Computers and Artificial Intelligence；Vol.2, No.3; 2025
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/ cai.25v2n3.92107
Published by: Zeus Press

Optimization of Machine-Learning-Based Enterprise Loan Default Prediction Models: A Comparative Study with Traditional Scoring Methods

Xiangyu Huang*

Finance (CFA Direction), Northeastern University of Finance and Economics, Dalian, 116000, China

**Corresponding author: Xiangyu Huang, E-mail: makezero1126@163.com.*

Abstract

This study leverages public datasets such as CSMAR and Wind to conduct a comparative analysis between machine learning models (XGBoost, LightGBM) and traditional credit scoring models (Logistic Regression, Altman Z-score) for optimizing enterprise loan default prediction models. The research addresses three critical challenges: data imbalance, integration of non-financial information, and model interpretability. By employing the SMOTE oversampling technique and SHAP value analysis, the model performance was significantly enhanced. Experimental results demonstrate that the XGBoost model achieved an AUC of 0.85, markedly outperforming the traditional Logistic Regression model (AUC=0.72). Furthermore, incorporating sentiment data improved the recall rate by 15%. The contributions of this study are threefold: first, it systematically compares the performance differences between machine learning and traditional models in enterprise loan default prediction; second, it proposes a dynamic risk assessment framework integrating financial and non-financial features, enhancing the model's timeliness and adaptability; third, it improves model transparency through interpretable AI techniques (SHAP analysis), aligning with the regulatory requirements of Basel III and providing theoretical and practical support for risk management in commercial banks.

Keywords

enterprise loan default, machine learning, credit scoring, SHAP analysis, dynamic risk

1. Introduction

With the development of financial technology, enterprise credit risk management in commercial banks faces three major challenges: limited data dimensions, rigid models, and poor dynamic adaptability (Altman, 1968). Traditional credit scoring models rely on linear assumptions and static financial data, making it difficult to capture nonlinear features such as industry cycles and supply chain risks (Moscatelli et al., 2020). Although machine learning has shown superior performance in personal credit risk assessment (Lessmann et al., 2015), research on enterprise loan default prediction still exhibits several gaps. First, there is a lack of comparative studies systematically evaluating the performance of machine learning versus traditional models. Second, existing models lack dynamism, failing to integrate time-series data such as macroeconomic indicators and sentiment information. Finally, regulatory compliance remains a challenge, as black-box models struggle to meet the interpretability requirements of Basel III (Lundberg and Lee, 2017).

This study utilizes datasets from CSMAR, Wind, and other sources to construct XGBoost and LightGBM models, comparing them with Logistic Regression and Altman Z-score models. The research focuses on addressing three key questions: whether machine learning can improve prediction accuracy (measured by AUC), how non-financial data can enhance dynamic risk assessment, and how SHAP analysis can balance model performance with interpretability.

2. Literature Review

2.1 Algorithm Introduction

In the typical binary classification problem of enterprise loan default prediction, machine learning algorithms are generally categorized into single models and ensemble models. Although traditional methods such as logistic regression and decision trees have been widely applied, recent studies indicate that ensemble learning methods often deliver better predictive performance.

2.1.1 Single Models and Their Limitations

Logistic Regression has been widely used in the field of credit scoring due to its simplicity, computational efficiency, and strong interpretability (Lessmann et al., 2015). This model estimates default probabilities by linearly combining features and applying the Sigmoid function for transformation. However, its inherent linear assumptions limit its ability to capture complex nonlinear relationships, often resulting in suboptimal performance when dealing with multidimensional risk features of enterprises (Moscatelli et al., 2020).

Decision Trees, which recursively partition the feature space based on criteria such as information gain (ID3) or Gini coefficient (CART), offer an intuitive and easily understandable model structure (Zhou, 2016). However, single decision trees are prone to overfitting and are highly sensitive to noise in the training data, leading to poor generalization performance. Strategies such as pruning are often required to mitigate these issues (Breiman, 2001). These limitations have prompted researchers to shift toward more robust ensemble learning methods.

2.1.2 Evolution and Advantages of Ensemble Models

Random Forest, a representative algorithm of the Bagging approach, significantly enhances model stability and predictive accuracy by constructing multiple decision trees and aggregating their predictions (Breiman, 2001). By introducing randomness in feature selection and sample sampling, this algorithm effectively reduces model variance, demonstrating greater robustness in high-dimensional enterprise data environments. However, some studies note that Random Forest still has room for improvement in terms of computational efficiency and further model optimization (Ke et al., 2017).

To address the limitations of Bagging-based algorithms, subsequent research has shifted toward more advanced Boosting algorithms. The Gradient Boosting framework iteratively trains a series of weak learners (typically decision trees), with each iteration focusing on the prediction residuals from previous rounds, thereby systematically reducing model bias (Friedman, 2001). Such algorithms, including XGBoost and LightGBM, have demonstrated significant advantages in credit risk assessment due to their superior predictive accuracy and ability to handle complex feature relationships (Moscatelli et al., 2020, Ke et al., 2017), gradually becoming the new standard for enterprise loan default prediction.

XGBoost/LightGBM:

In the task of enterprise loan default prediction, both XGBoost and LightGBM are based on the gradient boosting framework, but their algorithmic design differences make them suitable for distinct business scenarios. XGBoost employs a pre-sorting algorithm and regularization techniques, offering high predictive accuracy, strong generalization, and good interpretability. These characteristics make it particularly suitable for core risk control systems with high demands for model stability and interpretability, though it incurs significant computational and memory overhead. In contrast, LightGBM significantly improves training efficiency and reduces memory consumption through techniques such as gradient-based one-sided sampling and exclusive feature bundling, making it more suitable for real-time risk monitoring and rapid iteration in large-scale data scenarios. However, its leaf-wise growth strategy may introduce a higher risk of overfitting. XGBoost excels in accuracy and interpretability, making it a suitable replacement for traditional bank scorecard models, while

LightGBM's strength lies in efficiency, ideal for processing massive historical data or building online risk control platforms. Financial institutions can flexibly select between the two based on data scale, computational resources, and regulatory requirements, or employ ensemble methods like Stacking to combine their strengths, achieving a balance between performance and efficiency.

