Loan Default Prediction Based on Machine Learning Approaches

Xinyu Cai
College of Business
Jiaxing University
Jiaxing, Zhejiang, China
caixinyu@zjxu.edu.cn

Wenbo Dai*
College of Business
Jiaxing University
Jiaxing, Zhejiang, China
cxy96201@163.com

Jingyu Lu
College of Business
Jiaxing University
Jiaxing, Zhejiang, China
Lucy0420Date@outlook.com

Abstract

To address the credit risk losses incurred by commercial banks due to loan defaults, this study utilizes the loan default prediction dataset from the Alibaba Tianchi platform to develop machine learning models for predicting customer defaults, aiming to mitigate credit risk. Given the characteristics of class imbalance and high dimensionality of loan data, data preprocessing and exploratory data analysis are conducted. Based on a comparative analysis of various models, seven machine learning algorithms that demonstrate superior performance are selected for experimental comparison, including Decision Tree, Random Forest, AdaBoost, Bagging, XGBoost, LightGBM, and CatBoost. The results indicate that ensemble learning algorithms exhibit higher accuracy and predictive performance compared to single algorithms, with the CatBoost model performing best across various indicators, including AUC. The study identifies key features highly correlated with loan defaults, including loan grade, annual income, loan amount, credit history length, and debt-to-income ratio.

CCS Concepts

• **Computing methodologies** → Machine learning; Machine learning approaches.

Keywords

Machine Learning, Cross-Validation, Ensemble Learning, Loan Default

ACM Reference Format:
Xinyu Cai, Wenbo Dai, and Jingyu Lu. 2025. Loan Default Prediction Based on Machine Learning Approaches. In *2025 2nd International Conference on Generative Artificial Intelligence and Information Security (GAIIS 2025), February 21–23, 2025, Hangzhou, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3728725.3728813

1 Introduction

Over the past several years, the financial market in China has witnessed astonishing growth. This surge can be attributed largely to the consistent and vigorous enforcement of inclusive finance policies and the resultant remarkable expansion of personal loan services. The People's Bank of China reports that the balance of personal loans had swelled to 69.8 trillion yuan by the end of 2023, reflecting a year-on-year increase of 7.5 percent [1]. But although this sector is expanding very quickly, we all now must question its underlying health and sustainability. Is it actually servicing a real demand for credit from Chinese residents? Or is it a credit bubble waiting to burst? On the surface, one might be tempted to respond that it is likely to be a bubble because the ratio of non-performing loans at commercial banks is also trending upward [2].

The ordinary ways of appraising the risk of defaulting on a loan—that is, determining how likely a borrower is to repay or not—rely on expert judgment and basic statistical tools such as credit scorecards [3]. These tools are simple in comparison to the kinds of unfathomable, complex, and large datasets that even an average lender has to manage these days. Some experts maintain that the simple scoring systems of yesteryear might have worked better at capturing the essential, if messy, borrower-to-lender relationships that now just seem to elude even the smartest models. Big data technology and artificial intelligence have brought about a new frontier in financial risk management—machine learning. What sets machine learning apart from previous solutions is its ability to plow through huge piles of data and automatically find the patterns that lie hidden within [4].

Using machine learning to predict when a borrower will default on a loan has attracted a lot of attention around the world. This attention started with Khandani et al. (2010), who showed that not only were machine learning methods better, but they also had the potential to be much better than traditional methods at predicting when consumers would default on their loans [5]. Khandani then led a follow-up study with Lessmann et al. (2015) that looked at 41 kinds of machine learning methods (or algorithms) and traditional methods to see how well they predicted default rates [6]. The attention has recently turned to ensemble learning methods, which are a kind of machine learning method that works much like an orchestra works [7].

In credit evaluation, several machine learning methods have been explored in domestic research. For personal, small loan credit risk evaluation, Luo and Chen (2017) used logistic regression to analyze the risk, identifying key feature variables in the process [8]. According to Li et al. (2018), support vector machines offer distinct advantages when dealing with small, imbalanced datasets [9]. More recently, ensemble learning approaches have significantly influenced China's academic research. For instance, Zhou et al. (2020) achieved improved predictive results by optimizing random forest algorithms, while Shen and Zhang (2020) developed a hybrid SVM-KNN model for banking credit assessment systems [10] [11]. In spite of these advancements in research, some challenges persist. A considerable amount of research has been focused on individual, narrowly selected machine learning models, which has led to a

*corresponding author.

