International Journal of Innovation and Research in Educational Sciences
Volume 10, Issue 3, ISSN (Online) : 2349–5219

Loan Default Prediction Based on Ensemble Learning

Yifan Huang[1*], Yanxi Shao[1], Dapeng Tang[2], Jie Huang[3] and Sijia Chen[1]
[1]School of Finance, Zhejiang University of Finance and Economics, Hangzhou, China.
[2]School of Data Science, Zhejiang University of Finance and Economics, Hangzhou, China.
[3]School of Economics, Zhejiang University of Finance and Economics, Hangzhou, China.
*Corresponding author email id: 17816740085@163.com

Date of publication (dd/mm/yyyy): 28/05/2023

***Abstract* – In recent years, with the booming development of modern technology and information technology, intelligent risk control has become an indispensable part of the healthy development of the financial industry. The core issue in the field of intelligent risk management is to accurately identify potential risks in loans, not to issue loans to borrowers with a high default rate, or to track users who have issued loans in real time, so as to more effectively ensure the interests of lending institutions. Therefore, it is a very important research topic to use the massive data information of lenders and data mining technology to predict the default behavior of loan users. According to the characteristics of unbalanced loan data categories and high feature dimensions, we clean the data and select the features with strong predictive ability by filtered feature selection. Subsequently, based on the comparison of various models, this paper selects the best performing Adaboost model for loan default prediction model construction and conducts model evaluation. The analysis found that all the indicators of Adaboost are high, indicating that the model has better performance and can be used to accurately predict loan defaults. This is used as a basis to provide reference for commercial banks and other lending platforms when offering credit products to borrowers.**

***Keywords* – Loan Default Prediction, Filtered Feature Selection, Model Comparison, Adaboost.**

I. INTRODUCTION

With the rapid development of the Internet in China, Internet finance has also started to develop rapidly, and new credit transaction methods such as Jingdong, and crowdfunding have emerged in the consumer credit market in China. In addition, along with the emergence of inclusive finance and Internet consumer finance, a variety of credit products have been introduced to the public, and the P2P lending market has been developing rapidly. While these new businesses bring convenience to people's lives, they also lead to many credit problems, such as loan defaults, platform fraud and information asymmetry, which makes financial risk control and research malefactor a necessity. Compared with traditional finance, Internet finance faces bottlenecks such as inadequate laws, privacy protection, lagging regulation and high risk. Therefore, strengthening credit risk control, building a complete personal credit system, and helping the orderly development of the online credit market have far-reaching effects on the development of China's Internet finance industry.

Due to the development of the market economy and the expansion of the consumer credit market, the credit business for housing, automobiles, consumer durables and educational assistance has developed rapidly, but as the scale of consumer credit continues to expand, the problems and risks in the credit business have become increasingly evident, and various credit defaults have aroused the concern of all walks of life, especially in the financial field. At present, the four state-owned commercial banks are still the main source of funds for personal loans in China, which are state-owned assets. The more lenders default, the greater the risk of bank lending, the more serious the loss of state-owned assets, thus affecting the state budget and the effectiveness of the actual assets can play. It will affect the state's social and economic management functions, and in serious cases, it will

affect workers' income, unemployment, and even cause social unrest. It can be seen that serious personal loan defaults will have a huge harmful effect on society, which is a problem that all of us can not ignore.

Essentially, commercial banks are financial institutions, which mainly operate by means of risk to make profits, so controlling the occurrence of credit risk is fundamental to their development and key to the success or failure of business management. For the traditional financial risk control, risk identification ability is relatively limited, mainly by virtue of expert experience and comprehensive human costs. Nowadays, data mining technology can deeply analyze the huge amount of risk data, so as to more accurately dig out the risk information implied by the data, effectively solving the problem that the traditional model cannot handle the huge amount of data and the deep hidden level of information.

II. RELATED WORK

A. *Study of the Factors Influencing Default*

Back in 2008, Miao used a sample of 1690 students who entered repayment period to reveal statistical associations between students' repayment status (compliance or default) and their influencing variables through factor analysis, discriminant analysis, cluster analysis, and logistic model analysis to infer the causes of student default occurrence. It was found that the most important factors influencing student default were the total amount of student loans and the amount of repayment per unit of time [1].

