Personal Credit Loans Risk Prediction Based on NS3-LightGBM

Juncheng Zhang

School of Big Data and Software, Chongqing University, Chongqing, 40044, China

dailan@ldy.edu.rs

Abstract. In recent years, the demand for personal credit loans has been increasing day by day. For banks, how to accurately identify and effectively predict whether borrowers will repay on time is a highly concerning issue. To effectively address a series of key problems existing in traditional loan risk prediction models, such as insufficient prediction performance, single hyperparameter optimization objectives, and poor model interpretability, this research integrates machine learning algorithms such as LightGBM and NSGA-III and builds an algorithm called NS3-LightGBM for predicting the probability of borrowers repaying on time. Testing and empirical analysis are done to confirm the suggested model's ability to make predictions. The suggested model has an accuracy rate of more than 86%, according to the results. in prediction, and its prediction ability is better than traditional machine learning models. In addition, through a more detailed analysis of the impact of various features on the prediction results, it is found that monthly income, the number of months on-time repayment per year, and whether there is a fixed job are key features that affect the possibility of borrowers repaying on time.

Keywords: personal credit loans, risk prediction, machine learning.

1 INTRODUCTION

Credit loans are one of the main businesses of banks, and one of its core businesses is judging the repayment ability of borrowers. Personal consumer credit is a product of financial innovation, which refers to the loan services provided by financial institutions to meet personal consumption needs [1]. This service has developed based on traditional financing business bets. Its growth is mainly attributed to two factors: the maturity of personal consumer credit business and internet technology. A large number of personal consumer credit products have emerged online and are widely [2]. Many financial institutions, banks, and enterprises have launched online loan services, making it a hot topic in Internet finance. According to a survey, the scale of personal credit loans consumption loans in China has exceeded that of personal loans [3]. With the growth of personal consumer credit demand, related risks are gradually emerging, among which personal credit loan risk is one of the primary risks [4]. According to official data, as of the end of March 2020, the number of operational online lending

A. Bhunia et al. (eds.), *Proceedings of the 2024 2nd International Conference on Finance, Trade and Business Management (FTBM 2024)*, Advances in Economics, Business and Management Research 304,
https://doi.org/10.2991/978-94-6463-546-1_7

institutions in China had dropped significantly by 86% year-on-year, leaving only 139 institutions left [5]. Many well-known platforms, such as INVESTING.com, JIMBUDONG.com, and RENRENJIAOGAI.com, have withdrawn from the market due to personal credit loans risk. According to the He and other research, consumer credit business always brings risks to financial institutions [6]. The establishment of a loan credit information database helps to reduce these risks. By assessing the credit risk of borrowing, lending platforms can more accurately understand the credit reliability of borrowing, thus preventing potential losses caused by the trust crisis [7].

Currently, consumer finance companies use various statistical and machine learning methods to predict loan risk, commonly referred to as credit scoring. Through the scores on the credit scorecard, this paper can intuitively analyze which individuals are more likely to be approved for a loan. The on-time repayment rate of customers is constantly changing, so it is necessary to regularly update each credit scorecard. The future stability of the credit scorecard is crucial, as accuracy declines could result in loans being allocated to worse-performing customers. To account for the time required for redeveloping, validating, and implementing credit scores, ensuring model stability and performance is very important.

This article utilizes the NS3-LightGBM algorithm for personal credit loan reputation score prediction, involving three main machine learning methods: LightGBM, and NSGA-III. Among them, LightGBM is an effective analysis method for credit risk evaluation. For example, based on the LightGBM algorithm, research on personal loan default prediction by Ma Xiaojun et al. confirmed that the LightGBM model outperformed other machine learning models, including decision trees and random forests, in terms of prediction performance [8]. Therefore, this paper uses it as the basic classifier for personal credit loan default risk prediction. Generally, the setting of hyperparameters in machine learning models has a significant impact on their classification performance. To illustrate the non-dominated sorting genetic algorithm (NSGA's) beneficial effects in this respect, De et al. suggested a fast NSGA-III based on NSGA [9]. Thus, this study builds a personal credit loan default risk prediction model based on LightGBM-NSGA-III by optimizing the hyperparameters of the LightGBM model using the NSGA-III method. This method increases the model's interpretability while simultaneously increasing prediction accuracy. The three aforementioned elements comprise the primary contributions of this research: (1) the Detailed design of a feature system for predicting credit default; (2) Through the integration of multiple machine learning algorithms, it effectively addresses a series of key issues, such as prediction efficiency, this paper generalization performance, and lack of interpretability, that exist in existing machine learning algorithms when applied to credit default prediction research.