GAIIS 2025, Hangzhou, China

ACM ISBN 979-8-4007-1345-3/2025/02
https://doi.org/10.1145/3728725.3728813

significant gap in research that comprehensively and systematically compares a diverse range of established methods. Even more glaring is the lack of research that discusses the detailed, real-world application and interpretation of these models in a business context. Furthermore, there is minimal research that concentrates on the unique aspects of the financial ecosystems in China. To address these problems, this study not only seeks to fill the gaps in the existing research but also aims to serve the financial practitioners in China by providing a detailed risk management framework in which to compare various machine learning and statistical models used as loan default prediction tools.This research contributes several important components:

(1) A framework for comparison is developed in this research for prediction of defaults on loans. This framework incorporates seven machine learning models that are widely recognized and used—decision trees, random forests, AdaBoost, Bagging, XGBoost, LightGBM, and CatBoost. With the help of this comparison framework, a very systematic assessment can be made of the performance of the different prediction models.

(2) This analysis uses a large, real-world loan dataset and five-fold cross-validation to ensure robust and consistent evaluations. Owing to the dataset's inherent imbalance, the study relies on AUC as the primary metric to maintain high standards of reliability.

(3) This study distinguishes the main causes of consumer credit defaults by analyzing feature importance across models. It provides actionable intelligence to help lenders better their credit underwriting policies.

(4) It also gives a deep exploration into the strengths and weaknesses of the leading models, especially CatBoost, in real-life business conditions, providing clear guidance on how to use these models for financial risk management.

These findings enhance loan default risk management in commercial banking and offer new perspectives on integrating financial technology with credit risk assessment. Furthermore, the approaches and results presented here hold considerable value for tackling broader challenges in financial risk prediction.

2 data and methods

2.1 Data Source and Preprocessing

This study uses the loan default prediction dataset from the Alibaba Tianchi big data platform's financial risk control competition [12]. The dataset includes over 1.2 million real loan records from a credit platform, with 47 distinct feature variables. To ensure both model reliability and computational efficiency, 800,000 records were randomly selected for training, while the remaining 200,000 records were reserved for testing.

The dataset includes a range of feature categories: basic borrower attributes (e.g., age, gender), credit history metrics (e.g., FICO scores, delinquency records), loan-specific details (e.g., principal amount, interest rates), and a set of anonymized variables. The variables are shown in Table 1. The target variable, "isDefault," is a binary classification label, where 1 denotes a loan default and 0 signifies repayment as per the contractual terms.

The data preprocessing phase involved several systematic steps, outlined as follows:

(1) Handling Missing Data. Different methods were employed to address missing values for continuous and discrete variables. For continuous variables, the median was used for imputation, while the mode was applied for discrete variables. This approach helps maintain the natural distribution of the data and reduces the risk of bias compared to simpler imputation methods.

(2) Managing Outliers. Outliers were identified using the 3σ principle, and extreme values were capped rather than removed through Winsorization. This approach prevents unusually high or low values from distorting the analysis while preserving potentially valuable data points.

(3) Converting Data Formats. Date features, such as "issueDate," were transformed into timestamps, improving computational efficiency and enabling more effective analysis of time-based patterns.

(4) Encoding Categorical Variables. One-hot encoding was used to transform categorical variables, such as "grade" and "employmentTitle," into a format suitable for mathematical analysis, without assuming any inherent order among the categories.

