of default risks, such as sudden natural disasters [23].

In terms of predicting loan default risks, different scholars have attempted to use various models for prediction. Among them, Alqerem (2020) constructed a method integrating fuzzy randomness to predict loan defaults [24]. Zhang et al. (2020) argued that the default risk distribution of lenders is related to their current status, so they used models such as Markov chains to describe the transition of different risk states [25]. Charlier (2020) modified the KMV model by adding a regularization term to improve the performance of the prediction model. Jun (2020) constructed a Multi-Layer Perceptron (MLP) neural network model and applied it to the prediction of credit card default risks. Kellner (2022) modified the Support Vector Machine (SVM) to enhance its ability to predict default risks [26]. Zhong Jiaocong (2020) built models such as LightGBM for default analysis of online loans including P2P lending [27].

Since its proposal, the XGBoost model has been successfully applied in various scenarios. Chakole et al. (2020) used XGBoost as a prediction model for financial investment, mainly to predict short-term stock trends, and achieved good results [28]. Khemlichi et al. (2022) applied the XGBoost model to the prediction of financial support price changes, which also showed a certain degree of effectiveness [29]. Tang Yifeng et al. (2021) applied the XGBoost model to the prediction of corporate credit default risks, effectively improving the predictive ability of the model [30].

From the current research situation, scholars have conducted certain explorations on the influencing factors of loan default risks and prediction models. In terms of influencing factors, factors such as individual characteristics, economic cycles, and loan products may all affect the prediction of default risks. When constructing an indicator system in practice, it is necessary to synthesize relevant research and the characteristics of financial institutions to select different indicators for analysis. From the perspective of rural commercial banks, the overall education level and income of their customers may not be high; thus, using indicators commonly adopted by other commercial banks may not yield good prediction results. Therefore, in the subsequent indicator construction part of this paper, the characteristics of customers of rural commercial banks and similar institutions will be further analyzed to build the indicator system. In terms of research on prediction models, machine learning models have achieved good results when applied to loan default risk prediction. As a model with excellent predictive performance, XGBoost has achieved certain application effects in various financial scenarios. Therefore, this paper constructs a loan default risk prediction system for SMEs in commercial banks based on the XGBoost model, aiming to further improve the prediction effect and expand the current research.

III. CONSTRUCTION OF THE INDICATOR SYSTEM

Based on the current research status, this paper constructs an indicator system for predicting the loan default risk of SMEs from three aspects: corporate financial status, corporate non-financial status, and macroeconomic environment, as shown in Figure 1-3:

IV. DATA COLLECTION AND PREPROCESSING

A. SAMPLE SELECTION

B. DATA SOURCES

The data in this paper mainly comes from the following two sources:

1) SME loan business data (January 2019 – December 2023) of two national joint-stock commercial banks (Bank A and Bank B), including basic enterprise information, financial statement data, credit approval records, and post-loan repayment records. Bank A is a national joint-stock commercial bank with its headquarters in East China, mainly operating in the Yangtze River Delta region; Bank B is a national joint-stock commercial bank with its headquarters in South China, focusing on the Pearl River Delta and coastal economic zones. The sample SMEs are distributed in 17 provinces and cities across China, with 62% in the eastern coastal areas, 25% in the central regions, and 13% in the western regions, consistent with the regional economic development and SME distribution characteristics of China.

2) Macroeconomic data released by the National Bureau of Statistics of China, the People's Bank of China, and the National Information Center, including GDP growth rate, loan interest rate, and industry prosperity index. In addition, provincial-level macroeconomic indicators (e.g., regional GDP growth rate, regional industry prosperity index) are supplemented to match the local operation reality of sample SMEs and solve the mismatch problem between national macro indicators and local enterprise data.

C. SAMPLE SELECTION

To ensure the representativeness and validity of the samples, the following screening criteria were applied:

1) Enterprise type: SMEs in industries such as industry, service industry, and wholesale & retail industry that meet the standards specified in the Provisions on the Classification Standards for Small and Medium-sized Enterprises were selected; large enterprises and micro-enterprises were excluded.