2 LITERATURE REVIEW

Early scholars mostly used traditional statistical methods or default risk measurement models based on option theory to warn of borrow. This paper default risk. Among them, traditional statistical models mainly include the Z-Score and its improved models constructed by Altman; default risk measurement models mainly include structural and simple models [10]. The structural model is represented by the BSM model constructed by Merton [11]. In practical applications, the sample size of default data required to construct structural models is insufficient, making it difficult to achieve ideal warning effects [12]. The reductionist model may not be able to consider all possible influencing factors, resulting in a series of problems such as unstable estimation in the model [13].

In recent years, by actively introducing machine learning algorithms for research in the financial field, algorithms have shown significant improvements. For example, Wang Di et al. constructed a debt default risk prediction model based on machine learning methods and found that the prediction accuracy exceeded 80% [14]. Ke presented a LightGBM model with more consistent classification outcomes in 2017 that was based on the GBDT algorithm [15]. The LightGBM algorithm was used in a study on personal loan default warnings by Ma Xiaojun et al., who confirmed that the LightGBM model outperforms other machine learning models including decision trees and random forests in terms of warning performance [8]. In 2024, Deng Shanguang and others optimized the LightGBM algorithm to obtain LightGBM-NSGA-II-SHAP, which was used to study personal credit loan issues and achieved good results [16].

3 FORECASTING MODEL

To address a series of key issues, such as efficiency, poor generalization performance, and lack of interpretability in existing forecast models, this study builds a personal credit loans default forecast model based on LightGBM as the base model. Due to the large number of hyperparameters in the model and their significant influence on the model's forecast results, NSGA-III is used to optimize the important hyperparameters in LightGBM to improve the model's forecast effect and generalization ability. In addition, we compared this method with various other methods to demonstrate its superiority in predicting personal credit loan reputation scores.

3.1 LightGBM

LightGBM is a fast, efficient, and flexible machine-learning algorithm based on a gradient-boosting framework [17]. It uses a structure similar to an ensemble learning algorithm, but through a series of improvements, provides higher performance and flexibility compared to traditional decision trees or random forests. The main feature of LightGBM is its efficient implementation, especially for large-scale datasets, it can achieve training speeds in milliseconds. This article uses LightGBM for personal credit loan prediction.

The main steps of the LightGBM algorithm include:

1. Feature selection: LightGBM uses a method called feature importance score to select the most important features, which helps reduce the need for feature engineering and improve model performance.

2. Tree construction: LightGBM uses a tree structure based on Gradient Boosting Decision Trees (GBDTs) for modeling.

3. Gradient computation: Unlike random forests, LightGBM uses a method called "Gradient-based" to compute gradients, which provides better performance and stability.

4. Boosting process: LightGBM uses a method called "Gradient Boosting" to combine the predictions of multiple decision trees, which can improve the model's generalization ability and prediction accuracy.

3.2 NSGA-III

When the LightGBM model is used for personal credit loan default risk forecast, different hyperparameter configurations will produce different warning effects. This article optimizes the model's critical hyperparameters using the non-dominated sorting genetic algorithm (NSGA-III) to create an accurate and effective risk forecast model.

An enhanced variant of NSGA-II, NSGA-III is a significant algorithm in the field of multi-objective optimization [18]. To preserve population variety, it introduces widely dispersed reference sites, which innovates the selection process. Its primary benefit is its capacity to efficiently strike a balance between diversity and convergence to identify superior solutions on the Pareto frontier. The following are the primary NSGA-III steps: initialize the population, perform non-dominated sorting on individuals in the initial population, selection operations, crossover and mutation, and update.