2.2 Feature Engineering

Feature engineering is critical for enhancing model performance, and this study employed several strategies to optimize the features: (1) Feature Selection. Feature selection was conducted using correlation analysis and Variance Inflation Factor (VIF) methods. To prevent instability in the models, variables with high correlation (correlation coefficient > 0.9) or significant multicollinearity (VIF > 10) were removed. (2) Feature Construction. Leveraging domain knowledge, new variables were created, including the debt-to-income ratio (DTI) and loan-to-income ratio. These constructed features capture additional credit risk factors that may not be evident in the original data. (3) Feature Standardization. Continuous variables were standardized using Z-scores to make them comparable across different scales. Additionally, monetary and income-related variables were log-transformed to mitigate distortions caused by heteroscedasticity during model fitting. (4) Class Balancing. Given the typical class imbalance in loan default prediction, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm was applied to augment the minority class samples, creating a more balanced distribution between positive and negative cases.

Figure 1 illustrates the correlation patterns between features. The color intensity in the figure corresponds to the strength of the correlation between pairs of features, with darker shades indicating stronger relationships.

2.3 Model Selection and Construction

This study evaluates the performance of seven widely used machine learning models:

(1) Decision Tree. The Decision Tree model employs the CART (Classification and Regression Tree) method, using the Gini index to decide how to split the data. It builds a tree-like structure by recursively selecting the best feature to partition the data, starting with a root node and generating terminal leaf nodes. Various versions of decision trees, such as ID3, C4.5, and CART, use different criteria for splitting: information gain, gain ratio, and Gini index,

Table 1: Dataset Features and Descriptions.

Feature Name	Feature Description
id	Unique identifier assigned to the loan application
regionCode	Regional code
isDefault	Whether the loan defaulted or not
loanAmnt	Loan amount
term	Loan term (in years)
interestRate	Loan interest rate
installment	Installment amount
dti	Debt-to-Income ratio
homeOwnership	Home ownership status provided by the borrower at registration
issueDate	Month the loan was issued
purpose	Loan purpose category as stated by the borrower in the loan application
postCode	First 3 digits of the postal code provided by the borrower in the loan application
delinquency_2years	Number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
ficoRangeLow	Lower boundary of the range the borrower's FICO belongs to at loan issuance
ficoRangeHigh	Upper boundary of the range the borrower's FICO belongs to at loan issuance
openAcc	Number of open credit lines in the borrower's credit file
pubRec	Number of derogatory public records
pubRecBankruptcies	Number of public record bankruptcies
initialListStatus	Initial listing status of the loan
revolBal	Total credit revolving balance
revolUtil	Revolving line utilization rate
totalAcc	Total number of credit lines currently in the borrower's credit file
applicationType	Indicates whether the loan is an individual application or a joint application with two co-borrowers
earliesCreditLine	Month the borrower's earliest reported credit line was opened
employmentTitle	Job title
employmentLength	Employment length (in years)
annualIncome	Annual income
n series anonymous features	Anonymous features n0-n14, representing processed behavioral count features of borrowers

respectively. Pruning techniques are typically applied to address the issue of overfitting.

(2) Bagging. Bagging (Bootstrap Aggregating) is an ensemble technique that reduces generalization error by combining predictions from multiple models. It works by creating bootstrap samples from the original data and training separate base learners (usually decision trees) on each subset. The final prediction is made by aggregating the individual model predictions—voting for classification tasks and averaging for regression tasks.

(3) Random Forest. Random Forest is an ensemble of 100 decision trees. It improves upon Bagging by incorporating both bootstrap sampling and random feature selection. During the construction of each tree, a random subset of features is selected for splitting nodes, enhancing model diversity and improving generalization. By aggregating predictions across all trees, Random Forest helps reduce overfitting.

(4) AdaBoost. AdaBoost (Adaptive Boosting) enhances the performance of weak classifiers, typically decision trees, by iteratively adjusting the weights of misclassified samples. Each iteration increases the weight of misclassified samples and decreases the weight of correctly classified ones. The final classifier is a weighted combination of all weak classifiers, with their influence determined by their individual performance.

(5) XGBoost. XGBoost uses a second-order Taylor expansion to approximate the loss function and applies regularization to control model complexity. It leverages a pre-sorted decision tree algorithm to determine the best split points. Although this can be computationally intensive, XGBoost overcomes these challenges through parallel computing, cache optimization, and out-of-core techniques.