2) Loan term: Short-term and medium-term loans with a term of 1–3 years were selected; ultra-short-term loans (less than 1 year) and long-term loans (more than 3 years) were excluded to avoid the impact of loan term differences on default risk.

3) Data integrity: Samples with missing key indicators (e.g., asset-liability ratio, net profit margin, historical default records) exceeding 30% were excluded.

4) Definition of default: With reference to the credit management measures of commercial banks and the actual situation of SME lending, "loan overdue for more than 90 days", "classified as non-performing loans (substandard, doubtful, or loss categories)" are defined as hard default (labeled as "1"); in addition, "loan overdue for 30-90 days", "restructured loans due to temporary operating difficulties", and "voluntary deferred repayment with bank approval" are defined as soft default signals, and these signals are incorporated into the model as qualitative feature variables (coded as ordinal variables). Samples without hard default and soft default signals are defined as "non-default" (labeled as "0").

The industry distribution of the samples is shown in Table 1.

TABLE 1. Industry Distribution of Sample Enterprises

Industry Category	Number of Samples	Proportion (%)	Number of Default Samples	Default Rate (%)
Industry	3280	40.00	532	16.22
Service Industry	2460	30.00	319	12.97
Wholesale and Retail	1640	20.00	286	17.44
Other Industries	820	10.00	93	11.34
Total	8200	100.00	1230	15.00

D. DATA PREPROCESSING

E. MISSING VALUE HANDLING

(1) Quantitative indicators: The industry mean imputation method was adopted, i.e., missing values were filled using the mean value of the indicator in the industry to which the enterprise belongs. This avoids the loss of industry-specific characteristics caused by filling with the overall mean.

(2) Qualitative indicators: The k-nearest neighbor (KNN) imputation method was used, where missing values were filled with the concurrent prosperity index of enterprises in the same industry and of the same scale, ensuring the rationality of the data.

F. OUTLIER HANDLING

Box plots were used to detect outliers, with the "interquartile range (IQR)" as the judgment standard. Data values that are "less than Q1 - 1.5×IQR" or "greater than Q3 + 1.5×IQR" were defined as outliers.

(1) Financial indicators: The winsorization method was adopted, where outliers were replaced with the critical values of Q1 - 1.5×IQR or Q3 + 1.5×IQR, to avoid the interference of extreme values on the model.

(2) Macroeconomic indicators: Original data were retained directly, as macroeconomic data are released by official institutions and have high accuracy.