In 2011, Chen adopted a combination of theoretical and empirical research, quantitative and qualitative methods, through the form of internal data collection to analyze the personal status of credit card default cardholders in a branch, and used SPSS15.0 software to analyze the data and explore which factors among the personal characteristics of cardholders have a significant impact on the data were analyzed using SPSS15.0 software to explore which of the cardholders' personal characteristics have a significant impact on "credit card default" and the weighting of the respective impacts to provide a basis for the final targeted recommendations [2].

In 2013, Zhang used the logistic method, which is a common method for analyzing loan defaults at home and abroad, and used a stepwise regression to eliminate the insignificant factors in the regression until all the factors were significant. Through parameter estimation, we found that four factors, namely, borrower's income level, education level, industry of borrower's occupation, and borrowing amount, have significant effects on default [3].

In 2015, Deng developed a logistic model to analyze whether the influencing factors of farmers' loan default are significant. The influencing factors were subdivided into 16 influencing factors based on five major aspects, which were age, gender, etc. of the sample farmers. The model showed that age, household size, loan amount, whether they belonged to the credit village, and whether they had overdue records had significant effects [4].

In 2021, based on the data of defaulted credit bonds in China from 2014 to 2020, Tan searches for the influencing factors at four levels: macro factors, industry factors, corporate factors and bond factors, taking into account the development of the bond market and the availability of data. The recovery rate model was established by multiple linear regression, and the analysis found that among the macro factors, the 10-year bond yields 1-3 years before default showed a significant positive correlation with the recovery rate [5]. The empirical

analysis was conducted based on the credit data of micro and small enterprises in a commercial bank in Shaoxing at the end of 3 quarters of 2020. After that, a binary logistic model was used to compare the significance of each indicator. The empirical analysis showed that the length of cooperation between the bank and the enterprise had a significant positive influence on credit default of micro and small enterprises, the loan term and the guarantee method had a significant negative influence on whether to default [6].

In 2022, Yang used Q Bank's online loans for micro and small enterprises as a sample, and divided the overall sample into 12 independent variables based on four dimensions. The important factors influencing whether the micro and small enterprises defaulted on their online loans were analyzed by the logistic regression model. The study showed that loan maturity, annual interest rate, credit use, and credit risk control of the managing institution were significantly associated with loan default [7].

B. *The Study of Default Prediction Model*

Back in 2013, Kong selected 1463 listed companies in Shanghai and Shenzhen A-shares and applied the Merton model based on market information and the Logistic model based on accounting information to measure the default risk of the companies respectively. The results of correlation analysis showed that the two models were less consistent in measuring default, and further analysis based on ROC curves and accuracy ratios showed that the logistic model was significantly better than the Merton model in predicting default [8].

In 2014, Hong used log-log regression method, logistic regression method, and classification and regression tree method to construct prediction models for home mortgage loan default loss rate, and compared the prediction effects of these three types of models and historical data averaging method models for in-sample and out-of-sample data. The results showed that logistic regression method has the best prediction effect [9].

In 2022, Guo used corporate bonds maturing in 2017-2020 as the research object, selected macroeconomic indicators, bond issuance factor indicators and issuing corporate bond subject-related indicators in the year of bond maturity, and applied Logistic regression, multilayer perceptron, radial basis function, discriminant analysis, support vector machine, and other seven machine learning algorithms to build a classification model to assess the credit risk of corporate bonds and give default predictions [10]. Zhang used 2019 listed companies in China as a sample, and mainly studied whether the default distance variable in the Merton model has significant explanatory and predictive power for corporate credit risk, by comparing the explanatory power and predictive effect of the financial model, the extended Merton default distance model and the Merton Naive model and the hybrid model on credit risk, it was found that there was no significant difference in the predictive power and explanatory power of the two Merton models that adopted different assumptions and calculation methods [11].