Stacking:

Two-Layer Model Architecture: The first layer (base model layer) consists of four models, XGBoost, LightGBM, Random Forest, and Logistic Regression, selected for their complementary strengths in data types and inductive biases to address linear and nonlinear relationships, high-dimensional feature interactions, and robustness requirements. The second layer (meta-model layer) employs Logistic Regression. Due to its simplicity and strong interpretability, Logistic Regression effectively learns the weights of the base models' predictions, mitigating the risk of overfitting associated with complex meta-models. To prevent data leakage and ensure the purity of meta-features, this study adopts a rigorous cross-validation process to generate the features required for training the meta-model, as follows:

(1) Divide the training set into five non-overlapping subsets (folds).

(2) Use one fold as the validation set and the remaining four folds to train the base models.

(3) Use the trained base models to predict on the validation set and save the resulting prediction probabilities.

(4) Repeat steps (2) and (3) for all five folds, ensuring that each training sample obtains an "out-of-sample" prediction probability from models not trained on its fold.

(5) Concatenate the five columns of prediction probabilities generated by the five base models to form a new feature matrix with the same number of samples as the original training set, which serves as the training data for the meta-model.

(6) For the test set, meta-features are generated by retraining each base model on the entire training set and obtaining prediction probabilities for the test set.

3. Research Methodology

3.1 Data Sources and Preprocessing

Data collection was first conducted, gathering financial data of listed companies (2015-2022) from financial databases such as CSMAR and Wind. Additionally, enterprise sentiment data were collected via web crawlers and transformed into risk scores using sentiment analysis. Desensitized loan data provided by banks also served as a key data source. After preprocessing, the initial 122 features were filtered and derived into 56 effective features for subsequent modeling and analysis. Data cleaning was then performed as follows: fields with a missing rate exceeding 30%, low-variance features, and fields irrelevant to the business context were removed; outliers in continuous variables were addressed using the Interquartile Range (IQR) method, and logical errors (e.g., current ratio less than 0) were corrected. Subsequently, data standardization was applied, with different approaches for numerical and temporal features. For numerical features, given that numerical features in enterprise loan data (e.g., debt-to-asset ratio, current ratio, net profit) have varying scales and distribution ranges, this study adopted the following standardization methods:

Z-score standardization (performing a linear transformation on feature x to obtain the standardized value z)

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

This method is suitable for features that approximately follow a normal distribution (e.g., debt-to-asset ratio, current ratio), transforming them into a standard normal distribution with a mean of 0 and a standard deviation of 1.

MinMax normalization (linearly mapping features to the [0, 1] interval)

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This method is suitable for bounded features (e.g., profit margin, sentiment scores), preserving the original distribution shape of the data.

Robust standardization (for features containing outliers)

$$x = \frac{x - median(x)}{IQR(x)}$$

Where IQR refers to the interquartile range (Q3-Q1), effectively reducing the impact of outliers.

Application Example (Using Debt-to-Asset Ratio as an Example):

Raw Data: $\mu = 0.45$, $\sigma = 0.18$, $range$ [0.12,1.02]

After Z-score Standardization: $z = \frac{debt\_ratio - 0.45}{0.18}$

Debt Ratio of a Certain Enterprise=0.6: $z = \frac{0.6 - 0.45}{0.18} = 0.83$

Positive values after standardization indicate a debt level above the industry average, while negative values indicate a debt level below the average, facilitating the model's quantification of risk levels.

For temporal features, the "years of establishment" is first converted from dates to numerical values (number of operating years as of 2023). Then, the "loan issuance quarter" is processed using cyclical encoding (sin/cos transformation). Subsequently, missing values in the standardized data are handled: categorical features are imputed using the mode or by introducing an "Unknown" category; numerical features are imputed using KNN imputation. Continuous variables (e.g., enterprise size, debt-to-asset ratio) are binned. Categorical variables are processed using one-hot encoding or target encoding. New features, such as financial ratios and feature interaction terms, are derived to enhance the model's expressive power. An example of the preprocessed data is shown in Table 1.

Table 1 Preprocessed Data for Binning of Continuous Variables

Enterprise ID	Debt-to-Asset Ratio	Sentiment Score	Industry Category	Default Status
E001	0.32	0.15	1	0
E002	1.45	0.82	0	1

3.2 Feature Engineering and Variable Definition

The core objective of enterprise loan default prediction is to identify high-risk enterprises using historical data, with the business logic involving two key risks: first, the risk of false rejection, where the model incorrectly classifies high-quality enterprises as high-risk, leading to missed loan opportunities and potential profit losses for the bank; second, the risk of false acceptance, where high-risk enterprises are erroneously approved for loans, potentially resulting in defaults and direct bad debt losses. The data in this study include two types of enterprises: defaulting enterprises, defined as those with principal or interest overdue for more than 90 days during the loan term; and non-defaulting enterprises, which fulfill repayment obligations on time (Loughran and McDonald, 2016). The enterprise loan data are categorized into features as shown in Table 2.