(6) LightGBM. LightGBM employs a histogram-based decision tree algorithm with a leaf-wise growth strategy, distinguishing itself through fast training speed and memory efficiency. It optimizes the histogram algorithm and adopts leaf-wise growth to reduce computational complexity while maintaining high prediction accuracy.

(7) CatBoost. The symmetric tree structure within CatBoost effectively processes features with the Ordered Boosting strategy by calculating occurrence frequencies and introducing hyperparameters to develop numerical representations since due to its ordered boosting method the gradient estimate bias reduction enables enhanced accuracy and generalization results.

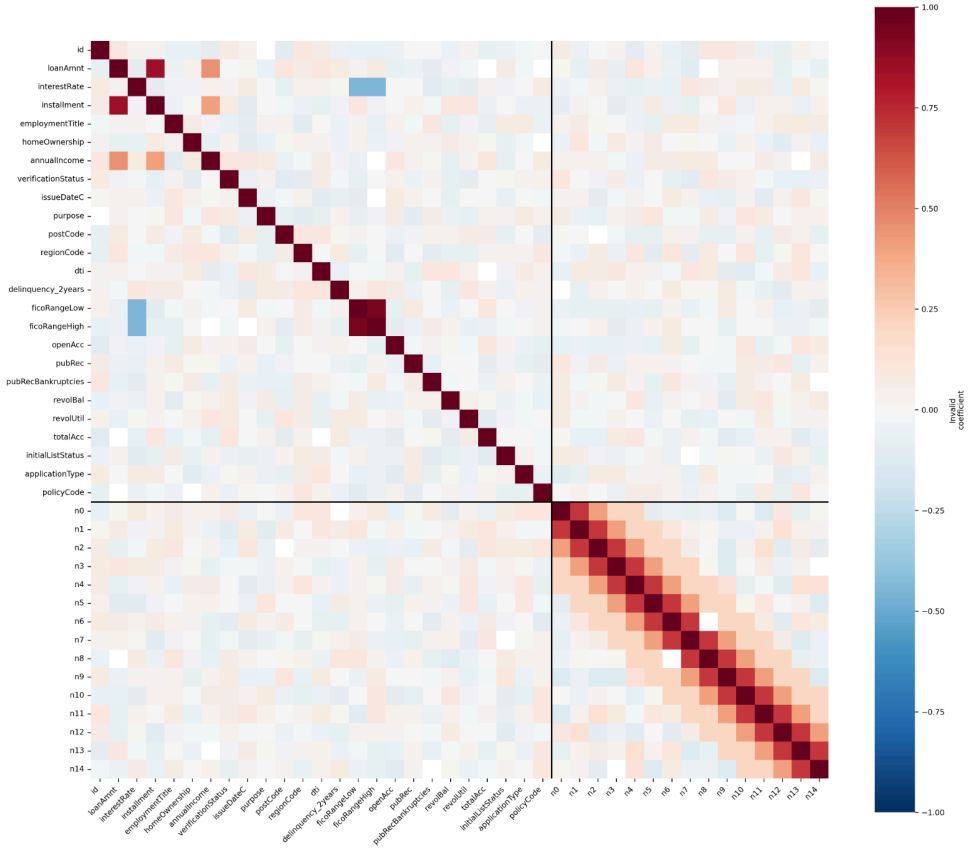


Figure 1: Correlation Heatmap of Features.

2.4 Model Evaluation Metrics

Performance metrics serve a crucial role in robust model evaluation for loan default prediction because they help understand the meaning and significance of in situations where class imbalance exists. A Threshold-Independent and Skewed-Distribution Resistant Measure of Discrimination Power exists in AUC which Enables Evaluating a Model for its Ability to Separate Default from Other Cases at All Classification Thresholds. Accuracy as a performance measure functions as a general indicator but delivers deceptive results for unbalanced data since it mostly identifies the dominant class. On the other hand, Precision is the accuracy of positive predictions (fraction of predicted defaults that are true defaults), while Recall is the model's ability to detect all actual defaults, both of which are important in managing financial risk where identifying defaults is crucial. The F1 Score serves as an optimal metric to balance false positive errors against missed defaults because it calculates the harmonic mean between Precision and Recall. The Kolmogorov-Smirnov (KS) Statistic functions as a vital indicator of discriminative power in risk assessment by measuring the separation between default and non default cumulative distribution functions which provide an integrated and sophisticated perspective on predictive performance which aligns with academic standards of discourse.