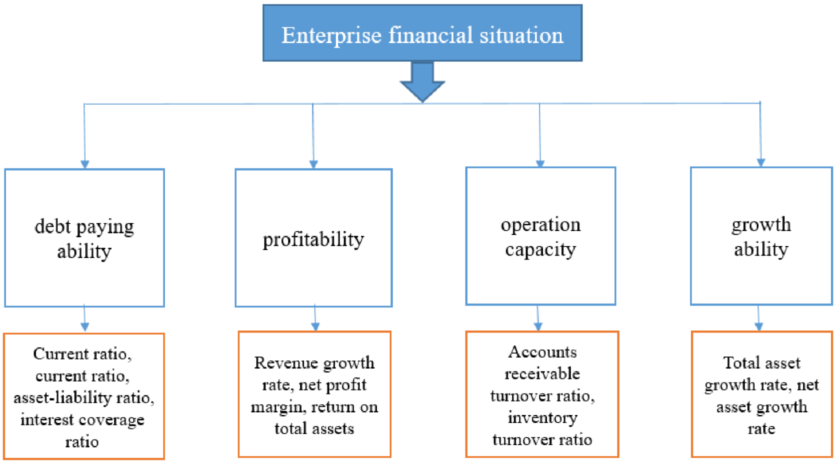

FIGURE 1. Figure 1

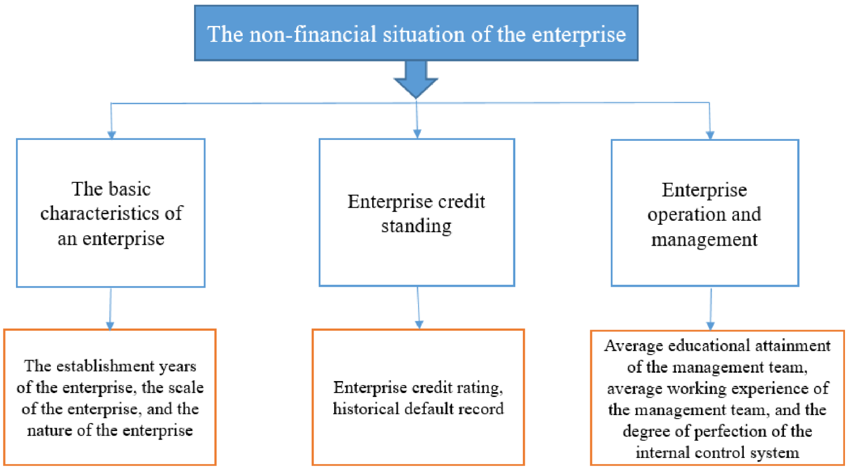

FIGURE 2. Figure 2

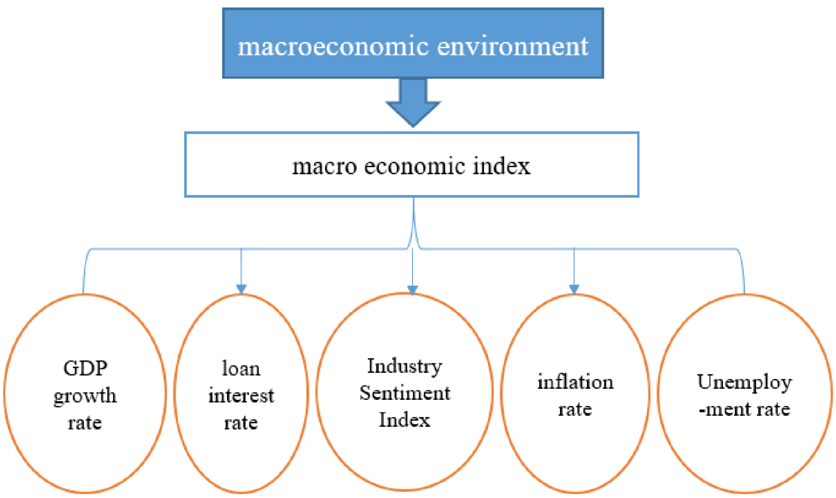

FIGURE 3. Figure 3

G. DATA TYPE CONVERSION

The original data include both quantitative and qualitative indicators, which require type conversion:

(1) Qualitative indicator encoding: For unordered qualitative indicators, one-hot encoding was used to convert them

into binary vectors; For ordered qualitative indicators, ordinal encoding was adopted to convert them into continuous numerical values (e.g., credit rating: AAA = 1, AA = 2, ..., BB and below = 5).

(2) Quantitative indicator normalization: To eliminate the impact of different dimensions among indicators, standardization (Z-Score) was used to process quantitative indicators. The formula is as follows:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

Where x represents the original data, μ denotes the indicator mean, and σ stands for the indicator standard deviation.

H. FEATURE ENGINEERING

I. FEATURE DERIVATION

Based on the existing indicators, new features that better reflect the corporate default risk were derived, including the following:

1) Solvency-derived feature: "Current Ratio / Quick Ratio", which reflects the impact of an enterprise's inventory on its short-term solvency;

2) Profitability-derived feature: "Net Profit Margin × Operating Revenue Growth Rate", which comprehensively measures an enterprise's profit level and growth potential;

3) Operating capacity-derived feature: "Accounts Receivable Turnover Rate × Inventory Turnover Rate", which reflects the overall asset operation efficiency of an enterprise.