3.3 Algorithm Ideas

This study used NS3-LightGBM to identify personal credit loan default risk. The research framework is shown in Fig.1 and is mainly divided into the following three parts:

Data collection and preprocessing: First, obtain raw data from the database and perform feature selection and cleaning. After that, split the created dataset equally into training and test datasets and provide the information for the appropriate preparation.

Model training and optimization: Train a LightGBM model using the training dataset, and then use the NSGA-III method to determine the best combination of hyperparameters to maximize the LightGBM model's performance.

3. Model testing and performance evaluation: Verify the optimized personal credit loans default identification model's recognition performance using the test dataset. Evaluate the model's performance in terms of accuracy and efficiency to ensure its accuracy and efficiency.

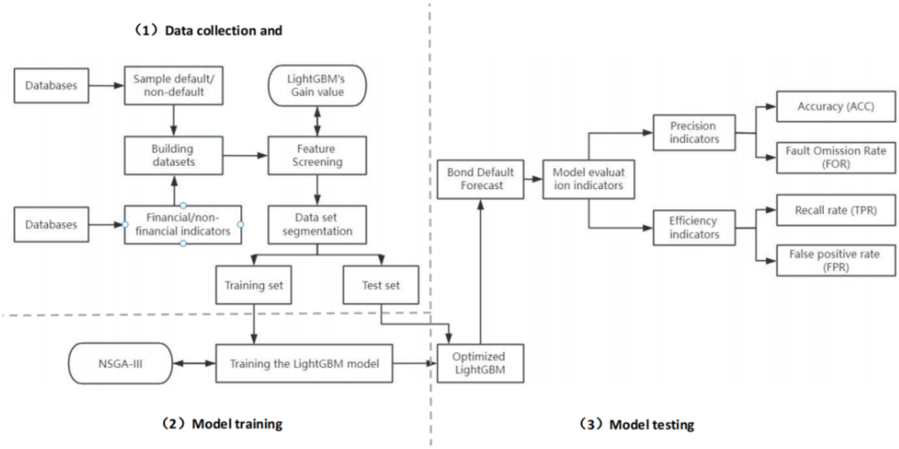

Fig. 1. Main idea of NS3 LighrGBM algorithm.

3.4 Data Description, Data Exploration & Data Preprocessing

This dataset was obtained from Kaggle. host machine learning competitions. The data comes from a machine learning competition called Credit Risk Model Stability.

Considering the wide variety of data types, as shown in Table 1, this article will divide the data into five categories of 17 small indicators: Existing situation, repaying ability, personal information, Personal reputation, and Purpose of loan, for further analysis:

Table 1. Overall dataset classification.

classify	Feature name	code
Existing situation	outstanding obligation	A1
	value in pledge	A2
	Guarantor assets	A3
Repaying ability	Monthly work income	B1
	Pension income	B2
	Debt situation	B3
Personal information	age	C1
	permanent occupation	C1
	family status	C3
Personal reputation	Whether there have been any other loans	D1
	Last loan maximum balance	D2
	Annual number of monthly payments on time	D3
	Number of years of violations	D4
	Last loan time	D5
Purpose of loan	lawful	E1
	rational	E2
	Use risk	E3