In 2023, Zhang used four machine learning models, Logistic Regression, Random Forest, XGBoost and Light GBM, and integrated them with Voting voting algorithm to predict credit default based on historical data published by financial institutions. The numerical results show that the performance of the fusion based on the Voting voting algorithm has better prediction accuracy than the single machine learning model [12]. Li takes Jiangxi as an example to establish an enterprise credit risk evaluation index in line with the actual development of cross-border e-commerce in Jiangxi, and then uses BP neural network, K-nearest neighbor and support vector machine machine learning classification model to predict credit default of small and micro enterprises in Jiangxi. The empirical results show that the three deep learning classification models have good accuracy [13].

III. EMPIRICAL ANALYSIS

A. *Data Source*

In this paper, we use the loan record data of the credit platform provided in the Tianchi competition, the original dataset is 1.2 million, containing 47 columns of data features, such as loanAmnt, term, interestRate, installment, grade, subGrade, employmentTitle, employmentLength, which contains 15 columns of anonymous features for some lender behavior count features, and also desensitizes the employmentTitle, purpose, postcode and title fields. In order to improve the computing efficiency of the model and to take into account the time factor, a total of 100,000 data were randomly selected according to the feature issueDate (month of loan issuance), and the extracted data set contains data from 2007 to 2018.

Table 1. Data source (First five lines).

Loan Amnt	Interest Rate	Installment	Grade	...	Earlies Credit Line	Title
35000	19.52	917.97	E	...	Aug-01	1
18000	18.49	461.9	D	...	May-02	1723
12000	16.99	298.17	D	...	May-06	0
11000	7.26	340.96	A	...	May-99	4
3000	12.99	101.07	C	...	Aug-77	11

B. *Data Cleaning*

In daily life, the data obtained is often incomplete, there are some missing values, and some features have large differences between them, and are not stable enough to have many outliers, so it is particularly important to pre-process the data. Therefore, it is necessary to deal with this problem before modeling to ensure that the quality of the data can meet the task of data mining. In this paper, we have done three parts of data cleaning: filling in missing values, exploring outliers and transforming data features.

The first is the treatment of missing values, there are many missing values in this dataset, in order to avoid the impact of missing values on the prediction performance of the model, the missing values need to be filled. The approach taken in this paper is to eliminate this feature if there are more missing values (more than 30% missing); if there are fewer missing values, the Lagrangian interpolation method is used to interpolate the missing data.

Next, outliers are handled. Through data exploration, it is found that some features contain some outliers, however, in the field of financial risk control, the outliers may represent some special cases and can be regarded as a risk, therefore, this paper does not handle the outliers.

Finally, there are five category features in the dataset, namely grade, subGrade, employmentLength, issueDate and earliesCreditLine, which need to be converted into numerical features for better measurement model construction. For the features with obvious sequential relationships, they are custom coded; for the date variables issueDate and employmentLength, four numerical features are constructed using them, namely issueDateDT (the difference in days between the time of loan origination company and the time of the earliest loan origination company), numbers (the difference between the time of loan origination issueDate_now_year

(the number of years between the loan origination time and the earliesCreditLine_now_year (the number of years between the earliesCreditLine_now_year), and the original two date features are deleted after the construction. The original two date features are deleted.

C. *Factor Selection*

Not all of these features are useful for subsequent modeling, and according to their usefulness for modeling or not, the features can be classified into two categories: useful and useless, which can be called "relevant features" and "irrelevant features", respectively. Then the process of selecting relevant features from these features is called feature selection. Feature selection before modeling can not only reduce the dimensionality of the data, reduce the processing time of useless features, i.e., reduce the complexity of the computation time, but also improve the credibility of the model, and can more fully understand the information implied in the features.

In this paper, we mainly use filtered feature selection for factor selection. The filtering method first selects features for the dataset according to some rules (scoring each feature according to divergence or relevance, setting a threshold or the number of thresholds to be selected, and thus selecting features that satisfy the conditions), and then train the learner, and the feature selection process is independent of the subsequent learners. This is equivalent to "filtering" the initial features with the feature selection process and then training the model with the filtered features.

(1) *Removal of Covariance Features*

Covariance features are features that are highly correlated with each other. In the field of machine learning, high variance and low model interpretability lead to reduced generalization ability on the test set. In this paper, we find covariance features based on a specified correlation coefficient value (set a threshold of 0.9). For each pair of correlated features, it identifies one of them to be removed.