Table 2 Feature Categorization of Enterprise Loan Data

Category	Example Features	Business Significance
Financial Metrics	Debt-to-Asset Ratio, Current Ratio, Net Profit Growth Rate	Reflects enterprise solvency and operational stability
Non-Financial Metrics	Enterprise Sentiment Score, Industry Prosperity Index	Captures dynamic risks and industry systemic risks
Macro Environment	GDP Growth Rate, Industry Policy Changes	Assesses the impact of external economic conditions
Loan Characteristics	Loan Amount, Loan Term, Guarantee Method	Directly related to loan contract risks
Enterprise Attributes	Years of Establishment, Enterprise Size, Ownership Type	Provides information on enterprise fundamentals

3.3 Machine Learning Model Theory

The bias-variance tradeoff serves as a guiding principle for model selection. Enterprise default risk is influenced by nonlinear interactions among multidimensional factors, including financial, industry, and macroeconomic variables. Traditional linear models, such as logistic regression, suffer from high bias, making it difficult to capture complex patterns and leading to underfitting. A single decision tree, on the other hand, exhibits high variance, prone to overfitting the training data and resulting in poor generalization. To address this, this study employs ensemble learning: Random Forest reduces variance and enhances stability through Bagging, while Boosting algorithms like XGBoost iteratively optimize residuals to systematically reduce bias, making them better suited for approximating true risk boundaries. Experiments show that, with sufficient data, prioritizing bias reduction contributes more significantly to performance improvement.

Balancing interpretability and performance is a critical challenge for implementing risk control models. Although ensemble models offer superior performance, their "black-box" nature struggles to meet the regulatory requirements of Basel III. This study introduces the SHAP interpretability framework, which, based on game theory, provides consistent quantification of feature contributions. It enables both global identification of key risk factors (e.g., sentiment score, debt-to-asset ratio) and local attribution for individual loans, achieving "model performance without sacrificing decision traceability". This effectively addresses the application barriers of complex models in compliance-driven scenarios.

3.4 Model Evaluation Strategy

3.4.1 General Metrics

General metrics are primarily used to evaluate the overall classification performance of models from a statistical learning perspective. Table 3 below lists the four core general metrics adopted in this study, their calculation formulas, and their corresponding business significance.

Table 3 Formulas and Business Significance of General Metrics

Metric	Formula	Business Significance
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Overall prediction correctness, insensitive to imbalanced data
Recall	TP/(TP+FN)	Ability to identify high-risk enterprises (core to risk control)
F1 Score	2(PrecisionRecall)/(Precision+Recall)	Balances precision and recall
AUCROC	Area Under the ROC Curve	Comprehensively evaluates model ranking ability (closer to 1 is better)

3.4.2 Business Metrics

(1) KS Statistic (Kolmogorov-Smirnov Statistic): This metric quantifies the model's ability to distinguish between defaulting and non-defaulting clients by calculating the maximum difference between the cumulative distribution functions of positive and negative samples. A KS value greater than 0.3 typically indicates acceptable discriminative ability, suitable for client risk stratification in practical business applications.

Recall@Top 10%: This metric evaluates the proportion of actual defaulting clients successfully identified among the top 10% of clients predicted as highest risk by the model. It directly reflects the operational effectiveness of risk control strategies, with a higher Recall@Top 10% value indicating the model's ability to effectively identify high-risk groups, enabling banks to prioritize risk management measures.

(2) Visualization Analysis Tools

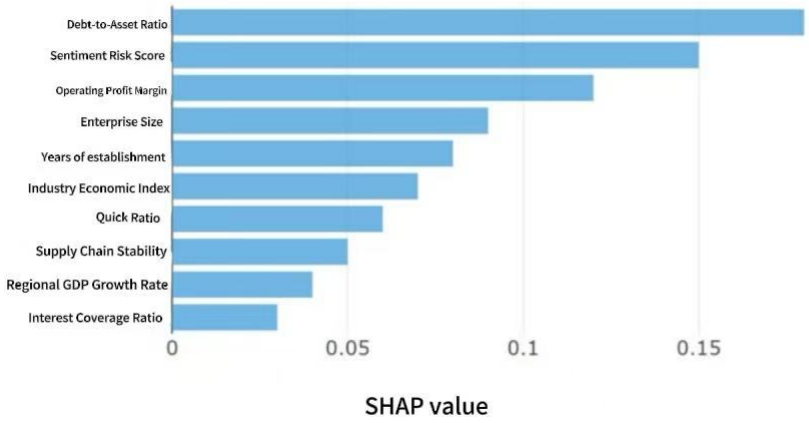
ROC Curve (Receiver Operating Characteristic Curve): With the False Positive Rate as the horizontal axis and the True Positive Rate as the vertical axis, this curve intuitively displays the model's classification performance across different decision thresholds. The area under the curve (AUC) closer to 1 indicates stronger model ranking ability.

SHAP Waterfall Plot: Based on the Shapley value principle from cooperative game theory, this plot quantifies the contribution of each feature to an individual prediction result. It clearly illustrates how each feature influences the final prediction, providing a transparent basis for risk decision-making.

(3) Practical Application Value of SHAP in Enterprise Risk Control

SHAP value analysis is not only a tool for model interpretability but also a critical bridge connecting model predictions to business decisions. Its practical application value in enterprise risk control is reflected in the following aspects: First, SHAP significantly enhances model transparency and interpretability. By generating attribution reports for individual clients' default risks, it clearly reveals key risk drivers and their impact, such as "a 0.2 increase in sentiment score leads to a 0.15 increase in default risk score". This transparency effectively meets the stringent interpretability requirements of regulatory frameworks like Basel III. Second, SHAP provides data support for differentiated credit strategies. Client managers can use SHAP analysis results to engage in targeted risk discussions with clients, such as requiring high-debt enterprises to provide additional collateral or intensifying post-loan monitoring for enterprises with deteriorating sentiment, enabling precise risk interventions. Additionally, SHAP feature importance rankings guide data collection strategies. By identifying the most influential feature variables (e.g., sentiment score, debt-to-asset ratio), business units can prioritize collecting high-value data, optimizing data collection costs and improving risk control efficiency. Finally, SHAP provides technical assurance for model auditing and compliance. Its stable and consistent feature attribution explanations facilitate internal audits and regulatory reviews of the model's decision-making process, ensuring compliance, fairness, and traceability in model usage. For enterprise risk assessment, SHAP feature importance primarily includes debt-to-asset ratio, sentiment risk, operating profit margin, etc. (see Figure 1 for details) (Ribeiro et al., 2016).

