DLINE JOURNALS

Journal of Information Security Research

Print ISSN: 0976-4143
Online ISSN: 0976-4151

JISR 2024;15 (4)
https://doi.org/10.6025/jisr/2024/15/4/155-162

Financial Company Risk Prediction in the AI Era

Xiaojun Li, Xiyan Han*
School of Economy & management
Weifang University, 261061
Weifang, Shandong, China
hanxiyan176@163.com

ABSTRACT

Integrating the Internet and financial companies has expanded the market and avenues for personal loans. While the scale of personal loans has rapidly expanded, it has also brought higher default rates. This paper constructs a personal loan default risk prediction model based on an improved LightGBM model to control default rates and reduce financial company risks. The model's accuracy in predicting default risks is enhanced by optimizing model parameters and supplementing the evaluation system witha particle swarm optimization algorithm. Experimental results show that compared to four other risk prediction models, this model performs better in predicting default risks. The introduced indicators effectively reduce prediction errors, resulting in higher model accuracy and a better fit to real-world scenarios.

Received: 6 June 2024
Revised: 3 August 2024
Accepted: 21 August 2024

Keywords: Artificial Intelligence, Financial Companies, Credit Risk, Risk Prediction

1. Introduction

Consumer finance is an important area of business development for financial companies. People's consumption concepts have gradually evolved under national policies promoting consumption and economic development. Consequently, consumer finance has experienced rapid growth and has become a driving force for enhancing social consumption capacity. With the development and popularity of the internet, internet finance emerged, finding ample development space and market opportunities with the support of the internet and information technology. Credit plays a crucial role in Internet finance. Unlike traditional credit, internet credit platforms have higher efficiency in approval processes, providing more flexible and diverse loan services with a wider range of loan terms for customers. These advantages have rapidly expanded the market share of Internet finance in a short period, and it continues to grow quickly [1]. While financial companies provide a significant amount of consumer funds to customers, they also face the risk of personal loan defaults. An increase in default rates raises the risk of loan defaults for financial companies, making their operations more challenging.

To better control the risk of personal loan defaults, financial companies adopt credit scoring methods based on customers' personal information, consumption behaviour, and credit history to determine credit ratings and predict their repayment ability as a basis for assessing the likelihood of default. However, based on the recent financial company development reports, the average default rate for personal loans has not shown a clear downward trend and remains relatively high for some companies [2]. This indicates that the current credit reporting system cannot cover all individuals' credit information, and there are significant differences in financial companies' risk assessment capabilities. As a result, predicting personal loan risks has become an important research topic in financial company risk prediction.

The development of artificial intelligence technology provides new directions for risk prediction in financial companies and enables more accurate forecasting of risks. Therefore, this paper constructs a personal loan default risk prediction model based on the LightGBM model and introduces the particle swarm optimization algorithm to optimize the model, reducing and avoiding the problem of the model getting trapped in local optima and achieving better results in predicting personal loan default risks.

2. Research Development and Current Status of Financial Company Risk Prediction

Risk prediction and assessment for financial companies have always been a research focus both domestically and internationally. Scholars proposed a multivariate discriminant analysis model based on statistical knowledge in the early stages. This model selected five highly correlated indicators from numerous financial indicators as important predictors of bank bankruptcy risk. Experimental results showed that the model had significant predictive effectiveness [3]. Other scholars combined scoring methods and constructed corresponding credit scoring systems based on logistic regression models and discriminant analysis, showing higher importance in scoring results [4]. In addition, some researchers introduced entropy weighting into default risk prediction research, achieving comprehensive calculation of default risk by weighting risk indicators. Regarding loan default behavior, some scholars considered using consumer funds as an important factor affecting loan default behavior. Based on this, they constructed a consumer behavior change analysis model with time series characteristics, enabling the prediction of loan default risk [5]. Some researchers based their risk control on prior knowledge, combining Bayesian network models to construct loan risk control models. This allowed for the analysis of the association between default behavior and related factors to explain the default mechanism [6]. It was also noted that there is continuity between consumers' daily consumption behaviors, and abnormal reflections in daily behaviors may indicate the occurrence of default behavior. Thus, predicting default behavior can be achieved by monitoring abnormal points [7]. Furthermore, some researchers pointed out that machine learning can effectively improve the predictive performance of models. They introduced weight vectors into the random forest model, improved the voting decision mechanism, and reduced the model's prediction errors for default behavior [8]. Other researchers combined neural networks with logistic regression models to improve the prediction of loan default behavior. To further enhance the predictive performance of the model, some scholars introduced methods such as random sampling into the XGBoost model, enhancing the predictive capability and effectiveness of the model [9]. Additionally, researchers have developed LightGBM models specifically for P2P default behavior, significantly improving the predictive effectiveness and accuracy of the model for default risk. Although the LightGBM model emerged relatively late, it has shown good predictive performance and strengthened risk monitoring capability [10].