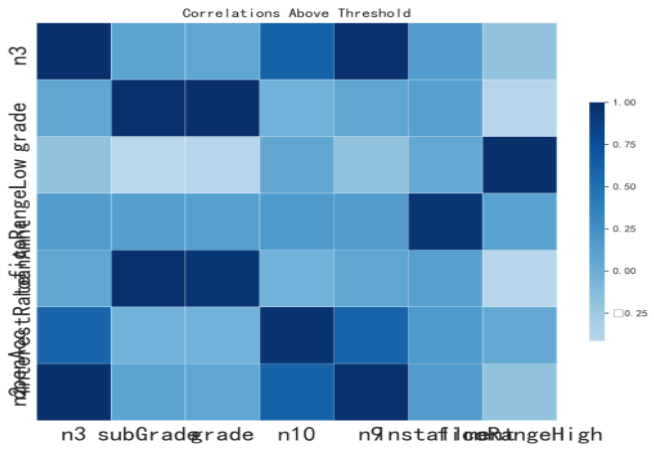

Fig. 1. Characteristic heat map for exceeding the threshold.

Co-collinearity can be well visualized using heat maps. Figure 1 shows all features with at least one correlation above a threshold (0.9), for a total of 28 covariance features, removed.

(2) *Remove Zero Importance Feature*

For supervised machine learning problems, where the labels of the training models exist and are non-deterministic, the features with zero importance are found according to the gradient boosting machine (GBM)

learning model. In this paper, we mainly use a tree-based machine learning model (e.g., boosting ensemble) to find the feature importance. The absolute value of this importance is not as important as the relative value, and the relative value can be used to determine the most relevant features for a task. It is also possible to use feature importance in feature selection by removing zero-importance features. In a tree-based model, zero-importance features are not used to segment any nodes, so they can be removed without affecting model performance. The gradient boosting machine from the LightGBM library in python is used to obtain the feature importance. To reduce the variance, the obtained feature importance is averaged over 10 training rounds of GBM. In addition, the model is trained using early stopping (early stopping) to prevent over fitting on the training data.

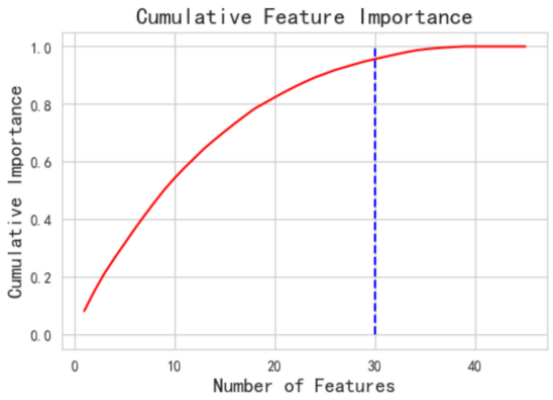

Fig. 2. Cumulative Importance of Number of Features Chart.

The model identifies six zero-importance features and removes them. Figure 2 shows the cumulative importance of the corresponding number of features, and the blue vertical line marks the threshold for a cumulative importance of 95%.

(3) *Remove Low Importance Features*

The next approach is based on zero importance identification, using the feature importance from the model for further selection. The features with the lowest importance are found by the previous model, and these features do not contribute to the specified total importance. A threshold of 0.95 is set to find the least important features, which achieve 95% importance even without these features.

Table 1. Importance table of all features.

Feature	Importance	Normalized_importance	Cumulative_importance
issueDateDT	149.6	0.08122	0.08122
dti	126.1	0.06846	0.14967
...	...	...	...
policyCode	0	0	1
n12	0	0	1

Table 1 shows the importance ranking of all features. Based on the previous cumulative importance graph and this information, the gradient boosting machine considers many features irrelevant to learning. Sixteen factors of low importance are identified, and the remaining 19 factors can reach 95% importance.

(4) *Removing a Single Unique Value Feature*

A feature with only a single unique value cannot be used for machine learning because the variance of this feature is zero, and then the feature is meaningless. For example, if a feature has only one value, then a tree-based model can never discriminate. Therefore these features need to be eliminated.

