Innovative Design and Intelligent Manufacturing
L.C. Jain et al. (Eds.)

doi:10.3233/FAIA241096

Real-Time Data Analytics and Predictive Maintenance in Commercial Bank Credit Risk Management

Jing FU[a,1]
[a]*Software Engineering institute of Guangzhou, Guangzhou 510980, China*

Abstract. In order to identify customers with default risk and avoid credit risk, the application of real-time data analysis and predictive maintenance in credit risk management of commercial banks has been proposed. This paper will use CatBoost algorithm to study credit risk of credit card. This paper first preprocesses the data of 24 real-time variables, such as credit line, gender, age, education, marital status, repayment amount, repayment status, and bills payable, and selects 19 of them as the input variables of the model to establish a credit card user credit risk prediction model based on CatBoost algorithm. The results show that the accuracy of CatBoost is 91.73%, which is the highest among the five models, and the accuracy of Logistic is 74.39%, which is the lowest among the five models. Compared with other algorithms, CatBoost algorithm has higher classification accuracy for credit default prediction of credit card users. Conclusion: The model based on CatBoost algorithm has higher classification accuracy and can provide reference for commercial banks to predict credit card risk.

Keywords. Credit card; credit risks; machine learning; CatBoost algorithm

1. Introduction

Today, China's domestic economic environment has entered a new normal, small and medium-sized enterprise loans, financing and development is facing an unfavorable environment, commercial bank operations are also affected by the external social environment, their own business and strategic adjustments are more difficult, but also further affects the development of enterprise credit business [1]. In this regard, small and medium-sized enterprises financing difficulties, financing expensive, commercial banks small and medium-sized enterprises credit management challenges, both sides have their own reasons. Influenced by the social environment, small and medium-sized enterprises do not pay enough attention to their own credit rating, commercial banks do not pay attention to small and medium-sized enterprise credit business credit system construction. At present, China's financial market is still in the primary stage compared with foreign developed countries, the relevant system construction and laws and regulations are still missing, enterprise credit business risk management in the absence of laws and regulations is difficult [2]. From the point of view of the current public, the masses generally lack of credit consciousness, social credit awareness is weak. In this case, the

[1] Corresponding Author. Jing FU, fj17991008@163.com

financial institutions credit business risk management is insufficient, the lack of credit consciousness of small and medium-sized enterprises, the lack of supervision of the loan guarantee institutions, resulting in the financial credit business has a serious information asymmetry, the guarantee institutions cannot perform their due duties in the credit business, small and medium-sized enterprises in charge of the insolvency of the debt "run away", the news of bad debts in credit loans are repeated. The news of bad debts of credit loans has been repeatedly reported [3].

Under the influence of the economic environment, commercial bank credit business is facing great challenges [4]. The adverse impact of the macro-economy, affecting the level of enterprise cash flow management, and risk management of bank loans, many enterprises in this period through the financial packaging, peer-to-peer insurance and other ways to speculate to obtain commercial bank credit lines, due to the pressure of the development of top-down business, commercial banks often ignore the enterprise qualifications, credit tracking and detailed examination, increasing the risk factor of the credit business [5].

2. Literature Review

The relationship model between probability and company characteristic variables. Hassan, M. et al. were the first to carry out univariate bankruptcy prediction research, and found that the financial ratio of enterprises can reflect the financial situation of enterprises, and can predict the future credit development of enterprises [6]. Dzhaparov, P. et al. used financial data to study corporate default and established a financial crisis prediction model [7]. He took 79 bankrupt companies and non-bankrupt companies with the same industry and similar asset scale in the United States as samples, and through a single test of more than 30 ratios, studied the critical value of the financial ratio of bankrupt enterprises to normal enterprises for default, and used this value to predict enterprise bankruptcy. It is easy to see that a single financial ratio is difficult to fully reflect the financial information and credit status of the enterprise, and its forecasting ability is poor, so it is gradually replaced by multivariable analysis. Singh, J. and others proposed the famous Z-score model, modified and expanded the original model, and established the second-generation ZETA model [8]. The scoring model proposed by Cornwell, N. et al. assumes that the collected data samples obey normal distribution and the covariance between the data is equal. On this premise, a linear discriminant model is established [9]. The Z-score model takes 66 listed companies in the United States as the research object, and divides them equally into bankrupt companies and non-bankrupt companies. It uses 22 financial ratios, and finally determines the 5-variable Z-score discriminant model after statistical test analysis and screening. Later, it is improved to the 7-variable Zata model. Similar studies include Horrigan, West, Altman and Katz.

The credit card user credit risk prediction model to be established in this paper is a binary classification model, the output of the model is either "1" or "0", and "1" means that the user will default next month and is a "bad customer". "1" means that the user will default next month and is a "bad customer", commercial banks need to pay attention to this kind of customers, and can understand the reasons for overdue credit card holders and their willingness to repay by phone, etc., while "0" means that the user will not default next month and is a "good customer". A "good customer", this kind of credit

situation of excellent customers can be moderately increasing the credit limit, in order to bring more revenue for the card issuer.