3.5 Feature Value Filtering

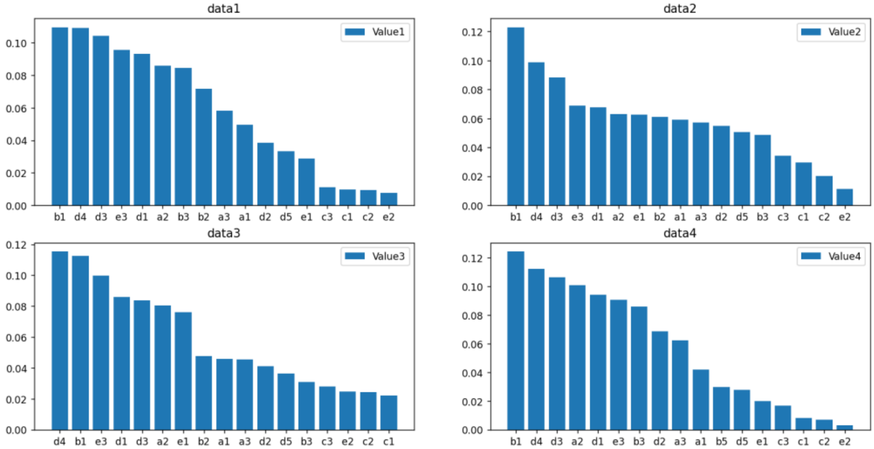

Fig. 2. Ranking of feature degree for four datasets.

In this article, to effectively improve the training efficiency of the proposed personal credit loans prediction model, before model training and optimization, original features are screened based on feature importance determined by calculating the LightGBM model gain value. Fig. 2 shows the processing results of four data sets. The four data sets are chronologically followed by the next quarter. The hoaxis represents the importance of each feature, ranked from top to bottom according to the sorting results. Features are selected from the top 60% based on the sorting results for feature screening. Following feature screening, the dataset is split into a test set and a training set at a 3:2 ratio. The test set is used to evaluate the model's prediction, and the training set is used to train the personal credit loans prediction model.

3.6 Model Training and Optimization

Table 2. Optimization results of hyperparameters for four datasets.

Data	learing_rate	num_leaves	max_depth	subsample
Data1	0.015	29	9	0.55
Data2	0.181	32	9	0.69
Data3	0.049	51	5	0.59
Data4	0.177	48	3	0.54

When optimizing the hyperparameters of LightGBM using NSGA-III, this paper selects four hyperparameters from the numerous model parameters: learning_rate (Lr), num_leaves (NL), max_depth (MD), and subsample (SS) based on reference by Lei [19]. A smaller Lr can lead to better classification performance, but a too-small value may cause overfitting. NL and MD are important parameters that determine the forecast performance and generalization ability of the model. A reasonable setting can suppress overfitting. SS can be used to accelerate model training. For the selected optimization objective function, the false negative rate (FOR) is minimized and the true positive rate (TPR) is maximized. The optimization results are shown in Table 2.

4 RESULTS

4.1 Comparing Models and Testing Methods

In addition to the NS3-LightGBM model used in this article, we also used NSGA-II in conjunction with LightGBM to obtain the NS2-LightGBM model. And LightGBM, Adaboost, ANN, and SVM as comparative models [20].

The proposed prediction model and comparison method have test results shown in the table on the testing set. Among the selected accuracy evaluation indicators, ACC represents the proportion of correctly predicted defaulted samples and non-defaulted samples among all test samples, which can reflect the accuracy of the model in predicting both defaulted and non-defaulted samples. FOR represents the proportion of actual defaulted samples incorrectly identified as non-defaulted. Specifically, when the ACC is larger or the FOR is smaller, the model's prediction accuracy is higher.

4.2 Analysis of Results

Table 3. Comparison results.

DATASET	NS3-LightGBM	NS2-LightGBM	LightGBM	Adaboost	ANN	SVM
			Panel A: ACC			
Data1	92.99%	86.64%	81.41%	78.97%	73.58%	71.43%
Data2	88.40%	84.11%	83.62%	74.69%	74.07%	69.77%
Data3	87.01%	83.54%	81.46%	73.31%	70.34%	69.33%
Data4	85.47%	82.26%	82.84%	70.64%	69.62%	68.38%
Average	88.47%	84.13%	82.52%	74.18%	70.94%	69.23%
			Panel B: FOR			
Data1	2.71%	8.99%	16.12%	26.82%	10.85%	8.99%
Data2	5.52%	10.18%	17.51%	25.19%	13.83%	10.18%
Data3	7.35%	10.82%	10.39%	35.92%	12.42%	10.82%
Data4	10.65%	15.73%	16.17%	32.99%	14.01%	15.73%
Average	6.56%	11.43%	15.05%	30.23%	12.78%	11.43%
			Panel C: TPR			
Data1	90.96%	83.96%	70.08%	53.40%	62.92%	83.96%
Data2	90.71%	82.11%	71.56%	55.75%	61.08%	82.11%
Data3	82.76%	81.38%	71.28%	50.97%	53.75%	81.38%
Data4	79.76%	72.29%	73.56%	49.24%	57.56%	72.29%
Average	86.05%	79.94%	71.62%	52.34%	58.83%	79.94%