3 experimental results and analysis

3.1 Model Performance Comparison

The research examined seven machine learning models to predict loan defaults by implementing 5-fold cross validation and determining average performance metrics for all models which are displayed in Table 2 where all evaluation metrics show CatBoost to deliver leading performance since it achieved 0.7488 AUC. An approximately 6% increase emerges from the baseline decision tree model in these results. The research findings match those of Zhang et al. (2021) who documented CatBoost's effectiveness in credit risk assessment [13]. The performance of XGBoost and LightGBM models matched each other but CatBoost demonstrated superior results. Traditional ensemble methods, including Random Forest and AdaBoost, surpassed the single decision tree but fell short of the gradient boosting models in predictive power.

It is worth highlighting the computational efficiency of these models. Although CatBoost delivered the highest predictive performance, it demanded the longest training time. In contrast, LightGBM balanced strong predictive accuracy with notably faster training times, a feature that could prove valuable in business contexts requiring frequent model updates. For a visual comparison of

Table 2: Performance Comparison of Models.

Model	AUC	Accuracy	Precision	Recall	F1 Score	KS Value	Training Time (s)	Prediction Time (s)
Decision Tree	0.7066	0.8014	0.7576	0.7242	0.7406	0.3124	12.5	0.3
Random Forest	0.7091	0.803	0.7689	0.7301	0.7489	0.3182	45.7	1.2
AdaBoost	0.7159	0.8031	0.7701	0.7356	0.7524	0.3318	78.3	0.8
Bagging	0.7146	0.8024	0.7695	0.7342	0.7514	0.3292	62.1	1.5
XGBoost	0.7341	0.8153	0.7912	0.7589	0.7746	0.3682	156.4	0.6
LightGBM	0.7333	0.8147	0.7905	0.7581	0.7739	0.3666	34.2	0.4
CatBoost	0.7488	0.8226	0.8037	0.7741	0.7886	0.3976	203.7	0.7

the models' performance, Receiver Operating Characteristic (ROC) curves are provided in Figure 2.

3.2 Feature Importance Analysis

Figure 3 highlights the key predictors of loan defaults, with loan grade, annual income, loan amount, credit history length, and debt-to-income ratio standing out as the top five factors. Notably, behavioral factors such as the number of recent credit inquiries (inq_last_6mths) and credit card utilization (revol_util) also play a crucial role, ranking high among the most influential predictors. These findings highlight the importance of recent credit behavior of borrowers in predicting default risk, which is not captured by traditional credit scoring models which implies that current risk assessment approaches should be updated by combining dynamic behavioral data with standard financial metrics.

3.3 Model Interpretability Discussion

Ensemble learning models such as CatBoost have strong predictive power, but their 'black box' nature makes them difficult to use in finance. To address this lack of interpretability, this study used SHAP (SHapley Additive exPlanations) values, a method that explains how individual features contribute to specific predictions. The SHAP value analysis for a typical default case is shown in Figure 4.

Figure 4 shows that in this case, a high default risk is driven by a large loan amount, low annual income, and short credit history. Such detailed, case by case breakdowns not only help explain the model's decision making, but also offer practical guidance to credit evaluators. CatBoost's strength in handling categorical variables (occupation types, loan purposes) without complex feature engineering was further analyzed which is especially useful for financial datasets, which tend to have a lot of categorical elements.

Overall, gradient boosting based ensemble methods, especially CatBoost, are good at predicting loan defaults since in accuracy and feature importance, they outperform traditional approaches. Still, practical use demands a careful balance of accuracy, computational speed, and interpretability.

4 conclusion and discussion

In this study, we examined and compared seven machine learning algorithms for predicting loan defaults using a comprehensive loan dataset where several key insights surfaced through meticulous experimentation.

The CatBoost algorithm stood out first, achieving an AUC of 0.7488, which is about 6 percentage points higher than the baseline decision tree model. All three were far ahead of traditional machine learning approaches, with XGBoost and LightGBM following closely behind which highlight the pronounced strengths of gradient boosting based ensemble methods in tackling complex financial risk prediction challenges.

