Research on Credit Risk Assessment of Commercial Banks Based on Machine Learning

Yan Chen*

Shanghai International Studies University School of Business and Management ShangHai, China
** Corresponding author: yancychen87@gmail.com*

Abstract

The globalization and liberalization of the financial industry have intensified the operating risks of financial institutions. In the context of slowing macroeconomic growth, the rise in the rate of non-performing loans highlights the rising risks of the financial industry. This further illustrates the necessity and urgency of conducting credit risk analysis and early warning research. In order to alleviate the continuous increase of non-performing loans, this paper studies the credit risk analysis model based on machine learning. Through analysis, this paper proposes an XGBoost model for user loan risk prediction. This model has good prediction accuracy. Based on the results of the model, some suggestions are provided for the online lending platform to identify high-risk lending users.

Keywords-*machine learning; neural network; risk assessment;*

1. INTRODUCTION

Credit risk refers to the possibility that the borrower or the counterparty is unwilling or unable to perform the contract conditions due to various reasons, which will cause the bank or the counterparty to suffer losses. With the globalization and diversification of credit and financial markets, credit risks have become more complex, changeable and uncontrollable. Therefore, more and more banks are aware of the need to build a more professional credit risk management system to deal with the fact that financial risks are becoming more complex. Existing credit risk has been showing a growth trend around the world, and the existing risk management system still has many shortcomings in data collection, risk analysis and risk early warning [1]. In recent years, China's banking industry has exposed numerous scandals in credit management. Although electronic and informatized credit management is more convenient, it poses a higher challenge to professionalism. In order to improve the shortcomings of the traditional credit risk management system and make the credit risk analysis and early warning of commercial banks faster and more accurate, this paper proposes a credit risk assessment method based on machine learning which has important practical significance. And bank credit risk assessment is one of the problems that the current banking system needs to solve urgently [2].

2. COMMERCIAL BANK CREDIT BUSINESS PROCESSING FLOW

The main characteristics of corporate clients include large scale of assets, relatively secure debt repayment sources, sustainability of credit cooperation, and large dimensions of comprehensive contribution [3]. Corporate credit business is the dominant asset business of commercial banks, which can bring great benefits to commercial banks. Legal person credit business is a business that meets the credit needs of legal person customers. From the perspective of economic entities, the credit needs of legal person customers are more abundant. The commercial bank credit business processing process mainly consists of five stages, namely: (1) Target customers; (2) Pre-loan investigation; (3) Credit review and approval (credit risk assessment); (4) Signing of loan contracts and loan issuance; (5) Post-loan management. We can see the flow in Figure 1.

The overall process of credit risk management mainly used in China is shown in Figure 2. The first is to collect customer information. The second is credit risk assessment, to understand the past business conditions of the company, and to investigate, analyze and classify customers before lending. The third is customer credit management, which confirms the corresponding credit policy based on the customer’s credit rating. The fourth is loan management, which establishes the amount of

G. Vilas Bhau et al. (Eds.): MSEA 2022, ACSR 101, pp. 1119–1124, 2023.
https://doi.org/10.2991/978-94-6463-042-8_160

loans that customers can borrow in accordance with the level of credit granted to customers, and completes the loan operations in accordance with procedures [4]. The fifth is risk early warning and management, tracking the customer's repayment ability, continuously paying attention to the user's operation, and initiating risk early warning if a crisis situation affects the return of the loan. Finally, according to the above-mentioned information and experience, different strategies need to be formulated in the credit management process according to different situations, and different management decisions are made according to the industries of different types of enterprises, so as to ensure the timely return of loans and reduce bank losses.

Due to the particularity of China's credit risk management, the role of commercial banks in China's financial system is very special [5]. At present, the credit risk assessment and management of Chinese commercial banks generally suffer from problems in credit risk management operations, incomplete credit risk early warning systems, and poor comprehensive quality of credit personnel. Therefore, this article conducts an in-depth study on the credit risk assessment of commercial banks based on machine learning.