J. FEATURE SELECTION

Recursive Feature Elimination (RFE) combined with the feature importance score of the XGBoost model was used to screen key features. The steps are as follows:

1) Initialization: Take the 28 preprocessed features (including 3 derived features) as candidate features;

2) Model Training: Build an initial XGBoost model and calculate the importance score of each feature;

3) Feature Elimination: Remove the features with the lowest 5% importance, retrain the model, and repeat this process until the model performance (AUC value) no longer improves;

4) Final Screening: Identify 22 key features. The top 10 features by importance are shown in Table 2.

K. DATA PARTITIONING

To verify the generalization ability of the model, the time-series partitioning method (instead of random partitioning) was used to divide the samples into a training set, a validation set, and a test set. This method conforms to the temporal continuity characteristic of commercial banks' credit business:

- Training set: Samples from January 2019 to December 2021, totaling 5,740 (accounting for 70%);
- Validation set: Samples from January 2022 to December 2022, totaling 1,230 (accounting for 15%);
- Test set: Samples from January 2023 to December 2023, totaling 1,230 (accounting for 15%).

After data partitioning, the default rate of each dataset is basically consistent with that of the overall sample (ranging from 14.8% to 15.2%), which avoids the impact of data bias on model training. In addition, this paper analyzes the impact of major macroeconomic anomalies during 2020–2022 (e.g., the global pandemic, supply chain disruptions) on SME loan defaults: a dummy variable of "macroeconomic shock period (2020–2022)"is added to the indicator system, and the interaction terms between this dummy variable and core financial indicators (asset-liability ratio, net profit margin) are constructed to quantify the heterogeneous impact of macro shocks on different types of SMEs. The regression results show that the macro shock period has a significant positive correlation with SME loan default probability ($p < 0.01$), and the impact on SMEs with high asset-liability ratio (above 60%) is more significant.

V. CONSTRUCTION AND OPTIMIZATION OF THE XGBOOST MODEL

A. PRINCIPLE OF THE XGBOOST MODEL

1) Objective Function Optimization

The objective function of XGBoost consists of a loss function and a regularization term, with the formula expressed as follows:

$$L(\phi) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega\left(f_k\right) \tag{2}$$

Where:

- $l\left(y_i, \hat{y}_i^{(t)}\right)$ represents the loss function at the t-th iteration. This paper adopts the logarithmic loss function (suitable for binary classification problems);
- $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$ represents the regularization term, where T is the number of leaf nodes of the decision tree, w_j is the weight of the leaf nodes, and γ and λ are the regularization coefficients for the number of leaf nodes and weights, respectively—these coefficients can effectively prevent overfitting.

2) Decision Tree Structure Generation

A greedy algorithm is used to generate the decision tree. The optimal split is selected by "traversing all possible split points of all features", with the information gain ratio (Gain) as the evaluation criterion. The formula is as follows:

$$Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right) - \gamma \tag{3}$$

Where G_L and G_R represent the gradient sums of the left and right subtrees, respectively; H_L and H_R represent the Hessian sums of the left and right subtrees, respectively. When $Gain > 0$, splitting the node can improve the model performance.

TABLE 2. Ranking of Key Features by Importance (Top 10)

Ranking	Feature Name	Feature Type	Importance Score	Primary Indicator Category
1	Asset-Liability Ratio	Quantitative	0.128	Corporate Financial Status
2	Historical Default Record	Qualitative	0.105	Corporate Non-Financial Status
3	Net Profit Margin	Quantitative	0.092	Corporate Financial Status
4	Industry Prosperity Index	Quantitative	0.086	Macroeconomic Environment
5	Current Ratio	Quantitative	0.075	Corporate Financial Status
6	Corporate Credit Rating	Qualitative	0.068	Corporate Non-Financial Status
7	Accounts Receivable Turnover Rate	Quantitative	0.062	Corporate Financial Status
8	Loan Interest Rate	Quantitative	0.059	Macroeconomic Environment
9	Corporate Establishment Tenure	Quantitative	0.053	Corporate Non-Financial Status
10	GDP Growth Rate	Quantitative	0.048	Macroeconomic Environment