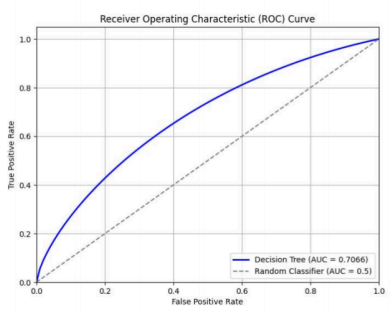
Second, feature importance analysis within the CatBoost model showed that loan grade, annual income, loan amount, credit history length, and debt to income ratio were the top five drivers of loan defaults which not only confirm the continued importance of core metrics in traditional credit evaluations, but also highlight the critical role of behavioral factors (e.g., recent credit inquiries and credit card utilization) in predicting default risk.

Third, while ensemble models are often referred to as opaque 'black boxes', SHAP (SHapley Additive exPlanations) values greatly helped to explain the workings of the CatBoost model which provides a more complete picture of the model's broader decision making framework and provides precise risk assessments for individual cases, providing tangible support for credit decision processes.

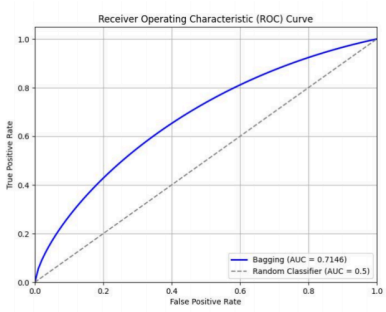
Finally, on the efficiency front, CatBoost had the best predictive power but took the longest training time while on the other hand, LightGBM offered strong accuracy with fast training, making it a compelling choice for financial institutions that need to update their models quickly.

The implementation of CatBoost displayed effective handling of categorical inputs through its ability to perform without requiring complex preprocesses which becomes crucial in financial datasets containing many categorical features since it drives analytical efficiency.

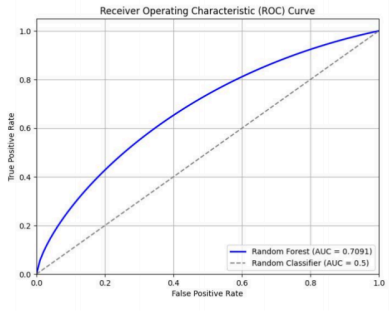
These recent advancements in technology have certain restrictions since the sizable data originates from a single credit platform resulting in constrained broader credit market representation. Research enhancement could be achieved by incorporating data from multiple platforms together with information from social media or consumer behavior patterns. The research design employed a static dataset which failed to account for the effects that changing economic conditions might have on default risks. The integration of Time series models specifically Long Short-Term Memory networks would capture temporal shifts to enhance predictive capabilities and, as feature engineering reached some milestones, new potential features should be investigated for their potential impact.



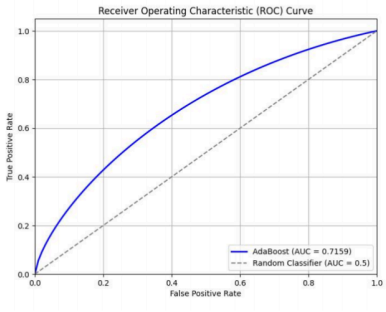
a: AUC Performance of Decision Tree Model



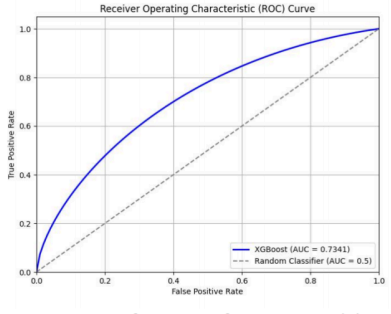
b: AUC Performance of Bagging Model



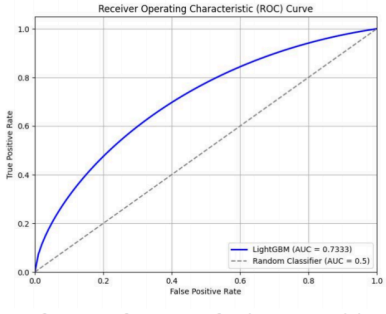
C: AUC Performance of Random Forest Model