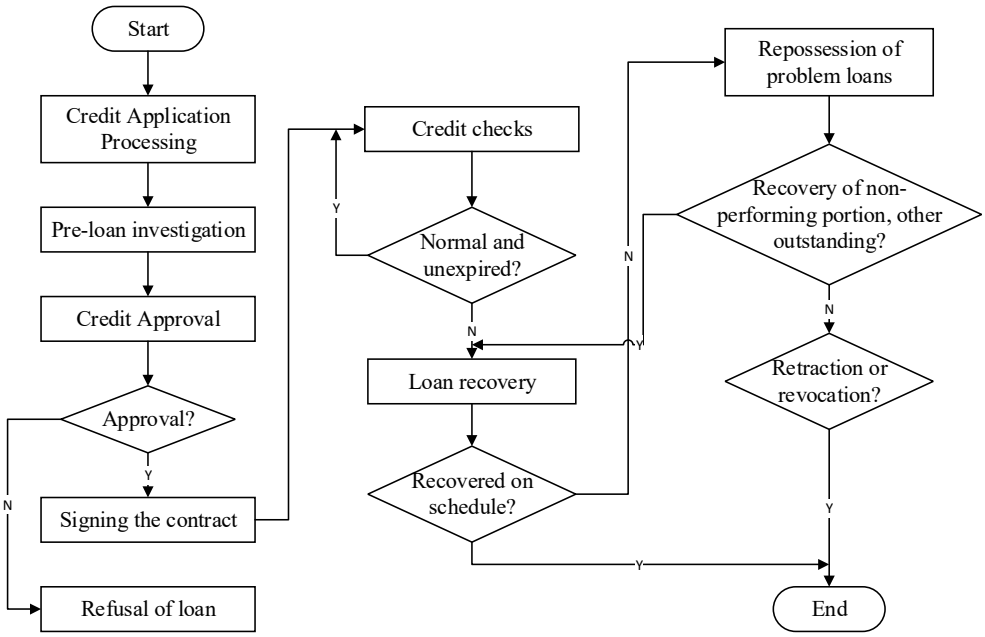

Figure 1. Credit business processing flow

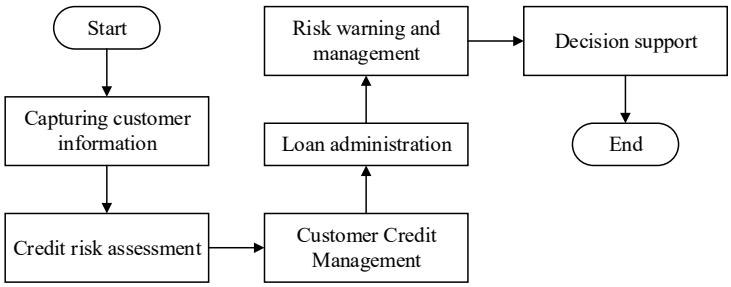

Figure 2. Credit risk management process

3. Analysis of Machine Learning Algorithms

After analysis and research, combined with the specific data set and sample size and variable type of this article, this article finally decided to adopt the XGBoost algorithm, and compare the prediction effect of the model with the prediction results of the logistic regression algorithm and the GBDT algorithm [6]. The basic principle of the algorithm and the pushing process are analyzed below.

3.1. Principle of XGBoost algorithm

The full name of XGBoost is Extreme Gradient Boosting, which is a C++ implementation of the gradient boosting tree algorithm. GBDT is a classic algorithm in Boosting. The main idea of the Booting algorithm is to serially combine multiple weak classifiers into a strong classifier [7]. GBDT adopts the idea of gradient descent when constructing the decision tree, that is, based on the previous decision tree, it is optimized and constructed in the direction of minimizing the objective function. XGBoost has been improved on the basis of GBDT.

XGBoost can implement parallel construction of decision trees, while adopting a second-order expansion method for the loss function, and adding L1 and L2 regularization to reduce overfitting.

Set sample set as $D = \{x_i, y_i\}(|D| = n, x_i \in R_m, y_i \in R)$,the model yields the predicted value $\widehat{y_i} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F,$ which $F = \{f(x) = w_{q(x)}\}$ is a collection of all decision trees, q is structural part of the tree, w is the leaf weight, and T is the number of decision trees.

Suppose the objective function is:

$$L(\phi) = \sum_i l(y_i, \hat{y}) + \sum_k \Omega(f_k) \quad (1)$$

Among them, $\Omega(f) = \Upsilon T + \lambda||w||^2/2, l(\phi)$ is a loss function, usually a convex function, measure the difference between the predicted value $\hat{y}$ and the true value $y_i.\Omega(\phi)$ is a regularization function, which can reduce the complexity of the model and alleviate over-fitting. The optimization goal is to minimize the objective function.

Perform the second-order Taylor expansion of the loss function at $\hat{y}_i\,(t)$ to get:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \approx \sum_{j=1}^{T}\left[g f_t(x_i) + \frac{1}{2}\mathrm{h} f_t(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (2)$$

In the above formula, $\mathrm{g_i}$ is the first derivative, h_i is the second derivative.