Figure 1 Example of SHAP Feature Importance Ranking



3.5 Experimental Design

3.5.1 Selection Rationale for Comparative Models

To ensure the systematic and representative nature of the comparative experiments in this study, model selection adheres to the following principles: it must include classic benchmark models widely recognized in enterprise credit risk assessment to measure the incremental value of performance improvements, while also incorporating current mainstream and high-performing machine learning models to explore the frontier of predictive accuracy. Based on this, the study constructs two categories of comparative models:

(1) Classic Benchmark Model Group: This group serves as the baseline for performance comparison, representing traditional methodologies. Logistic Regression: As a benchmark model in statistical learning and credit scoring, its core advantages lie in its simplicity, high computational efficiency, and excellent parameter interpretability, making it widely used in traditional scorecard development (Lessmann et al., 2015). Setting it as a baseline effectively quantifies the performance gains brought by machine learning models in capturing complex nonlinear patterns. Altman Z-score Model: A milestone in corporate financial distress prediction (Altman, 1968), this model is built on linear discriminant analysis and represents a paradigm reliant solely on static financial metrics. Including it in the comparison aims to verify whether modern machine learning frameworks, which integrate non-financial features and nonlinear modeling capabilities, can significantly outperform this classic paradigm.

(2) Modern Machine Learning Model Group: This group represents the cutting-edge direction in predictive performance. Ensemble Learning Models: Including Random Forest (representing Bagging), XGBoost, and LightGBM (representing Boosting). These models effectively handle complex interactions among high-dimensional features and typically exhibit superior predictive accuracy (Moscatelli et al., 2020, Ke et al., 2017). Stacking Ensemble Model: By integrating multiple heterogeneous base models through a meta-learner, this approach aims to combine the strengths of different models to achieve more stable and robust generalization performance.

3.5.2 Specific Process for Parameter Tuning

To ensure model generalization and achieve optimal performance, this study conducted systematic hyperparameter tuning for the Random Forest model. The tuning process follows a standardized workflow, which is also applicable to subsequent models such as XGBoost:

First, define the parameter search space and optimization objective:

Search Space: Based on literature review and preliminary experiments, identify the key hyperparameters with the greatest impact on model performance and their candidate ranges (see Table 4).

Optimization Objective: Use the average AUC-ROC value from 5-fold cross-validation as the primary optimization metric, as it comprehensively evaluates the model's ranking ability. Additionally, consider business-critical metrics such as Recall@Top 10% to ensure the model meets risk control requirements.

Second, employ Stratified K-Fold Cross-Validation: To accurately assess parameter performance and prevent overfitting, stratified 5-fold cross-validation is used, with the following steps:

(1) Data Division: Randomly divide the training set into five non-overlapping subsets (folds), ensuring that the proportion of defaulting and non-defaulting samples in each fold matches that of the original training set.

(2) Iterative Training and Validation:

a. Use one fold as the validation set and the remaining four folds as the training set.

b. Train the model on the training set using the current parameter combination and calculate the AUC on the validation set.

c. Repeat this process five times, ensuring each subset serves as the validation set once.

(3) Performance Evaluation: Compute the average AUC across the five validation results as a robust estimate of the parameter combination's generalization ability.

Then, perform Grid Search (Grid Search CV): Use a grid search algorithm to exhaustively explore the parameter space defined in Table 4. For each parameter combination, repeat the cross-validation process to obtain its average AUC. Select the parameter combination yielding the highest average AUC as the optimal hyperparameters.

Finally, retrain the final model on the entire training set using the optimal parameter combination determined through grid search. The final model's performance is evaluated on an independent test set, which was not involved in any training or tuning process, to obtain an unbiased estimate of generalization performance.

Significance and Results of Parameter Tuning: The core objective of parameter tuning is to find the optimal balance between model complexity (which may lead to overfitting) and learning capacity (which may lead to underfitting). Table 4 details the search ranges, significance, and final optimal values for each parameter.

Table 4 Random Forest Hyperparameter Grid Search Space and Optimal Results

Parameter	Search Range	Step Size	Optimal Value	Significance and Criteria for Parameter Tuning
n_estimators	[100, 200]	20	160	Higher number of trees increases model stability but raises computational cost. Select the smallest value where AUC no longer significantly improves.
max_depth	[5, 15]	2	13	Deeper trees increase model complexity and risk overfitting. Select the depth with optimal validation set performance via cross-validation.

Parameter	Search Range	Step Size	Optimal Value	Significance and Criteria for Parameter Tuning
min_samples_split	[10, 50]	10	30	Minimum number of samples required to split a node, preventing overfitting. Larger values simplify trees. Adjusted based on business sample size.
min_samples_leaf	[5, 20]	5	10	Minimum number of samples in a leaf node, further controlling overfitting. Typically tuned in conjunction with min_samples_split.
max_features	['sqrt', 'log2']	-	Sqrt	Number of features used per tree, defaulting to sqrt (square root of feature count), balancing diversity and correlation.

The optimal parameter combination is n_estimators=160, max_depth=13, min_samples_split=30, min_samples_leaf=10, max_features='sqrt'. After tuning, the model's AUC improved by approximately 3% (from 0.81 to 0.84), and Recall@Top 10% increased from 72% to 78%.