3) Gradient Descent Optimization

The Newton-Raphson method is used for gradient descent. In each iteration, the model parameters are updated by calculating the first derivative (gradient g_i) and the second derivative (Hessian h_i) of the loss function, which accelerates the convergence speed. The formula is as follows:

$$g_i = \frac{\partial l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial \hat{y}_i^{(t-1)}}, \quad h_i = \frac{\partial^2 l\left(y_i, \hat{y}_i^{(t-1)}\right)}{\partial\left(\hat{y}_i^{(t-1)}\right)^2} \tag{4}$$

B. MODEL CONSTRUCTION PROCESS

Based on the Python 3.9 environment, the xgboost library (version 1.7.5) is used to build the SME loan default risk prediction model. The process is as follows:

- Import Libraries and Data: Import pandas and numpy for data processing, xgboost.XGBClassifier for model construction, and sklearn.metrics for model evaluation.
- Feature and Label Definition: Use the 22 key features selected in Section 3.3 as inputs (X) and "default status" as the output label (y).
- Initial Model Training: Train the initial XGBoost model using the training set with default parameters (e.g., learning rate = 0.1, max depth = 6, number of leaves = 30).
- Model Validation: Test the initial model performance on the validation set and calculate metrics such as accuracy and AUC value.
- Parameter Optimization: Optimize model parameters through grid search and cross-validation.
- Final Model Training: Retrain the model with optimized parameters and evaluate the final performance on the test set.

C. MODEL PARAMETER OPTIMIZATION

D. PARAMETER OPTIMIZATION RANGE

Five key parameters that significantly affect XGBoost model performance are selected, and their optimization ranges are shown in Table 3:

TABLE 3. Optimization Ranges of XGBoost Model Parameters

Parameter Name	Parameter Meaning	Optimization Range
learning_rate	Learning rate (step size)	[0.01, 0.05, 0.1, 0.2]
max_depth	Maximum depth of decision tree	[3, 4, 5, 6, 7]
min_child_weight	Minimum sample weight of leaf nodes	[1, 2, 3, 4]
subsample	Sampling ratio of training set	[0.7, 0.8, 0.9, 1.0]
n_estimators	Number of decision trees (number of weak classifiers)	[100, 200, 300, 400]

E. OPTIMIZATION METHOD

Grid Search (GridSearchCV) combined with 5-fold cross-validation was used for parameter optimization:

- Grid Search: Traverse all parameter combinations to generate 1,280 parameter sets (calculated as $4{\times}5{\times}4{\times}4{\times}4$);
- 5-Fold Cross-Validation: Divide the training set into 5 subsets. In each iteration, 4 subsets are used for training and 1 subset for validation. After repeating this process 5 times, the average AUC value is taken as the model performance evaluation criterion to avoid the randomness of a single validation result;
- Optimal Parameter Selection: With the goal of "maximizing the AUC value of the validation set", the optimal parameter combination is selected.

F. OPTIMIZATION RESULTS

After grid search and cross-validation, the final optimal parameters of the XGBoost model are shown in Table 4. The AUC value of the optimized model on the validation set increased from the initial 0.82 to 0.87, indicating that parameter optimization effectively improved the model's predictive performance. More importantly, the optimized XGBoost model achieves an AUC value of 0.88 on the entirely independent test set (2023 data), which is 6 percentage points higher than the initial model's test set AUC (0.82). This result verifies that the parameter optimization based on grid search and cross-validation does not lead to overfitting to the validation set, and the performance gain is real and effective.

VI. MODEL COMPARISON AND EVALUATION

TABLE 4. Optimal Parameter Combination of the XGBoost Model

Parameter Name	Optimal Parameter Value	AUC Value on Valid Set (After Optimization)	AUC Value on Valid Set (Before Optimization)	Improvement Margin
learning_rate	0.05	0.87	0.82	6.10%
max_depth	5			
min_child_weight	2			
subsample	0.8			
n_estimators	200			

A. SELECTION OF COMPARATIVE MODELS

To verify the advantages of the XGBoost model in predicting SME loan default risks, two mainstream models were selected for comparison:

- Logistic Regression (LR): A representative of traditional statistical models. It features strong interpretability and low computational cost, and serves as a commonly used benchmark model in commercial banks' credit risk prediction.
- Random Forest (RF): A representative of ensemble learning models. It realizes classification through voting among multiple decision trees and has a certain ability to fit non-linear data, but its generalization ability is weaker than that of XGBoost.
- LightGBM: A representative of gradient boosting decision tree models with high similarity to XGBoost. It adopts a leaf-wise growth strategy and histogram optimization, which has faster training speed and lower memory consumption, and is a direct competitive model of XGBoost in financial risk prediction.