Definition $I_j = \{i|q(x_i) = j\}$ is the sample set on the leaf node j, then

$$L^{(t)} = \gamma T + \sum_{j=1}^{T}\left[\sum_{i\epsilon I_j} g_i w_j + \frac{1}{2} w_j^2 (\sum_{i\epsilon I_j} h_i + \lambda)\right] \quad (3)$$

When the tree structure q is known and the leaf node weight w_j of the above formula has a closed-form solution, the objective function [8]:

$$w^* = -\frac{\sum_{i\epsilon I_j} g_i}{\sum_{i\epsilon I_j} h_i + \lambda}, L^{(t)}(q) = \frac{1}{2}\sum_{j=1}^{T} \frac{(\sum_{i\epsilon I_j} g_i)^2}{\sum_{i\epsilon I_j} h_i + \lambda} + \gamma T \quad (4)$$

3.2. Model fusion method

Through the mutual fusion of multiple different forecasting models, the actual effect of further improving the performance of the forecasting model can be achieved. This method of fusion of predictive models has a universal meaning and application in the field of computing and machine intelligence learning. Commonly used methods of fusion of predictive models mainly include model-based averaging, bagging, and stacking methods. The technical foundation and structure of the model integration are shown in Figure 3. First, we train a part of a single classifier, and then combine multiple single classifiers with each other according to different mutual fusion strategies.

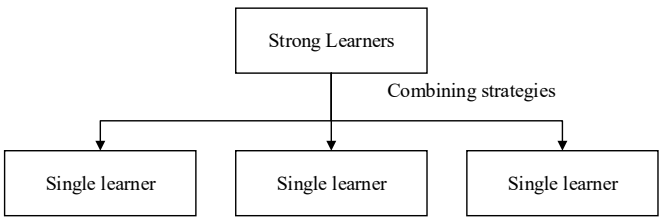

Figure 3. Model fusion structure basis

4. CREDIT RISK ASSESSMENT MODEL CONSTRUCTION AND ANALYSIS OF EXPERIMENTAL RESULTS

4.1. Construction of XGBoost-based credit risk assessment model

The parameters of the prediction model have a significant impact on the prediction effect of the model. Around this core, the key part of the user loan risk prediction lies in the pre-processing of the data and the adjustment of the parameters of the model itself, under the premise that the objective function is reasonably defined, in order to improve the prediction effect of the model. The parameters to be adjusted to the XGBoost model are as follows.

- Objective, the default value of this parameter is reg:linear, the other two commonly used values are binary logistic regression and softmax multi-classifier.
- Booster, used to select the model for each iteration, divided into tree-based models and linear models.
- eta, the learning rate, is used to improve the model by reducing the number of learning steps in each step, and is often taken to be between 0.01 and 0.2.
- max depth, the maximum depth of the tree, is used to prevent over-fitting.
- min_child_weight, the fitting parameter mainly indicates the smallest leaf node in a tree to fit the weight value, the main purpose of the fitting parameter is to prevent the occurrence of over-fitting, the parameter in the fitting of the weight value is too high will directly cause the tree over-fitting situation.

- gamma, this parameter specifies the need to split the nodes of the drop value of the loss function um, this parameter needs to be adjusted according to the loss function.
- silent, the parameter indicates the silent mode, taking a value of 1 when the model does not output the results of the run.
- nthread, this parameter is used to control the number of threads, for the number of cpu cores of the system.

After experiments, we decided to use the grid search method for parameter tuning, by adjusting the parameters and using a 10-fold cross-set for validation, dividing the data set into 10 samples without intersection, and selecting one of them respectively as the training set for the model [9]. The remaining 9 samples were used as the training set for the model and then the model was trained.

The final parameters of the model after adjustment are shown in Table 1.

TABLE 1. FINAL PARAMETER SETTINGS FOR THE PREDICTION MODEL

Parameter name	Parameter values
Objective	Binary: logistic
Booster	Gbtree
Eval_mrtric	Auc
Num class	2
Eta	0.1
Min_child_weight	3
Max_depth	8
Gamma	0.1
Cross_validation	10
Verbose	1
silent	0
nthread	12

4.2. Experimental analysis

This paper uses a dataset provided by an Internet lending platform, which is also unbalanced, to conduct the validation of the model effect and verify the prediction effect of the XGBoost model in this paper. If the prediction model proposed in this paper can have a good prediction effect on this test set, then it can be proved that the model is suitable for solving the loan prediction classification problem for unbalanced data.