3. Personal Loan Risk Prediction Model Based on Improved LightGBM

3.1. Risk Prediction Indicators for Financial Company Personal Loans

In general, the risk prediction results for personal loan defaults by financial companies are based on customers' basic information, income status, and reliable and effective third-party credit report information. Although this information can effectively reflect customers' economic level, repayment capability, and credit situation to some extent, it may not fully capture all relevant information about customers. Therefore, before constructing the risk prediction model, this paper needs to expand the existing risk prediction indicator system of financial companies, increase the dimension of the indicator system, and achieve a more comprehensive representation of customer-related information.

Regarding basic information, financial companies typically obtain information about applicants' age, education level, income level, household registration, and address. The income level of the applicants is directly influenced by their age and education level. The repayment capability can be assessed by combining information about their employment and housing assets. Applicants with more stable income and better asset performance tend to have a lower probability of default. In addition, this paper adds information about the applicant's driving experience and net asset ratio. The former can help judge the applicant's consumption ability based on the time they obtained their driver's license. At the same time, the latter can assist in assessing the applicant's ability to repay the loan.

In terms of loan elements, the applicant's repayment decision is influenced by factors such as loan basic information, interest rates, and repayment methods. The longer the loan term and the higher the applicant's repayment interest rate, the higher the likelihood of default. Additionally, considering that different repayment methods affect the applicant's cash flow, a repayment method indicator is added in this paper.

Financial companies often use the People's Bank of China and Tencent Credit Information System for third-party data to assess applicants' default risk. Good credit information can enhance applicants' trustworthiness in financial companies' eyes. This paper adds tax payment records as part of the income level calculation to gain a more comprehensive understanding of the applicant's real income level. Moreover, financial companies can judge whether applicants have economic disputes based on their judicial information.

In addition to the above three dimensions of information, this paper introduces transaction behavior indicators, including expenditure frequency, withdrawal amount, investment ratio, etc. The paper believes that applicants' reported income level information may have some untruthfulness. By analyzing their transaction behavior, the daily income level and consumption patterns can be effectively evaluated, thus determining whether the applicant truly possesses the ability to repay. Furthermore, financial companies can analyze the applicant's investment status and scale based on the investment ratio. Different investment ratios and methods carry varying risks, and the applicant's investment inclination may affect repayment decisions.

3.2. Risk Prediction Model Based on Improved LightGBM

LightGBM is a prediction model based on a single decision tree model. It assigns weights to decision tree models through ensemble learning, effectively improving the predictive performance of individual decision tree models. Regarding personal loan default risk prediction, LightGBM calculates and performs multiple learning rounds based on sample gradient information and then adjusts based on the consistency between the predicted results and the actual situation. Suppose there is a significant error between the two predictions. In that case, the model will increase the corresponding weight in the new round of training, thus improving the accuracy of sample prediction in each round of training and ultimately enhancing the final predictive performance.

Let each sample be represented by where each indicator in the predictive indicator system is described, and each indicator's default status is described by indicating normal or default status. The cost incurred when the LightGBM model predicts incorrectly is represented by. This value can be obtained through AUC calculation. To improve the predictive performance of each iteration, the update formula is shown in Equation (1):

$$f_t(m) = L(n, f_{t-1}) + \beta C_t(m) \tag{1}$$

The equation denotes the loss function, and the decision tree model for default risk is described, with the learning rate represented.

This model aims to reduce the value of the loss function by adjusting the weights of each round of decision tree models. Therefore, the formula for computing the negative gradient is shown in Equation (2):

$$g_k^t = -\left[\frac{\partial L(n_k, f_t(m_k))}{\partial f(m_k)}\right] \quad (2)$$

Where represents the negative gradient of the sample at iteration, indicating the optimal improvement direction for the model.