			Panel D: FPR			
Data1	5.43%	11.80%	16.34%	14.53%	14.19%	26.32%
Data2	11.82%	14.61%	9.50%	19.75%	15.44%	27.41%
Data3	11.81%	15.06%	12.79%	22.70%	12.12%	12.96%
Data4	9.50%	12.07%	16.37%	22.15%	15.27%	14.55%
Average	9.64%	12.19%	13.88%	18.56%	20.87%	12.19%

Analyzing Table 3 and comparing the prediction accuracy indicators of different prediction models, it is possible to conclude the following:

1. Regarding the prediction results of four different datasets (data1, data2, data3, and data4 in sequential-quarter order), LightGBM (with average ACC and average F1-score) achieved prediction accuracy indicators of 82.52% and 15.05%, respectively. These are superior to those of Adaboost, ANN, and SVM. This indicates that LightGBM has better accuracy when applied to personal credit loan default prediction compared to these classical machine learning models.

2. In this study, utilizing the NSGA-III multi-objective optimization algorithm, LightGBM's hyperparameters were optimized., resulting in significantly improved prediction accuracy indicators. In addition, the NS3-LightGBM prediction accuracy indicators are higher than those of LightGBM, indicating that NSGA-III can effectively improve the recognition accuracy of LightGBM.

3. Comparing the prediction results for four different time windows, the accuracy rate of predicting personal credit loan defaults in the previous period is better than that of the other three-time windows. The closer to the occurrence of personal credit loan defaults, the higher the model's prediction accuracy, providing empirical support for the study of credit default timeliness.

4. Comparing NS3-LightGBM with NS2-LightGBM, NS3-LightGBM's prediction accuracy indicators are 88.74% and 6.56%, respectively, which are superior to NS2-LightGBM. This proves that NSGA-III can improve the method's performance better than NSGA-II.

Secondly, regarding prediction efficiency indicators, the greater the TPR or smaller the FPR, the better the prediction efficiency. Based on the TPR and FPR indicators in the table, it is possible to conclude the following:

1. For the four prediction time windows, LightGBM's TPR and FPR indicators are superior to those of Adaboost, ANN, and SVM.

2. The proposed NS3-LightGBM model has higher average TPR and FPR results for all four prediction time windows compared to other models, suggesting that the LightGBM base model's prediction efficiency can be effectively increased by NSGA-III.

Overall, this study's suggested NS3-LightGBM technique outperforms other comparison models in terms of prediction efficiency and accuracy. Using the NSGA-III algorithm for hyperparameter optimization can effectively improve model execution. Therefore, personal credit loan default risks can be successfully identified by the NS3-LightGBM technique described in this work. In addition, LightGBM outperforms Adaboost, ANN, and SVM in personal credit loan default prediction test data sets closest to the time of personal credit loan default occurrence.

5 CONCLUSION

This study proposes the NS3-LightGBM model to identify the default risk of borrowers. This paper repays on time, using existing circumstances, repayment ability, personal information, and personal reputation as warning features. Using the LightGBM model as a classifier to identify repayment default risk, the NSGA-III algorithm is used to optimize the hyperparameters of LightGBM, improving the accuracy and efficiency of the LightGBM model's forecast. Based on the experimental results of this study, the following conclusions are drawn: By comprehensively comparing the accuracy and efficiency indicators of the four datasets in the experiment, this study proposes a method for forecasting default risks for borrowers. These papers repay on time, which is superior to other benchmark methods in terms of accuracy and efficiency. The best warning effect was achieved in the dataset Data1 of the quarter before default, indicating that the closer to the outbreak of personal credit loan default, the higher the model's warning performance. Through analyzing the impact of features on the model's prediction results, it was found that features such as monthly work income, monthly on-time repayment, monthly violations, and risk level have a high correlation with default probability