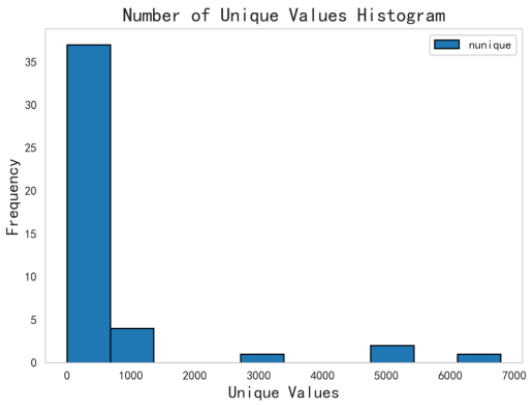

Fig. 3. Histogram of single values.

Figure 3 shows a histogram of the number of unique values for each category, identifying four factors with a single unique value and eliminating them.

After data cleaning and feature selection, the 15 most valuable features were retained from the original features, which will be used for modeling the various prediction models to facilitate the subsequent comparison of the prediction ability of various models.

C. *Model Construction*

The loan default prediction problem of loan users studied in this paper is a dichotomous problem, i.e., two outcomes of default and non-default. Therefore, five commonly used evaluation metrics for classification algorithms, Accuracy, Precision, Recall, the summed average of Precision and Recall (fl-score) and AUC (Areaunder Curve), are used to evaluate the model classification effectiveness.

In the original 1.2 million data set, 100,000 data were arbitrarily selected in the chronological order of loan disbursement as the data set studied in this paper. In addition, the descriptive statistics of the data in Chapter 3 shows that the data set is unbalanced (the ratio of positive to negative cases is about 4:1), therefore, before training the model, a balancing process is needed. The oversampled data are divided in the ratio of 4:1 and used for model training and prediction respectively.

Firstly, weak classifiers like decision trees and logistic regression are used for model construction, and the classification effect is found to be poor. Therefore, integrated algorithms are considered, and the decision tree based bagging and random forest in integrated learning bagging algorithm, and XGBoost and Catboost in boosting algorithm are established respectively, and three models with default parameters are built, and the effects of these models are compared according to the evaluation indexes, as shown in Table 3, and the appropriate model is synthesized and selected as the prediction model.

Table 3. Effect comparison of model.

Type	Classifier	Accuracy	Precision	Recall Rate	F1-score	AUC
Weak	SVM (Linear)	0.6602	0.6269	0.7866	0.6977	0.6977

Type	Classifier	Accuracy	Precision	Recall Rate	F1-score	AUC
Classifier	SVM (SGD)	0.6385	0.6362	0.6424	0.6393	0.6393
	Native Bayesian	0.6473	0.6350	0.6885	0.6607	0.6607
	Logistic Regression	0.6687	0.6574	0.7008	0.6784	0.6784
	Decision Tree	0.6925	0.7061	0.6568	0.6806	0.6806
	KNN	0.7187	0.6575	0.9099	0.7634	0.7634
Bagging	Bagging	0.8342	0.8770	0.7764	0.8237	0.8237
	Random Forest	0.7877	0.7925	0.7780	0.7852	0.7852
Boosting	Ada Boost	0.8334	0.8512	0.8070	0.8285	0.8285
	Gradient Boosting Classifier	0.7610	0.7414	0.7995	0.7693	0.7693
	Cat Boost	0.8356	0.8868	0.7684	0.8233	0.8233

Since this paper studies the default situation and aims to be able to predict defaulting users, the indicator of accuracy rate, although it represents the overall correct rate and has some reference value, is not an appropriate indicator for what is studied in this paper, so this indicator is not considered for the time being. The recall rate can be interpreted as the proportion of actual defaulted customers that can be predicted by the model, while the accuracy rate refers to the proportion of customers predicted by the model to be defaulted that actually have defaulted, and we tend to pay more attention to whether defaulted customers can be identified. If the model identifies defaulting users accurately, it can avoid most of the losses and harm caused to the bank due to the occurrence of non-performing loans, while for the case that the prediction is default but the customer actually does not default and the bank refuses to issue a loan, the bank will also have certain losses, such as a decrease in business volume leading to a decrease in revenue, but compared with the two, the former often has greater losses, so among the three indicators, namely, the accuracy rate, the recall rate and the F1 score, the recall rate is considered relatively more important and is the indicator to focus on. The AUC value, on the other hand, tends not to be affected when the positive and negative sample ratios change, while the other metrics will be more affected, so the AUC value should also be focused on.