Based on the computation result from Equation (2), the model can obtain the risk prediction model integrated with the new decision tree model using Equation (3):

$$c_j^t = \arg\min_c \sum_{m_i \in R_j^t} L(n_i, f_{t-1}(m_i) + c) \quad (3)$$

In the equation, different leaf nodes of the decision tree model are represented as, mainly describing different states of default samples, and the number of samples in different states is denoted as. The model is updated based on Equation (4):

$$f_t(m) = f_{t-1}(m) + \sum_{j=1}^{J} c_j^t I(m \in R_j^t) \quad (4)$$

Where when the leaf node with a decision tree sequence number is pointed to by the final sample path, its value becomes 1 or 0.

The final prediction result for individual loan default risk by the financial company is obtained through Equation (5):

$$F(m) = f_0(m) + \sum_{t=1}^{T}\sum_{j=1}^{J} c_j^t I(m \in R_j^t) \quad (5)$$

Based on the principles of the LightGBM model described above, it can be observed that the update rate of each round of the model depends on its learning rate. This value has a negative correlation with the precision of model iteration updates, meaning that a smaller value results in more fine-grained updates, but it also increases the likelihood of the LightGBM model encountering local optima during the update process. On the other hand, if the learning rate value is increased, the difficulty of the model reaching the optimal prediction state for default risk after each round of iteration updates becomes greater. Additionally, the depth and number of nodes in the risk prediction model's decision tree also affect the model's final prediction error. Therefore, this paper introduces the particle swarm optimization algorithm to improve the LightGBM model, and optimize parameters, and the optimization process is shown in Figure 1, the flowchart of the LightGBM model optimization based on the particle swarm optimization algorithm.

Firstly, the particle swarm initializes and randomly sets various parameters that particles represent. In this paper, the particle swarm consists of 500 particles and undergoes 1000 iterations. Secondly, based on the AUC value, the particle swarm predicts the performance of each particle's parameters in the LightGBM model. Afterwards, each particle's AUC value is recorded, denoted as, and according to Equations (6) and (7), the velocity and position of each particle in the swarm are updated accordingly:

$$v_m = w \times v_m + a_1 \times rand \times (P_{best}(m) - x_m) + a_2 \times rand \times (G_{best} - x_m) \quad (6)$$

$$x_{m+1} = x_m + v_{m+1} \quad (7)$$

Where is a constant a_1, a_2, $rand$ is a random number, and $rand \in [0,1]$.

Finally, based on G_{best}, the AUC variation value between iterations is evaluated, and when its value falls below a preset threshold, the corresponding parameters achieve the optimal result. When the number of iterations exceeds 1000, the parameters corresponding to are also the optimal results.

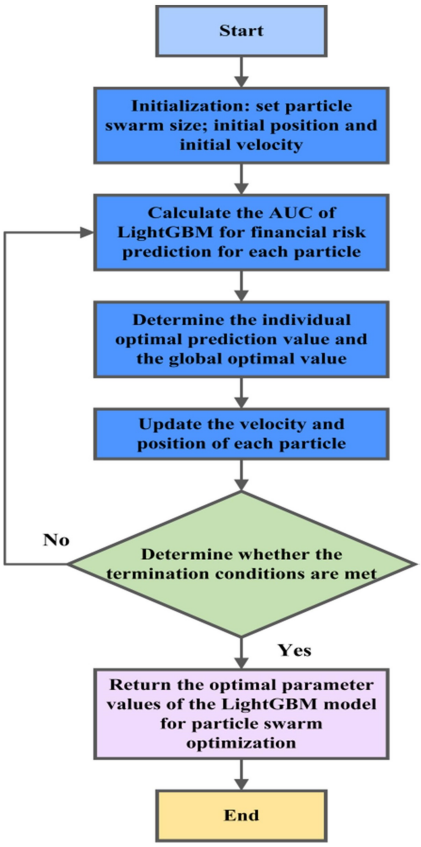

Figure 1. Flowchart of the LightGBM Model Optimization based on Particle Swarm Optimization Algorithm

4. Empirical Results Analysis of the Personal Loan Risk Prediction Model based on the Improved LightGBM

To verify the effectiveness of the personal loan risk prediction model based on the improved LightGBM, this paper selects 5 years of personal loan application data from a financial company and randomly samples data from normal and default loan categories with a ratio of 2:1, forming the training set and test set respectively. Meanwhile, for better validation of the model's predictive performance, this paper selects the BP neural network, decision tree, logistic regression, and XGBoost models as comparative models. As shown in Figure 2, the AUC and ROC curve comparison results of various risk prediction models in out-of-sample testing are presented.