REFERENCES

1. Haoyi, W., Zhangyang, X. Analyze the influence of various factors on personal loans. (eds.) Proceedings of the 2022 International Conference on Financial Technology and Business Analysis (part 4) (pp.167-178). International Accounting, Jimei University; (2022).
2. Jinzhi, L.: Research on the impact of internet consumption credit on college students' advanced consumption behavior – Based on the questionnaire data analysis of four universities in H City. Legal Expo, (31): 14-17 (2020).
3. Sanshao, Peng. Research on Advanced Ideas and Methods of Personal Loan Risk Management in Commercial Banks. (eds.) Proceedings of the 5th International Conference on Economic Management and Green Development (ICEMGD V) (pp.339-344). The University of Southern Queensland; (2021).
4. Zhuoran, L.: Research on the factors influencing personal consumer credit risk of XX bank Nanjing branch. Nanjing University of Posts and Telecommunications (2022).
5. Rockyan: 5000 institutions have exited the last P2P lending platform for online lending China Quality Miles, (8): 80-81 (2020).
6. Chengyijing, W., Haining, J., Xiaoyan, J. et al. Customer Credit Rating by Machine Learning.(eds.) Proceedings of 2022 International Conference on Company Management, Accounting and Marketing (CMAM 2022)(pp.388-396).Department of Social Science,University of California Irvine;Department of Social Audit,Nanjing Audit University;Department of Computer Science and Engineering,University of New South Wales, .(2022).
7. Yongsheng, Z., Jiali, C., Linyun, Z., et al.: Research on personal credit loans risk assessment based on improved random forest model. Credit report, 38 (1): 28-32 (2020).

8. Ma, X. J., Sha, J. L., Niu, X. Q.: Design and application of P2P project credit rating model based on LightGBM algorithm. The Journal of Quantitative and Technical Economics. 35: 144-160 (2018).
9. Deb, K., Jain, H.: An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: solving problems with box constraints. IEEE Transactions on Evolutionary Computation, 18(4), 577-601 (2014).
10. Altman, E. I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589-609 (1968).
11. Merton, R. C.: On the pricing of corporate debt: The risk structure of interest rates. Journal of Finance, 29(2), 449-470 (1974).
12. Huang, J. Z., Huang, M. How much of the corporate-treasury yield spread is due to credit risk? The Review of Asset Pricing Studies, 2(2), 153-202 (2012).
13. Duffie, D., Singleton, K. J.: Modeling term structures of defaultable bonds. Review of Financial Studies, 12: 687-720 (1999).
14. Wang, D., Zhanchi, W., Bangzhu, Z. Controlling Shareholder Characteristics and Corporate Debt Default Risk: Evidence Based on Machine Learning. Emerging Markets Finance and Trade(12), 3324-3339 (2022).
15. Meng, Q.: LightGBM: A highly efficient gradient boosting decision tree. Neural Information Processing Systems. Curran Associates Inc (2017).
16. Shangkun, D., Hong, N., Zonghua, L., Yingke, Z.: Explainable machine learning models for identifying default risk of corporate bonds. Computer Engineering and Applications. 1-15 (2024).
17. Meng, Q.: LightGBM: A highly efficient gradient boosting decision tree. Neural Information Processing Systems. Curran Associates Inc (2017).
18. Yihe, Q., Jinpeng, W. Multi-period portfolio optimization: A parallel NSGA-III algorithm with real-world constraints.Finance Research Letters104868, (2024).
19. Lei, S., Liang, X., Wang, X., Ding, J., Ge, X., Wang, F., Feng, J.: A Short-term Net Load Forecasting Method Based on Two-stage Feature Selection and LightGBM with Hyperparameter Auto-Tuning. In 2023 IEEE/IAS 59th Industrial and Commercial Power Systems Technical Conference (I&CPS), pp. 1-6. IEEE (2023).
20. Yuwei, Z., Haisong, H., Jianan, W.: Online tool wear status identification based on GA-LightGBM. Modular Machine Tool & Automatic Manufacturing Technique, (10), 83-87 (2021).