From the data in Table 3, it can be seen that among the weak classifiers, KNN works better, with the highest recall and AUC values, indicating that among the weak classifiers, KNN has the best prediction effect. In Boosting, AdaBoost is better than GradientBoostingClassifier and Catboost. Therefore, AdaBoost is finally used in this paper for the construction of the loan default prediction model.

The accuracy of the model obtained according to the default parameters is 83.34%. On this basis, the parameters were adjusted using sklearn's grid search. The specific process is as follows.

Step 1: Determine the number of estimators for learning rate and tree_based parameter tuning.

max_depth is generally selected between 3 and 10, starting at 5; min_child_weight picks a relatively small value, starting at 1; Initial gamma=0; subsample, colsample_bytree is generally taken as 0.5-0.9, with a starting value of 0.8; scale_pos_weight=0.

The first step is to find the optimal number of decision trees according to the default value of 0.1 for the lear-

-ning rate, incrementing from 100 to 900.

Step 2: max_depth and min_child_weight parameter tuning.

Start by tuning these two parameters, as they have a big impact on the final result. First coarse tuning the parameters from a large range, and then small fine tuning. After outputting the results, we get the optimal max_depth of 9 and min_child_weight of 1.

Step 3: Tuning of gamma parameters.

Search gamma from [0, 0.1, 0.2, 0.3, 0.4] and get the optimal gamma as 0.2.

Step 4: Adjust the subsample and colsample_bytree parameters.

The ideal values for the subsample and colsample_bytree parameters are 0.8 and 0.9. We then take values around this value in steps of 0.05. The ideal values for the subsample and colsample_bytree parameters are still 0.8 and 0.9. The ideal values for the _bytree parameter are still 0.8 and 0.9. Then they are taken as final values.

Step 5: Regularization parameter tuning.

The next step is to reduce the overfitting by regularization, here using an adjustment of the reg_alpha parameter. The ideal value here is 0.1.

Step 6: Reduce the efficiency of learning.

The learning rate is reduced by a factor of 10 to 0.001, trees' number is increased to 5000, and score is increased.

The final tuning results are shown in Table 4:

Table 4. Table of tuning results.

Parameter Abbreviation	Parameter Meaning	Initial Value Parameter	Adjustment Results
max_depth	Maximum Depth	None	9
min_child_weight	Decision Tree	1	1
Gamma	Number of	0	0.2
Subsample	Minimum Sample Weight of Leaf Nodes	1	0.8
colsample_bytree	Random Sample	1	0.9
reg_alpha	Spanning Tree Column Sampling	0	0.1
learning_rate	L1 of Weights	0.1	0.001
n_estimators	Regularization Term	60	5000

Finally, this paper uses accuracy, precision, recall and F1-score to evaluate the Adaboost model. The evaluation results are shown in Table 5:

Table 5. Table of tuning results.

Classification Report	Precision	Recall	F1-Score
Not Default	0.83	0.95	0.89

Classification Report	Precision	Recall	F1-Score
Default	0.94	0.81	0.87
Accuracy			0.88
Macro avg	0.89	0.88	0.88
Weighted avg	0.89	0.88	0.88

The accuracy, precision, recall, F1-score, and AUC values reach 88%, indicating that the model performs better and can be used to accurately predict loan defaults. Thus, it can be assumed that the integrated learning algorithm has improved the predictive capability of the model. In addition integrated learning has a strong advantage in the stability of the model in comparison, from which it can be inferred that the Adaboost model has a huge advantage in predicting loan default.