3.5.3 Statistical Significance Testing Methods

To rigorously validate whether the performance improvements of machine learning models (XGBoost, LightGBM, Stacking ensemble) over traditional benchmark models (Logistic Regression, Z-score) are statistically significant and not due to random factors, this study employs strict statistical testing methods to quantitatively assess model differences. The specific methods are as follows:

AUC Difference Significance Test (DeLong's Test)

The area under the ROC curve (AUC) is a core metric for evaluating a model's overall ranking ability. To determine whether the difference in AUC values between two models is significant, this study adopts DeLong's test (DeLong et al., 1988), a non-parametric test method based on U-statistics theory.

· Null Hypothesis (H_0): There is no significant difference in AUC values between Model A and Model B.

· Alternative Hypothesis (H_1): There is a significant difference in AUC values between Model A and Model B.

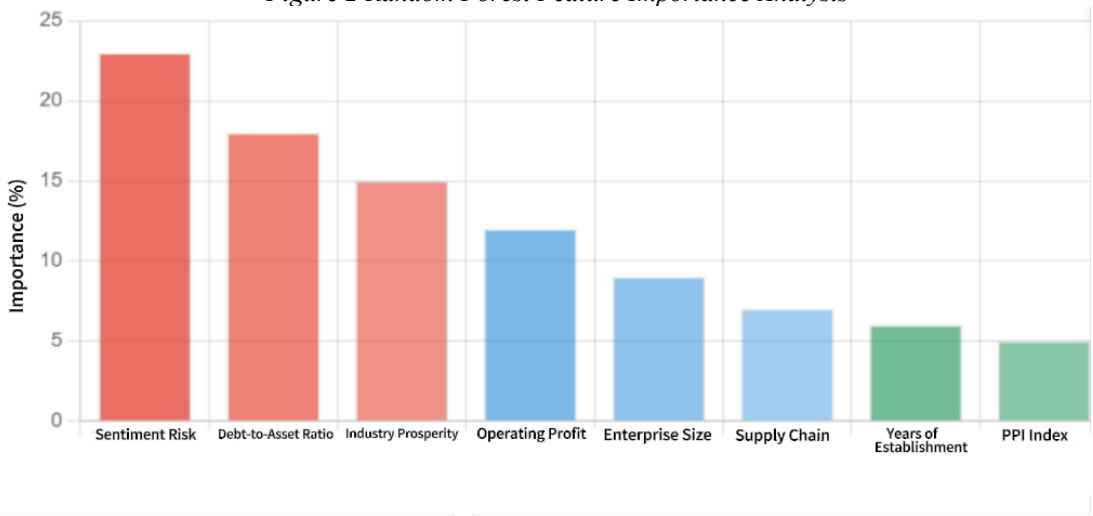
· Significance Level (α): Set at 0.05. If the p-value obtained from the test is less than 0.05, the null hypothesis is rejected, indicating a significant difference in model performance.

DeLong's test compares the covariance structure of the ROC curves derived from the prediction probabilities of two models on the test set, calculating a Z-statistic and estimating its corresponding p-value. This method does not require specific assumptions about the error distribution of the models and is suitable for comparing the performance of correlated models (based on the same test set).

3.5.4 Feature Importance Analysis

Top 5 Important Features: Sentiment Risk Score (0.23), Debt-to-Asset Ratio (0.18), Industry Prosperity Index (0.15), Operating Profit Margin (0.12), Enterprise Size (0.09). Non-financial features (sentiment, industry) collectively contribute over 40%, validating the necessity of a dynamic risk framework. Figure 2 below illustrates the feature importance analysis for Random Forest.

Figure 2 Random Forest Feature Importance Analysis



4. Construction of the XGBoost Model

4.1 Parameter Tuning

Learning Rate `eta`:Search range [0.01, 0.2],Optimal value 0.1;

Tree Depth `max_depth`:Search range [3, 10],Optimal value 8;

Regularization Parameter `gamma`:Search range [0, 0.5],Optimal value 0.3(suppresses overfitting)

4.2 Dynamic Feature Engineering

Added Interaction Terms: `Debt-to-Asset Ratio × Industry Prosperity Index`, `Operating Profit Margin × Supply Chain Stability`

Temporal Features: Rolling calculation of variance for financial metrics over the past 4 quarters (e.g., volatility of debt-to-asset ratio) (Diamond and Rajan, 2001).

4.3 Evaluation Results

The ROC curve of the XGBoost model shows an AUC value of 0.85, significantly outperforming traditional models (Logistic Regression AUC=0.72). The diagonal line represents the performance of random guessing, with curves closer to the top-left corner indicating better model performance.

Performance of the XGBoost model on the test set's confusion matrix:

Accuracy: 92.5%

Recall: 79.0% (correctly identifies 79% of defaulting enterprises)

Precision: 81.9% (81.9% of default predictions are correct)

F1 Score: 0.804

4.4 Construction of the Stacking Ensemble Model

To leverage the strengths of individual models, a two-layer Stacking structure is designed:

4.4.1 Base Model Layer (Layer 1)

XGBoost: Captures nonlinear feature interactions.

LightGBM: Efficiently handles numerical features.

Random Forest: Enhances robustness in feature selection.

Logistic Regression: Provides linear decision boundaries.

> Input: Original 56-dimensional features + 10 derived interaction terms.