From the results shown in figure 3, it can be observed that the AUC value of the personal loan risk prediction model based on the improved LightGBM is higher than 0.8. When considering other metrics, its predictive performance is found to be effective. The AUC values of other models are arranged in ascending order as follows: logistic regression model, BP neural network model, decision tree model, and XGBoost model, with the AUC value of the XGBoost model not exceeding 0.8. This indicates that the AUC values of the four models are all lower than the improved LightGBM model proposed in this paper. Additionally, the results of four

are all lower than the improved LightGBM model proposed in this paper. Additionally, the results of four important metrics show that the recall rate, precision rate, and accuracy of the improved LightGBM model are all above 0.9, while only the precision rate and accuracy of the XGBoost model achieve values above 0.9, which are still lower than those of the improved LightGBM model. As for the true negative rate, the improved LightGBM model reaches 0.85, while the values of other models are all lower. Therefore, it can be concluded that the improved LightGBM model has a higher predictive ability for personal loan default risk compared to the other four risk prediction models, demonstrating superior predictive performance and effectiveness.

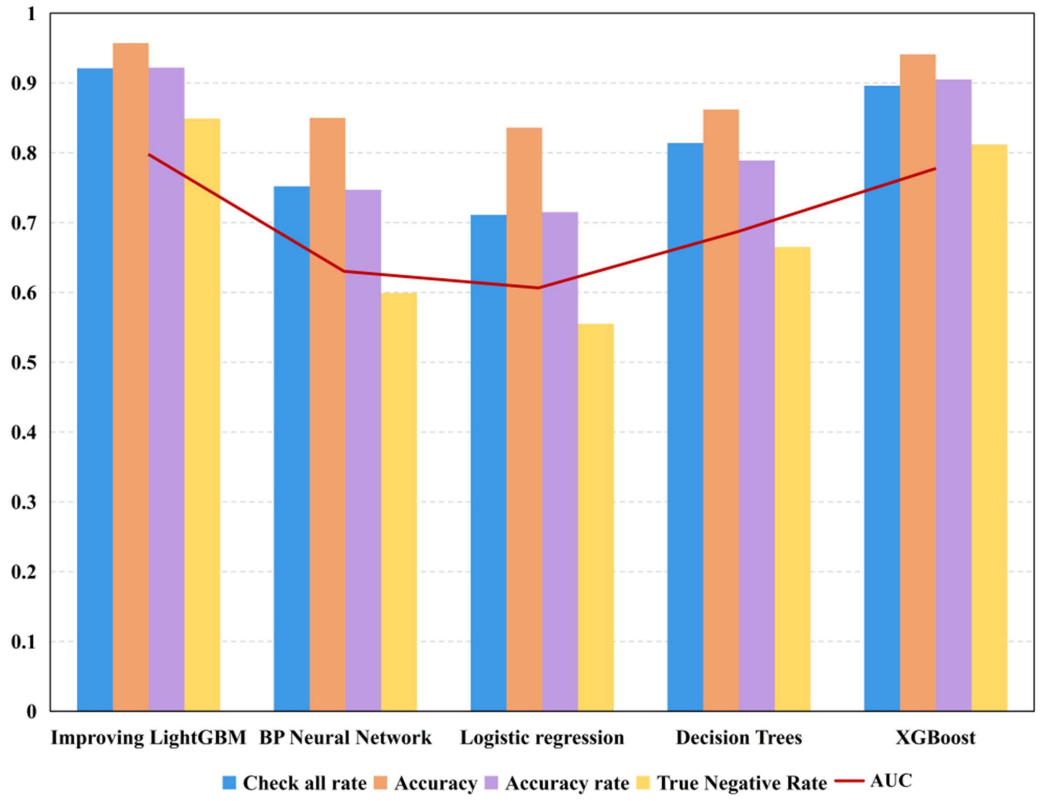

Figure 2. Comparison Results of Out-of-sample Testing AUC and Various Metrics for Different Risk Prediction Models

As shown in Figure 3, the results display the importance of the scores of indicators in the personal loan default risk prediction model based on the improved LightGBM. The overall data results show that the additional indicators supplemented by this paper based on the commonly used indicator system have played a positive role in predicting default risk. Compared with the results obtained from the income level proofs and corresponding materials collected by financial enterprises in the past, the proposed method of evaluating the applicant's income level based on the amount of tax paid can effectively reduce evaluation errors. The judicial litigation indicator reflects the applicant's credibility issues to some extent, providing multiple angles to verify the possibility of the applicant's default behavior. The net asset ratio reflects the applicant's repayment ability more accurately, while the transaction behavior indicates the applicant's repayment pressure. Lastly, the net income presents another perspective on the applicant's cash flow and debt-bearing capacity.