3. The Research Methodology

CatBoost algorithm is a new machine learning algorithm based on gradient enhancement library. When CatBoost processes various data types such as text and audio, it can convert data categories into numbers without any explicit pre-processing, and it can provide strong accuracy without a lot of data training. At present, some scholars have used CatBoost algorithm for e-commerce, disease prediction for research in water resources management and other fields, this paper applies CatBoost algorithm to the prediction of credit risk of credit card users [10]. The data used in the model is the data after feature selection. The input variables include 19 variables, including default in the next month, credit line, gender, age, education, marital status, repayment amount, repayment status, and account payable. Set the training set to account for 75% of the data set, and the test set to account for 25% of the data set. Use the grid search method to find the optimal parameters of the model, See Table 1 for specific parameter values, and then use the CatBoostClassifier module in Python software to establish the model. The following formula (1): No default

$$Y = \begin{cases} 1, & \text{break a contract} \\ 0, & \text{No breach of contract} \end{cases} \tag{1}$$

Table 1. Optimal parameters of CatBoost algorithm under grid search.

parameters	Parameter interpretation	parameter values
iterations	Maximum number of trees	100
depth	the maximum depth of the base model	6
learning_rate	the learning rate of the model iterations	0.03
loss_function	The loss function	Logloss

Finally, the training set is used to train the model, and the trained model is applied to the test set. According to the program running results, the accuracy of the credit risk prediction model obtained using the CatBoost classification algorithm is 91.72%, and the AUC value is 0.86. The AUC value represents the area under the ROC curve, and the AUC value of the model is greater than 0.8, indicating that the credit card credit risk prediction model based on the CatBoost algorithm is effective and acceptable [11].

Table 2. Prediction effect of classification model based on CatBoost.

V model	Accuracy	AUC value
CatBoost	91.72%	0.86

4. Analysis of Results

4.1 Methodology for Model Evaluation

When different classification algorithms are used to build credit risk prediction models for credit card users, the models will have different effects. In order to find out the algorithm most suitable for building credit risk prediction models for credit card users among several algorithms, it is necessary to evaluate the performance of each model. There are many criteria for evaluating model performance, including accuracy, confusion matrix ROC curve and AUC value, etc. For the credit risk prediction model of credit card users, the most important thing is the accuracy of the model prediction. The model accurately identifies customers who are likely to default and high-quality customers, which can help banks expand their interests and avoid economic losses [12].

(1) Confusion matrix

The accuracy of the model is generally the number of correctly predicted samples divided by the number of all samples, but when the classification of the sample has three or more than three samples, it is difficult to judge the model's good or bad by only knowing the model's prediction accuracy, so we do not know whether each category of the model can be accurately predicted, and then the confusion matrix can help to find out the hidden information, and the confusion matrix can clearly show the results of the actual classification and the prediction of the classification. The confusion matrix can clearly show the results of the actual classification and the predicted classification, taking binary classification as an example, the confusion matrix is usually presented in the form of Table 3.

Table 3. Confusion matrix (binary).

	0 (actual)	**1 (actual)**
0 (forecast)	TN	FN
1 (forecast)	FP	TP

Taking whether a credit card defaults as an example, "0" represents non default, "1" represents default, "TN" in the table represents the number of non-default customers correctly predicted, "TP" represents the number of default customers correctly predicted, and "FN" represents the number of default customers incorrectly predicted, that is, some default customers are predicted to be non-default customers, "FP" It means that the number of non-defaulting customers is incorrectly predicted, that is, some non-defaulting customers are predicted to be defaulting customers.

From this table, it is possible to calculate the accuracy of the classification, i.e., the number of correctly predicted samples divided by the number of all samples, i.e., the following equation (2).

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

After knowing the accuracy of the classification, the error rate of the classification can also be calculated, i.e., the following equation (3).

$$error = 1 - accuracy = \frac{FP+FN}{TP+TN+FP+FN} \tag{3}$$

(2) Completeness and accuracy

The detection rate, or recall rate, can be obtained by calculating the ratio of the number of correctly predicted "1s" to the number of actual "1s", i.e., equation (4) below.

$$recall = \frac{TP}{TP+FN} \tag{4}$$

The accuracy rate refers to the ratio of the number of correctly predicted "1" to the number of all predicted "1", i.e., the positive hit rate, i.e., the following formula (5).

$$precision = \frac{TP}{TP+FP} \tag{5}$$

Credit card default, for example, check the full rate of measurement is correctly predicted the number of default customers accounted for the actual number of all default customers in the proportion, and check the rate refers to the number of correctly predicted the number of default customers accounted for the prediction of the number of all default customers in the proportion of the classification of the check the full rate and the check the rate of contradiction exists, when the check the full rate rises, the actual "0" When the detection rate rises, the number of wrong samples that are actually "1" and predicted to be "1" increases, and then the detection rate decreases; when the detection rate rises, the number of wrong samples that are actually "1" and predicted to be "0" increases, and then the detection rate decreases. When the checking rate increases, the number of wrong samples with actual "1" and predicted "0" will increase, and then the checking rate will decrease. The