IV. CONCLUSION

Nowadays, with the development of credit loan business, many loan default problems have emerged, and in this era of rapid development of big data, using data mining technology to predict the default behavior of loan users is a very important research topic. Therefore, this paper uses the loan record data of credit platforms to study the loan default problem of loan users. The results of this paper are summarized as follows:

(1) In the exploratory analysis and feature engineering of the data, it is understood that the dataset used in this paper is an unbalanced dataset (the ratio of defaulted users to non-defaulted users is about 1:4), and it is initially judged that features such as loan grade, loan amount, loan term, loan interest rate and upper and lower range to which fico belongs are closely related to whether a loan user defaults; after feature selection a total of 26 features are retained 26 features are retained after feature selection, and the quality of the filtered features is high.

(2) When comparing the modeling effects of the weak classifier and the integrated model, it can be found that the prediction ability of weak classifier is poor and unstable when the amount of data is large, while integrated model performs better. After combining all the indicators, it is found that the models are in the optimal state, thus determining that the integrated learning model is more advantageous in predicting loan defaults.

(3) Since defaulters generally account for a relatively small percentage of the actual loan situation, i.e., there is a positive and negative sample imbalance in the dataset, the accuracy, precision, and F1 scores are affected at this time, while the recall and AUC values are almost unaffected, the Adaboost model is chosen as the final default risk prediction model for predicting whether a borrower will default.

In summary, the Adaboost-based loan default prediction model constructed in this paper can provide a reference for commercial banks and other lending platforms when offering credit products to borrowers.

REFERENCES

[1] Liao Maozhong. Research on the Factors Affecting Student Loan Default [D]. Huazhong University of Science and Technology, 2008.
[2] Chen Xiong. Research on personal influence factors of credit card default risk [D]. Zhejiang University, 2011.
[3] Zhang Jinyu. Research on the factors influencing the default of personal credit loans of commercial banks [D]. Nanjing Agricultural University, 2013.
[4] Deng Ying. Research on the influencing factors of loan default of rural commercial banks for farmers [D]. Fujian Agriculture and

Forestry University, 2015.
[5] Tan Huifang. Study on the factors influencing the recovery rate of defaulted bonds in China [D]. Southwest University of Finance and Economics, 2021.
[6] Sun Ying. Research on factors influencing credit default of micro and small enterprises [D]. Zhejiang University, 2021.
[7] Yang Chongyang. Research on factors influencing default of online loans for small and micro enterprises in Q Bank [D]. University of Electronic Science and Technology, 2022.
[8] Kong, D.Y., Li, X.F. Default prediction: market model or accounting model [J]. Investment Research, 2012, 31(09): 127-140.
[9] Hong Lu. Research on the prediction model of default loss rate of housing mortgage loans [J]. Financial Forum, 2014, 19(01): 60-66.
[10] Guo Hanwen. Research on corporate bond default prediction based on machine learning model [D]. Zhejiang University of Finance and Economics, 2022.
[11] Zhang, Ziwei. Credit risk assessment of listed companies' debt - a study based on the validity of different Merton models [J]. National Circulation Economy, 2022, No. 2309(05): 150-156.
[12] Zhao Chuan, Ju Hongmei, Wang Meiling. Research on credit default prediction based on machine learning [J]. Computer knowledge and technology, 2023, 19 (05): 9-11.DOI : 10.14004.
[13] Li Zhiqiang, Yu Xuanpu. Research on credit risk model optimization of cross-border e-commerce small and micro enterprises based on ADASYN [J]. Journal of Jiangxi Normal University (Philosophy and Social Sciences Edition), 2023, 56 (02): 118-127.

AUTHOR'S PROFILE

First Author
Yifan Huang, School of Finance, Zhejiang University of Finance and Economics, Hangzhou, China.

Second Author
Yanxi Shao, School of Finance, Zhejiang University of Finance and Economics, Hangzhou, China.

Third Author
Dapeng Tang, School of Data Science, Zhejiang University of Finance and Economics, Hangzhou, China.

Fourth Author
Jie Huang, School of Economics, Zhejiang University of Finance and Economics, Hangzhou, China.

Fifth Author
Sijia Chen, School of Finance, Zhejiang University of Finance and Economics, Hangzhou, China.