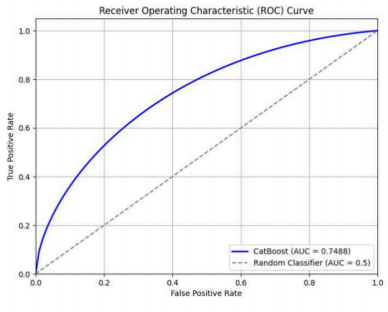
d: AUC Performance of AdaBoost Model



e: AUC Performance of XGBoost Model



f: AUC Performance of LightGBM Model



g: AUC Performance of CatBoost Model

Figure 2: ROC Curves Comparison of Models.

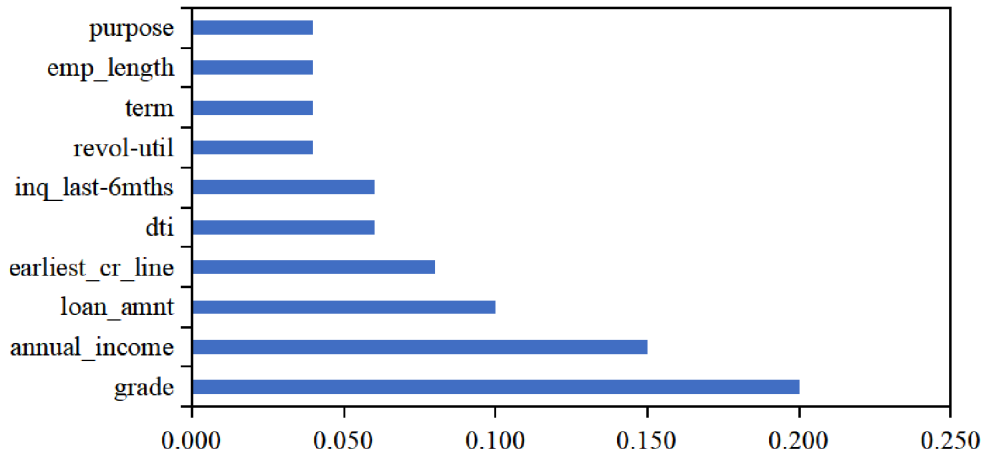


Figure 3: Top 10 Important Features of CatBoost Model.

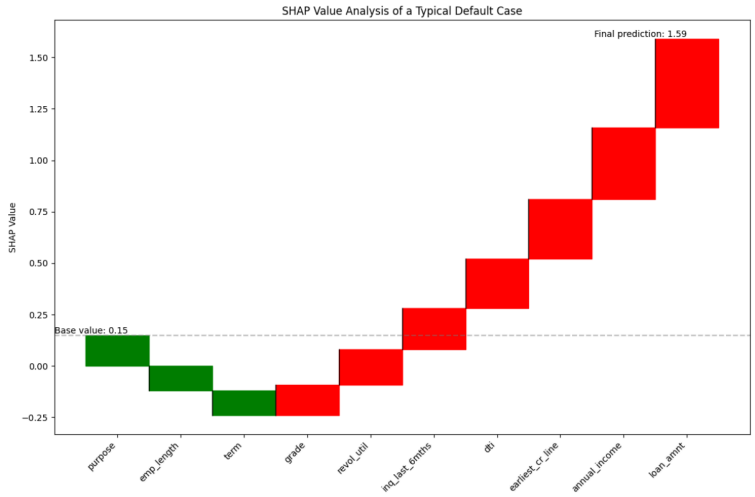


Figure 4: SHAP Value Analysis of a Typical Default Case.