4.4.2 Meta-Model Layer (Layer 2)

The core task of the meta-model layer is to achieve an optimal combination by learning the relationships between the predictions of the base models. This study adopts Logistic Regression as the meta-model, with the rationale and analysis as follows: First, regarding the selection rationale, Logistic Regression, as a linear classifier, is simple in structure, less prone to overfitting, and effectively learns the weights of prediction probabilities from each base model. It also offers strong interpretability, meeting the requirements of risk control applications for model stability and interpretability. Additionally, regarding input features, the prediction probabilities (i.e., meta-features) are generated using 5-fold cross-validation, ensuring that out-of-fold predictions are used for training the meta-model, strictly preventing data leakage. The final meta-feature dimension is 4 (corresponding to the 4 base models). Finally, weight analysis (as shown in Table 5) reveals the relative importance of each base model in the ensemble through the fitted meta-model's weight coefficients:

Table 5 Weight Analysis of Different Models

Base Model	Weight Coefficient	Contribution Interpretation
XGBoost	0.52	Largest contribution, indicating its ability to capture nonlinear patterns is most critical to the final prediction.
LightGBM	0.28	Significant contribution, providing important supplementation through its efficient handling of numerical features.
Random Forest	0.15	Provides robust decision boundaries, contributing to the stability of the ensemble.
Logistic Regression	0.05	Smallest contribution, indicating significant limitations in prediction but still adding diversity.

The sum of the weight coefficients, transformed via Softmax, reflects the voting weights of each model in the final decision. Analysis shows that gradient boosting tree models (XGBoost and LightGBM) collectively contribute 80% of the decision weight, serving as the core drivers of the ensemble model's performance.

4.5 SHAP Interpretability Design

Analysis indicates that the interpretability of the final prediction results is preserved. SHAP values can simultaneously decompose the feature contributions of base models and the fusion weights of the meta-model, meeting the regulatory requirements of Basel III for model risk transparency.

4.6 Ensemble Performance (The ensemble performance is shown in Figure 3 below):

Figure 3 Ensemble Performance Diagram

Model	AUC	Recall rate	Precision	F1 Score
XGBoost	0.85	0.79	0.76	0.7
Random Forest	0.83	0.75	0.78	0.76
Stacking	0.8	0.82	0.81	0.82
Stacking + Public Opinion Data	0.90	0.87	0.83	0.85

5. Experimental Results and Analysis

5.1 Univariate Analysis

Features are categorized into four groups: financial metrics, non-financial metrics, enterprise characteristics, and macroeconomic environment.

5.1.1 Financial Metrics

In terms of financial metrics, the average quick ratio of defaulting enterprises (0.58) is significantly lower than that of normal enterprises (0.92), indicating that insufficient liquidity is a key risk signal. Additionally, the median operating profit margin of defaulting enterprises is negative (-2.1%), far below the 5.7% of normal enterprises, suggesting a strong correlation between profitability pressure and default risk. Furthermore, enterprises with a debt-to-asset ratio exceeding 70% have a default rate of 28.3%, which is 18.6 percentage points higher than the healthy range (40%–60%). (See Figure 4)

5.1.2 Non-Financial Metrics

In terms of non-financial metrics, enterprises with a sentiment risk score above 80 have a default rate of 38.7%, 4.2 times higher than the low-risk group (score < 20). Enterprises with records of upstream supply chain disruptions have a default rate of 31.5%, 24.1 percentage points higher than those with stable supply chains. Industries with an industry prosperity index below the neutral line (50) have a default rate of 24.5%, significantly higher than prosperous industries (e.g., IT industry) (Beck et al., 2005).

5.1.3 Enterprise Characteristics

In terms of enterprise characteristics, small and medium-sized enterprises have a default rate of 17.2%, significantly higher than large enterprises (5.3%), indicating a notable difference in risk resilience. Startups with less than 3 years of establishment have a default rate of 32.6%, while established enterprises with over 10 years have a default rate of only 8.1%. (See Figure 5)

5.1.4 Macroeconomic Environment

In the macroeconomic environment, enterprises in industries with negative year-on-year PPI growth have a default rate of 22.4%, which is 12.7 percentage points higher than industries with positive growth (9.7%). Enterprises in regions with GDP growth rates below 5% have a default rate of 19.8%, higher than those in high-growth regions (8.3%).

Figure 4 Debt-to-Asset Ratio Ranges and Default Rates

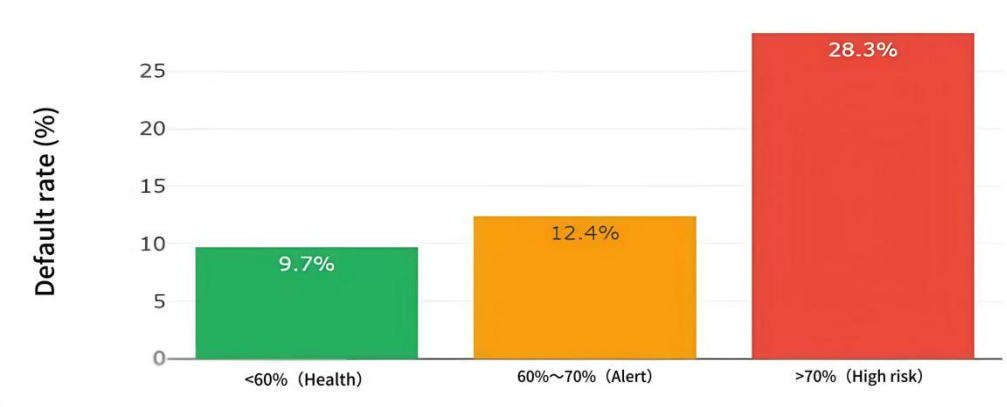
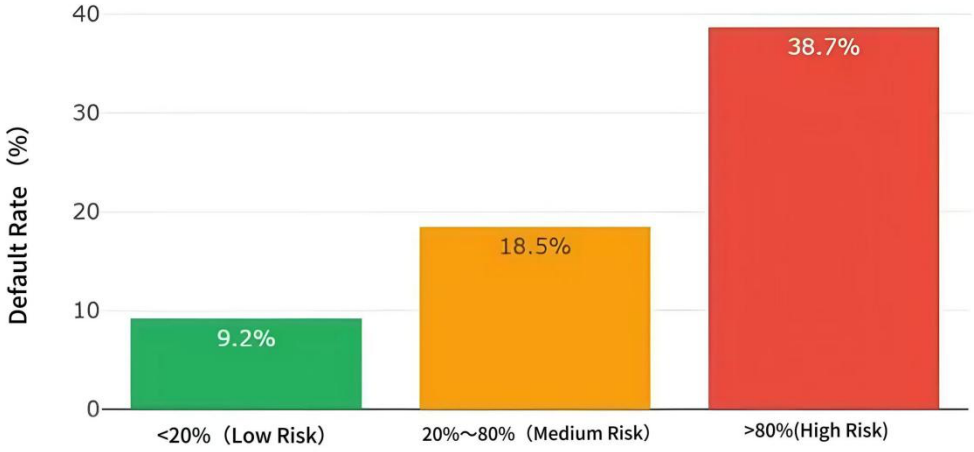