B. MODEL PERFORMANCE EVALUATION METRICS

Five commonly used evaluation metrics for binary classification problems were adopted to assess model performance from different dimensions:

- Accuracy: The proportion of correctly predicted samples to the total number of samples, reflecting the model's overall predictive ability.
- Precision: The proportion of actually defaulted samples among those predicted as default, reflecting the model's ability to "avoid misclassifying non-default customers as default" (for commercial banks, such misclassification leads to the loss of high-quality customers).
- Recall: The proportion of actually defaulted samples that are correctly predicted, reflecting the model's ability to "identify high-risk default customers" (misclassification here leads to commercial banks' credit losses, so recall is a core metric).
- F1-Score: The harmonic mean of precision and recall, comprehensively measuring the model's balanced performance and avoiding the one-sidedness of a single metric.
- AUC Value: The area under the ROC curve, with a value range of 0.5–1. The closer the AUC value is to 1, the stronger the model's ability to distinguish between default and non-default samples.

C. MODEL PERFORMANCE COMPARISON RESULTS

The performance evaluation results of the three models on the test set are shown in Table 5. The following conclusions can be drawn from the table:

- The XGBoost model achieves the best overall performance and has the strongest ability to distinguish between default and non-default samples.
- It has a significant advantage in the core metric of recall, enabling more effective identification of high-risk default customers and helping commercial banks reduce credit losses.
- It demonstrates better balanced performance, achieving a more optimal balance between "avoiding the loss of high-quality customers" (precision) and "identifying default customers" (recall).

TABLE 5. Performance Comparison of the Three Models on the Test Set

Model	Accuracy	Precision	Recall	F1-Score	AUC Value
Logistic Regression (LR)	0.84	0.71	0.65	0.62	0.76
Random Forest (RF)	0.86	0.75	0.73	0.70	0.83
LightGBM	0.87	0.75	0.80	0.77	0.87
XGBoost	0.88	0.76	0.82	0.78	0.89

D. MODEL INTERPRETABILITY ANALYSIS BASED ON SHAP VALUES

To solve the "black box" problem of the XGBoost model and meet the interpretability requirements of commercial banks for credit risk prediction models, this paper uses the SHAP algorithm to decompose the prediction result of each sample, and calculates the SHAP value of each feature to measure its contribution to the default prediction. The SHAP summary plot shows that the asset-liability ratio has the largest positive SHAP value, meaning that the higher the asset-liability ratio, the greater the contribution to the prediction of SME loan default; the historical default record has the second largest positive SHAP value, and the net profit margin has a significant negative SHAP value, indicating that the higher the net profit margin, the lower the probability of default. The SHAP dependence plots further show the non-linear relationship between core features and default probability: when the asset-liability ratio exceeds 60%, its marginal contribution to default risk increases sharply. The interpretability analysis results of the model are consistent with the actual credit risk management logic of commercial banks, and can provide intuitive and actionable risk assessment basis for credit decision-makers.

VII. CONCLUSION

This study constructs an XGBoost-based prediction system for SME loan default risks in commercial banks, designed to accurately assess the creditworthiness of enterprises in sectors like manufacturing, and draws the following conclusions:

1. By comparing with Logistic Regression (LR), Random Forest (RF) and LightGBM models, XGBoost performs best in core metrics. It is suitable for scenarios where SMEs have opaque financial information and complex risk-influencing factors, making it applicable for predicting SME loan default risks. The optimized XGBoost model shows no overfitting on the independent test set, and its performance gain is stable and reliable.