In the process of this comparison experiment, the training set was first uniformly processed using the previous random forest algorithm-based feature selection method, and the top 15 features in terms of feature importance were selected to construct a new test set to participate in this comparison experiment. In order to further validate the prediction ability of the XGBoost model based on machine learning on unbalanced data, we used the GBDT algorithm and the classical Logistic algorithm, which are currently commonly used on Kaggle, to conduct prediction comparison experiments on the same dataset, with different parameters of each algorithm set to the best value after parameter The best value after tuning [10].

In this paper, the performance of the test set was predicted using two algorithms, the current logistic regression algorithm model and the latest GBDT algorithm model, and compared with the proposed forest-based model algorithm. The AUC values and accuracy rates of the three models were then obtained as the evaluation metrics of the models in this comparison experiment.

A comparison of the performance metrics of the three prediction algorithms on the data set shows that the XGBoost algorithm proposed in this paper outperforms the other two classifiers in terms of AUC value and accuracy, and improves significantly in terms of AUC metrics, as shown in Figure 4. A comparison of the performance evaluation metrics of the three prediction algorithms on the prediction set shows that the XGBoost algorithm proposed in this paper outperforms the other two classifiers in terms of AUC value and accuracy, and achieves significant improvement in the AUC metrics when predicting user loan risk models.

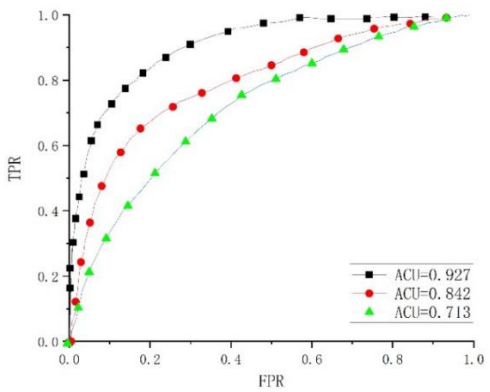

Figure 4. Comparison of AUC between logistic regression and GBDT and improved models on the test set

5. CONCLUSION

The control of credit risk, as a core element of risk management in Chinese commercial banks, has always been a key issue to be addressed. In recent years, state-owned commercial banks have made a lot of attempts and

practices in preventing loan risks and have achieved significant results and accumulated a lot of industry-related experience. Credit management has a decisive role to play in the bank's profitability, so credit risk analysis and early warning systems are essential. Modern data storage and computing technologies, as well as the development of artificial intelligence methods such as data mining and machine learning, are all providing safeguards and ideas for credit management. I hope to have more breakthroughs in my future studies and work, and to apply the findings of this study correctly in my work, to improve efficiency and control credit risks.

REFERENCES

[1] Y. Li. Risk assessment of Internet lending based on Bayesian net classifier. Northern Economic and Trade,2018(06):106-107.

[2] C. B. She. Study on the application of logistic regression on bank personal credit risk assessment. Technology and Innovation,2018(19):113-114+118-119.

[3] X. L. Zhong, F. Hou, S. L. Peng, et al. Risk control of big data for online personal credit. Journal of the University of Electronic Science and Technology (Social Science Edition),2018,20(05):7-11.

[4] W. X. Liao, B. Zeng, T. K. Liang, et al. A personal credit risk assessment method for high-dimensional data. Computer Engineering and Applications,2020,56(04):219-224.

[5] Y. Xu. Exploration of machine learning in financial risk management. Journal of Anshun College,2019,21(05):110-114.

[6] Z. Z. YANG, J. X. YUE, Z. T. ZHANG. Application of Bayesian algorithm in enterprise credit decision making. China Science and Technology Information,2020(24):104-105.

[7] T. Ning, D. Z. Miao, Q. W. Dong, X. S. Lu. Breadth and deep learning for overdue risk prediction. Computer Science,2021,48(05):197-201.

[8] Y. Zhang, X. P. Li. Research on risk management of personal consumer credit of commercial banks. Inner Mongolia Science and Technology and Economy,2021(06):74-75.

[9] W. Y. Yang. Research on credit strategy planning based on Bayesian neural network. China Business Journal,2021(10):85-87.

[10] Q. Rong, X. L. Huang. A preliminary study on credit risk scoring model for small and medium-sized enterprises in commercial banks. Journal of Science and Technology Economy,2021,29(19):219-220.