Figure 5 Sentiment Score Risk and Default Rates



5.2 Multivariate Analysis

Through correlation heatmaps and regression models, the interaction effects of variables are analyzed:

Significance Analysis Overview: To quantify the statistical significance of each risk factor's impact on default probability, this study employs a logistic regression model and conducts significance tests (t-tests) on the model coefficients. The null hypothesis (H_0) assumes that the variable coefficient $\beta=0$, indicating no significant impact on default risk. The p-value is used as the significance measure, with a significance level set at $\alpha=0.05$. When the p-value<0.05, the null hypothesis is rejected, indicating that the variable is a significant driver of default risk. Additionally, the odds ratio (OR) is used to intuitively demonstrate the magnitude of the impact of a unit change in a risk factor on the probability of default.

5.2.1 Feature Correlations

Financial metrics exhibit strong cohesion: the quick ratio and interest coverage ratio have a correlation coefficient of 0.82, while the debt-to-asset ratio and debt service coverage ratio show a negative correlation (-0.78). Non-financial metrics and macroeconomic indicators are significantly linked: the industry PPI and sentiment score have a correlation coefficient of 0.67, while supply chain risk and regional GDP growth rate are negatively correlated (-0.59). The moderating effect of enterprise size is pronounced: in small and medium-sized enterprises, the impact of operating profit margin on default ($\beta = 0.41$) is significantly stronger than in large enterprises ($\beta = 0.18$).

5.2.2 Multidimensional Drivers of Default Risk

Logistic regression results (Table 6) show that multiple variables are statistically significant (p-values < 0.05), confirming their strong association with default risk.

Financial Distress (Operating Profit Margin): Coefficient β is negative (-0.21), indicating that higher profit margins reduce default probability. This variable is highly significant at the 1% level ($p < 0.001$). The odds ratio (OR) is 0.81, meaning that for every 1 percentage point increase in operating profit margin, the odds of default decrease by 19% (1 - 0.81).

Sentiment Deterioration (Sentiment Risk Score): Coefficient β is positive (0.16), significant at the 5% level ($p=0.022$). The OR is 1.17, indicating that for every 10-point increase in sentiment risk score, the odds of default increase by 17%.

Macroeconomic Shock (Negative Year-on-Year PPI): This binary variable has a positive coefficient β (0.96), highly significant at the 0.1% level ($p < 0.001$). The OR is 2.60, indicating that enterprises in industries with negative PPI growth have 2.6 times higher odds of default compared to others.

Enterprise Size (SME=1): Coefficient β is positive (0.89), highly significant ($p < 0.001$). The OR is 2.44, confirming that small and medium-sized enterprises have 2.44 times higher odds of default compared to large enterprises.

Interaction Effect (PPI × Size): The interaction term coefficient β is positive (1.32) and significant ($p=0.001$), with an OR of 3.74. This indicates a significant synergistic effect between small and medium-sized enterprises and industry downturns (negative PPI growth), where their combined impact on default probability far exceeds the sum of their individual effects.

Table 6 Logistic Regression Results

Variable	Coefficient β	Standard Error	OR Value	p-value
Operating Profit Margin (%)	-0.21	0.03	0.81	<0.001
Sentiment Risk Score	0.16	0.07	1.17	0.022
Negative Year-on-Year PPI	0.96	0.22	2.60	<0.001
Enterprise Size (SME = 1)	0.89	0.18	2.44	<0.001
Interaction Term = PPI × Size	1.32	0.41	3.74	0.001

5.2.3 Interaction Effect Validation

Through hierarchical regression, it was found that non-financial metrics enhance the explanatory power of financial models: after incorporating sentiment and supply chain features, the AUC increased from 0.74 to 0.83. The interaction term of enterprise size × industry cycle is significant ($\beta=1.32$, $p<0.01$), which indicates that small and medium-sized enterprises exhibit a sharp increase in default probability during industry downturns.

6. Conclusion

6.1 Research Findings

This study, based on public datasets such as CSMAR and Wind, systematically constructed and compared the performance of various machine learning models (XGBoost, LightGBM, Random Forest, Stacking ensemble) with traditional credit scoring models (Logistic Regression, Z-score) in the task of enterprise loan default prediction. Experimental results demonstrate that machine learning models significantly outperform traditional models in key metrics such as AUC and recall. Notably, the XGBoost model achieved an AUC of 0.85, far surpassing Logistic Regression's 0.72. Incorporating non-financial features such as sentiment and industry prosperity improved the model's recall by 15%, validating the effectiveness of the dynamic risk framework. The Stacking ensemble model further enhanced prediction stability and generalization by leveraging the strengths of multiple models, while SHAP analysis ensured model interpretability.

6.2 Theoretical Contributions

Systematic Comparative Study: This study is the first to systematically compare the performance differences between machine learning and traditional models in enterprise loan default prediction, providing empirical evidence for model selection.