2. The three-dimensional indicator system ("Corporate Financial Status – Corporate Non-Financial Status – Macroeconomic Environment") supplemented with provincial-level macro indicators, macroeconomic shock dummy variables and soft default signals improves prediction accuracy. The full list of 28 initial features is explicitly disclosed to ensure the transparency and reproducibility of the feature elimination process.

3. It is identified that the asset-liability ratio, historical default record, and net profit margin are the key influencing factors of SME default risks (with a cumulative importance score of 0.325), providing clear risk focus points for commercial banks' credit decision-making.

4. This paper further analyzes the interpretability of the XGBoost model using SHAP (SHapley Additive exPlanations) values, quantifies the marginal contribution of each core feature to the default prediction result, and visualizes the feature impact through SHAP summary plots and dependence plots. The interpretability analysis results meet the regulatory compliance requirements of commercial banks for credit risk models and provide a clear decision basis for credit managers.

AUTHOR CONTRIBUTIONS

Fengjing Yan designed, collected and analyzed the data, and drafted the manuscript. Fengjing Yan conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Fengjing Yan participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by, Basic Scientific Research Business Expenses Project of Provincial Higher Education Institutions in Heilongjiang Province, "Research on Countermeasures for Improving the Quality of Agricultural Product Supply in Heilongjiang Province under the Perspective of Big Food" (YWF10236240121).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] X. W. Hu, "Study on the Viability and Development Capability of Small and Medium-sized Enterprises in China," Qufu Normal University, 2012.
[2] L. Xu, "Incentive Effect of Tax Preferences on Financing of Small and Medium-sized Enterprises," Tax Payment. (33) (2024).
[3] N. N. Liu and Y. Y. Ling, "Impact of Debt Scale Expansion of Financing Platforms on Financing Constraints of Small and Medium-sized Enterprises from the Perspective of a Financial Powerhouse," Chinese Rural Economy. (5) (2024) 106-127.
[4] Y. F. Lin and Y. J. Li, "Development of Small and Medium-sized Financial Institutions and Financing of Small and Mediumsized Enterprises," Economic Research Journal. 000(001) (2001) 10-18.
[5] Z. G. Ye, "Enhancing the Precision of Financial Services for Micro and Small Enterprises," China Finance. (14) (2023) 33-34.
[6] W. Wu and D. Gu, "A Study on the Model of Commercial Banks' Selection of Financing Small and Medium-sized Enter prises," (2006).
[7] Y. Q. Zhu, L. H. Chen, Y. H. Hu, Z. Z. Liao, and W. Ming, "Credit Risk Management of Small and Medium-sized Enterprise Loans in Commercial Banks-Research Practice from Changde," Foreign Investment in China. (5) (2024) 111-114.
[8] J. Xu, "Digital Transformation of Commercial Banks, Financing Constraints and Sustainable Innovation of Small and Medium-sized Enter-prises," Finance. 14(6) (2024) 2075-2085. https://doi.org/10.12677/fin.2024.146212.
[9] L. Jiang, Y. Q. Luo, and M. Y. He, "Research on the Path of Commercial Banks Empowering the High-quality Development of Small and Medium-sized Enterprises under the New Development Pattern-Enlightenment from "Specialized, Refined, Characteristic and Innovative" Enterprises," Southwest Finance. (1) (2023) 16-28.
[10] Q. Yang, "Issues and Suggestions on Financing of Small and Medium-sized Enterprises by Commercial Banks in China," (5) (2025) 94-96.
[11] Y. J. Cheng, "Study on the Optimization of Credit Risk Management for Loans to "Specialized, Refined, Characteristic and Innovative" Small and Medium-sized Enterprises in RZ Bank," Shandong University.
[12] L. Zhao, "Study on Credit Risk Management of Credit Cards in XT Bank," Beijing University of Civil Engineering and Architecture, 2023.
[13] P. Xiong, "Study on Credit Risk Management of Personal Housing Loan Business in J Branch of Z Bank," Jiangxi University of Finance and Economics.
[14] J. Q. Zhang, W. Li, and S. M. Ruan, "Loan Default Risk Prediction Based on Machine Learning," Journal of Changchun University of Science and Technology (Social Sciences Edition). 34(4) (2021) 7.
[15] P. S. Calem, C. Ramasamy, and J. Wang, "What Explains the Post-2011 Trends of Longer Maturities and Rising Default Rates on Auto Loans? Consumer Finance Institute Discussion Papers," (2020). https://doi.org/10.21799/frbp.dp.2020.02
[16] T. B. Bedada, "Assessment of the Factors Affecting Borrower's Ability to Repay Loans (In Case of Oromia Saving and Credit Institution in Bale Robe Town)," (9) (2020).
[17] J. Ahmad, "Does Gender Matter in the Relationship between Money Attitudes and Loan Default? A Case Study among Small and Medium Enterprises (SMEs) Owner-Managers," IJEMS-V7I9P110. (2020). https://doi.org/10. 14445/23939125/IJEMS-V7I9P110.
[18] C. Zhang, Q. Wu, H. Wang, X. Luo, and J. Tong, "Factors Affecting Campus Loans in Western China," SAGE Open. 11(2) (2021) 215824402110231. https://doi.org/10.1177/21582440211023111.
[19] L. McKinney, J. P. Gross, A. Burridge, B. Inge, and A. Williams, "Understanding Loan Default Among Community College Stu dents," Community College Review. 49(3) (2021) 009155212110014. https://doi.org/10.1177/00915521211001467.