The research findings deliver essential knowledge about enhancing financial institution risk management through advanced predictive modeling specifically for commercial banks. Gradient boosting ensemble methods including CatBoost, XGBoost and LightGBM demonstrate superior performance in loan default prediction by enhancing accuracy and managing categorical financial variables since recent credit inquiries and credit card utilization serve as behavioral indicators that help assess borrowers' current financial situation and credit health which significantly enhances predictive accuracy. SHAP (SHapley Additive exPlanations) values solve interpretability issues of complex 'black box' models by delivering transparent credit decision insights that fulfill regulatory standards. The study also shows a tradeoff between predictive power and computational efficiency: CatBoost is very accurate, but its longer training time may make it less practical in environments that require fast model updates, while LightGBM's faster processing is a viable alternative however, there are limitations, including the reliance on data from a single credit platform, which may limit generalizability, and the use of a static dataset that does not consider the impact of economic variability on default risk. Time series analysis along with other data sources needs to be used to perform dynamic tracking of changes in risk profiles over time in future research.

Future research should investigate multiple avenues to enhance loan default prediction models based on the findings from this study. An important avenue is to combine data from various credit platforms, possibly supplemented with other sources such as social media analytics and consumer behavior patterns, to build more

comprehensive and representative risk prediction models. Investigations may also consider hybrid approaches that combine time series techniques, such as Long Short Term Memory networks, with established predictive frameworks to better track the changing creditworthiness of borrowers over time. A detailed analysis of model performance across different demographic groups and under irregular market conditions is essential for ensuring the robustness and fairness of predictive outcomes. Furthermore, implementing mechanisms to dynamically update models would enable them to quickly adapt to shifting market dynamics, enhancing their real-world applicability. These efforts aim to refine more accurate, reliable, and widely applicable methods for predicting loan defaults, ultimately improving decision support for financial institutions' risk management and fostering innovation in credit risk assessment.

Acknowledgments

The Ministry of Education Industry-University Cooperative Education Project "Talent Cultivation in Business Administration Based on Innovative Collaborative Training Mode" (231104213285004);

Zhejiang Province Higher Education Society General Project "Research on Digital Reform Strategies of Higher Education Teaching in Zhejiang Province from the Perspective of New Quality Talent Cultivation" (KT2024355);

Jiaxing University Scientific Research Startup Project (CD71524005)

References

[1] People's Bank of China. Financial statistics report for 2023. 2024.
[2] China Banking and Insurance Regulatory Commission. Key regulatory indicators of commercial banks in 2023. 2024.
[3] Fang, K., Wu, J. Personal housing loan default prediction and interest rate policy simulation. Statistical Research, 2013, 30, 75-82.
[4] Zhang, J., Li, W., Ruan, S. Loan default risk prediction based on machine learning. Journal of Changchun University of Science and Technology (Social Sciences Edition), 2021, 34, 88-93.
[5] Khandani, A.E., Kim, A.J., Lo, A.W. Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 2010, 34, 2767-2787. [https://doi.org/10.1016/j.jbankfin.2010.02.017]
[6] Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 2015, 247, 124-136. [https://doi.org/10.1016/j.ejor.2015.05.030]
[7] Xia, Y., Liu, C., Li, Y., Liu, N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 2017, 78, 225-241. [https://doi.org/10.1016/j.eswa.2017.02.017]
[8] Luo, F., Chen, X. Credit risk assessment and application of personal small loans based on Logistic regression model. Financial Theory and Practice, 2017, 38, 30-35.
[9] Li, J., Ma, S., Jin, M., Chu, C. A crowdfunding default risk early warning model based on SA-SVM. Statistics & Information Forum, 2018, 33, 70-77.
[10] Zhou, Y., Cui, J., Zhou, L., Sun, H., Liu, S. Research on personal credit risk assessment based on improved random forest model. Credit Reference, 2020, 38, 28-32.
[11] Shen, Q., Zhang, L. A new method for bank credit risk identification: SVM-KNN combination model. Financial Regulation Research, 2020, 7, 23-37.
[12] Alibaba Tianchi. Financial risk control - loan default prediction challenge dataset. 2023. Available online: [https://tianchi.aliyun.com/competition/entrance/531830/information]
[13] Zhang, J., Li, W., Ruan, S. Loan default risk prediction based on machine learning. Journal of Changchun University of Science and Technology (Social Sciences Edition), 2021, 34, 88-93.