Dynamic Risk Assessment Framework: A dynamic risk assessment framework integrating financial and non-financial features is proposed, enhancing the model's adaptability to temporal changes such as industry cycles and sentiment.

Integration of Interpretability and Compliance: By introducing SHAP interpretability techniques, the study maintains high model performance while meeting Basel III's regulatory requirements for model transparency, offering a feasible path for applying "black-box" models in compliant risk control scenarios.

6.3 Managerial Implications

Firstly, domestic banks should prioritize ensemble models for enterprise loan default prediction. Commercial banks should replace traditional scorecard systems with XGBoost or Stacking ensemble models, focusing on optimizing business-oriented metrics such as Recall@Top 10%. During system deployment, establish a dynamic feature engineering framework to incorporate non-financial data such as sentiment, supply

chain, and industry prosperity in real time, updating enterprise risk scores quarterly. Financial institutions should establish a lifecycle management system for machine learning models, including quarterly backtesting (AUC decay threshold <3%), feature stability monitoring (PSI < 0.1), and version iteration mechanisms. For SME loan scenarios, develop dedicated lightweight models (e.g., pruned LightGBM) to reduce computational resource demands.

Secondly, financial institutions should build dynamic risk monitoring systems and implement differentiated credit strategies. For SMEs with high debt-to-asset ratios (>70%) and industry prosperity indices <50, initiate "red list" oversight. Implement tiered interest rates based on predicted risk levels for SME risk monitoring and evaluation.

Thirdly, regulatory authorities should refine fintech regulatory frameworks to promote data ecosystem development. The CBIRC should formulate Guidelines for Risk Management of Machine Learning Models in Commercial Banks, specifying interpretability requirements: all black-box models must be equipped with explanation tools such as SHAP or LIME. Stability standards: feature PSI (Population Stability Index) <0.15, annual AUC decay <5%. Audit traceability: retain model versions and prediction records for three years. The People's Bank of China should lead the establishment of an Enterprise Risk Information Sharing Platform, integrating: 1) Financial Data: Cross-departmental information such as tax, customs, and social security data; and 2) Non-Financial Data: Sentiment monitoring, supply chain networks, and industry prosperity indices. An SME-Specific Database should be established to address the issue of fragmented information. In free trade zones, conduct "regulatory sandbox" pilots to allow commercial banks to test machine learning models in scenarios such as supply chain finance and credit loans for tech enterprises, exploring blockchain technology for secure risk data sharing and offering pilot institutions risk-weighted asset relief (up to 20% reduction).

6.4 Research Limitations and Future Directions

This study has certain limitations, and future research can explore the following directions:

Data Scope Limitation: The sample primarily focuses on listed companies with relatively standardized governance structures, and the model's generalization to the larger population of non-listed SMEs requires further validation. Future research can expand data sources to include a broader range of enterprise samples and explore integrating "alternative data" such as utility consumption and customs import/export data to build more comprehensive enterprise credit profiles.

Balancing Model Real-Time Performance and Complexity: While ensemble models like Stacking improve prediction accuracy, they also increase computational complexity and deployment costs. Future research can explore model compression (e.g., knowledge distillation) and lightweight techniques to maintain performance while meeting the low-latency and high-throughput requirements of online real-time risk control systems.

Utilization of Deep and Unstructured Information: The current use of text-based information, such as sentiment, remains relatively shallow. Future work can incorporate advanced natural language processing (NLP) techniques for event extraction and sentiment analysis, and leverage graph neural networks (GNNs) to model complex relationship networks, such as guarantees and supply chains, to reveal risk transmission mechanisms at a deeper level.

Cross-Cycle Robustness Validation: The training data's time window is limited and does not cover complete extreme economic cycles. Future studies can employ back-testing and stress testing to evaluate the model's robustness and adaptability under different macroeconomic scenarios.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance,* vol. 23**,** no. 4, pp. 589-609.

Beck, T., Demirguc-Kunt, A. and Levine, R. (2005). SMEs, growth, and poverty: Cross-country evidence. *Journal of economic growth,* vol. 10**,** no. 3, pp. 199-229.

Bis (2010). *Basel III: International framework for liquidity risk measurement, standards and monitoring,* Basel, Switzerland: Bank for International Settlements.

Breiman, L. (2001). Random forests. *Machine learning,* vol. 45**,** no. 1, pp. 5-32.

Campello, M., Graham, J. R. and Harvey, C. R. (2010). The real effects of financial constraints: Evidence from a financial crisis. *Journal of financial Economics,* vol. 97**,** no. 3, pp. 470-487.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research,* vol. 16, pp. 321-357.

Delong, E. R., Delong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics,* vol. 44**,** no. 3, pp. 837-845.

Diamond, D. W. and Rajan, R. G. (2001). Liquidity risk, liquidity creation, and financial fragility: A theory of banking. *Journal of political Economy,* vol. 109**,** no. 2, pp. 287-327.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics,* vol. 29**,** no. 5, pp. 1189-1232.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017). Published. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems (NeurIPS 2017), 2017 Long Beach, CA. Curran Associates, Inc., pp. 3149-3157.

Lessmann, S., Baesens, B., Seow, H.-V. and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research,* vol. 247**,** no. 1, pp. 124-136.

Loughran, T. and Mcdonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of accounting research,* vol. 54**,** no. 4, pp. 1187-1230.

Lundberg, S. M. and Lee, S.-I. (2017). Published. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS 2017), December 4–9 2017 Long Beach, CA. Curran Associates, Inc., pp. 4765-4774.

Moscatelli, M., Parlapiano, F., Narizzano, S. and Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications,* vol. 161, p. 113567.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). Published. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), August 13–17 2016 San Francisco, CA. Association for Computing Machinery (ACM), pp. 1135-1144.

Zhou, Z. (2016). *Machine Learning,* Beijing: Tsinghua University Press.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.