In conclusion, the personal loan default risk prediction model based on the improved LightGBM can effectively enhance the risk prediction performance and accuracy, and the additional indicators in the supplementary indicator system play a positive role in improving predictive effectiveness and reducing prediction errors.

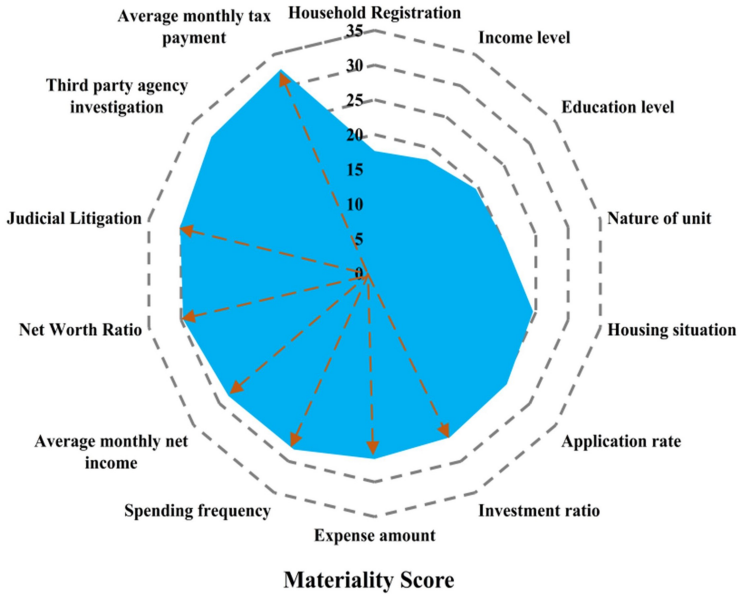

Figure 3. Importance Scores of Indicators in the Personal Loan Default Risk Prediction Model based on Improved LightGBM

5. Conclusions

Personal loan default risk prediction is a crucial factor affecting financial companies' operation and development. Generally, the prediction indicator system used by financial companies lacks comprehensiveness and multiple perspectives. Previous default risk prediction models also suffer from high error rates and significant variations in predictive performance. Therefore, this paper constructs a personal loan default risk prediction model based on improved LightGBM, optimizes the LightGBM parameters using the particle swarm optimization algorithm, and adds and supplements the existing prediction indicators. Experimental results show that the AUC value and other important metric scores of the improved LightGBM model are higher than those of the other four default risk prediction models. This indicates the proposed model exhibits better predictive performance for personal loan default risk. The added and supplemented indicators improve predictive effectiveness and accuracy, resulting in lower prediction errors and better alignment with real-world scenarios.

References

[1] Wu, F., Su, X., Ock, Y. S., et al. (2021). Personal credit risk evaluation model of P2P online lending based on AHP. *Symmetry, 13*(1), 83.

[2] Liu, Y., and Wei, Y. (2019). Copyright infringement problems of machine learning and their solutions. *Journal of East China University of Political Science and Law, 22*(2), 68–79.

[3] Chen, C. (2021). Research on the development path of consumer finance companies in China under the new normal. *Journal, 1*(9), 10–12.

[4] Zhao, B., and Gai, N. (2020). The impact of Internet consumer finance on the consumption structure of domestic residents—An empirical study based on VAR model. *Journal of Central University of Finance and Economics, 3*, 33–43.

[5] Zhu, Q., and Wang, L. (2016). Risk warning of internet financial private P2P lending platform. *Journal of Hebei University, 41*(4), 123–130.

[6] Sangwan, V., Prakash, P., and Singh, S. (2020). Financial technology: A review of extant literature. *Studies in Economics and Finance, 37*(1), 71–88.

[7] Jiang, F. (2020). Research on the personal credit evaluation model of online lending based on GA-SVM. *Control Engineering, 27*(6), 1025–1031.

[8] Bellotti, T., and Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*(2), 3302–3308.

[9] Zhang, H., and Dong, J. (2020). Prediction of repeat customers on e-commerce platforms based on blockchain. *Wireless Communications and Mobile Computing, 2020*, 1–15.

[10] Ma, X. J., Sha, J. L., and Niu, X. Q. (2018). Design and application of P2P project credit rating model based on LightGBM algorithm. *Journal of Quantitative Economy, 35*(5), 144–160.