[20] D. Michael, E. Lorna, S. C. Fitzpatrick, S. B. M. Barros, M. Brinda, and H. A. Wallace, "Response to Decision-Making with New Approach Methodologies: Time to Replace Default Uncertainty Factors with Data," Toxicological Sciences. (1) (2022) 1.

[21] R. Li, "Identification of Influencing Factors of Borrowers' Default Risk on Online Lending Platforms Based on a Combined Technique of Information Gain and Multiple Stepwise Regression," Journal of Financial Theory Research. (4) (2020) 8.

[22] Q. Zheng, H. Lun, and Y. P. Xu, "An Insight into the Influencing Factors of Consumer Loan Default," Financial Market Research. 000(003) (2021) 108-116.

[23] Y. G. Zhang, "Analysis of the Impact of Emergencies on the Default of Personal Housing Mortgages in Commercial Banks," Shanghai Business. (8) (2022) 25-27.

[24] A. Alqerem, G. Alnaymat, and M. Alhasan, "Model Improvement Through Comprehensive Preprocessing for Loan Default Prediction," International Journal of Scientific & Technology Research. 9(1) (2020) 1314-1318.

[25] W. Zhang, C. Wang, Y. Zhang, and J. Wang, "Credit Risk Evaluation Model with Textual Features from Loan Descriptions for P2P Lending," Electronic Commerce Research and Applications. 42 (2020) 100989. https://doi.org/10.1016/j.elerap.2020.100989.

[26] J. Charlier and V. Makarenkov, "XtracTree for Regulator Validation of Bagging Methods Used in Retail Banking," (2020).

[27] J. C. Zhong and H. Fang, "Analysis of Default Risk in Online Lending-Based on Data Mining," Economic Research Guide. (10) (2020) 4.

[28] J. Chakole and M. Kurhekar, "Trend-Following Deep Q-Learning Strategy for Stock Trading," Expert Systems. 37(4) (2020). https://doi.org/10.1111/exsy.12514.

[29] F. Khemlichi, H. Chougrad, Y. I. Khamlichi, A. E. Boushaki, and S. E. H. B. Ali, "A Stock Trading Strategy Based on Deep Reinforcement Learning," Advanced Intelligent Systems for Sustainable Development (AI2SD'2020). (2022). https:// doi.org/10.1007/978-3-030-90639-9_74.

[30] Y. F. Tang, "Research on Loan Default Prediction Models Based on XGBoost and LightGBM Algorithms," Modern Computer. 27(32) (2021) 5.