(3) ROC curve and AUC value

When evaluating the advantages and disadvantages of the two classifications, in addition to calculating the accuracy of the classification, we also need to look at the ROC curve of the classification. The ROC curve is an arc across the origin. Its horizontal axis FPR represents the ratio that is actually "0" but is predicted to be "1", while the vertical axis TPR represents the ratio that is actually "1" and is predicted to be "1". The ROC curve can be used to judge the pros and cons of classification. When the TPR is higher, the accuracy of the model is higher, while the FPR is lower, the misjudgment rate of the representative model is lower. Therefore, when the ROC curve is observed, the closer the curve is to the upper left of the coordinate axis, the better the classification effect and the better the performance of the model are.

In some cases, just observing different ROC curves can compare the advantages and disadvantages of the two models, so the AUC value is introduced. The AUC value refers to the area under the ROC curve, which is a specific value. The AUC value of general models is between 0.5 and 1. The larger the AUC value, the better the performance of the model. When the AUC value of the model is greater than 0.8, this type of model is effective.

4.2 Comparison of Models

In the previous article, CatBoost, Logistic regression, random forest, GBDT and XGBoost are used to predict the credit risk of credit card users, and Python software is used to give the accuracy of classification prediction, ROC curve and AUC value of the five models. The more the ROC curve deviates from the diagonal, the better, that is, the larger the area below the ROC curve, the better. The ROC curves of these five models are similar, and it is difficult to compare the performance of each model separately from the graph. Therefore, the AUC value is introduced. Next, the model performance will be compared from the accuracy and AUC value.

Table 4 shows the accuracy of the five models, of which the accuracy of CatBoost is 91.73%, which is the highest among the five models, and the accuracy of Logistic is 74.39%, which is the lowest among the five models. Therefore, CatBoost algorithm is feasible for credit default prediction of credit card users, and has higher classification accuracy than other algorithms.

Table 4. Accuracy of the model.

Model	CatBoost	Logistic	Random Forests	GBDT	XGBoost
Accuracy	91.73%	74.39%	84.87%	78.38%	86.28%
AUC value	0.86	0.75	0.79	0.78	0.78

Table 4 also shows the AUC values of the five models. The AUC value of CatBoost is 0.86, the AUC value of Logistic regression is 0.75, which is the lowest of the five models, the AUC value of both GBDT and XGBoost is 0.78, and the AUC value of random forest is 0.79. If only the AUC value is used to judge the performance of the model, From the table, we can see that the AUC value of the credit default prediction model for credit card users based on CatBoost algorithm is 0.86, which is the highest among the five models, so the performance of this model is the best.

Combining the accuracy and AUC value to determine the performance of these five models, it can be found that the overall performance of the credit risk prediction model for credit card users based on CatBoost algorithm is better than the four models of Logistic regression, random forest, GBDT and XGBoost. Therefore, it is feasible to use CatBoost algorithm to establish a credit risk prediction model for credit card users, and can effectively predict whether users have the possibility of default. The prediction accuracy is 5.54% higher than XGBoost, which can provide some reference for commercial banks' credit card business and help commercial banks to avoid credit risks to a certain extent, take relevant measures in advance for customers with default risk to reduce the bank's economic losses.

5. Conclusion

The default of credit card users has brought great economic losses to major card issuers, so how to identify customers with default risk, take corresponding measures in advance, and reduce the economic losses caused by credit risk has become the main work in the credit card business. This paper first preprocesses the data used, and then uses five

algorithms CatBoost, Logistic regression, random forest, GBDT, XGBoost to model respectively. Finally, by comparing the accuracy and AUC value of each model, it determines the superiority of CatBoost algorithm applied to credit risk prediction of credit card users.

Funding

This research was funded by Scientific research and teaching project of Software Engineering Institute of Guangzhou, grant number ky202325.

References

[1] Addy, W. A., Ugochukwu, C. E., Oyewole, A. T., Ofodile, O. C., Adeoye, O. B., & Okoye, C. C. (2024). Predictive analytics in credit risk management for banks: A comprehensive review. GSC Advanced Research and Reviews, 18(2), 434-449.
[2] Liang, Y., Quan, D., Wang, F., Jia, X., Li, M., & Li, T. (2020). Financial big data analysis and early warning platform: a case study. IEEE Access, 8, 36515-36526.
[3] Lin, M. (2022). Innovative risk early warning model under data mining approach in risk assessment of internet credit finance. Computational Economics, 59(4), 1443-1464.
[4] Wang, F., Ding, L., Yu, H., & Zhao, Y. (2020). Big data analytics on enterprise credit risk evaluation of e-Business platform. Information Systems and e-Business Management, 18(3), 311-350.
[5] Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology, 4(1), 111-138.
[6] Hassan, M., Aziz, L. A. R., & Andriansyah, Y. (2023). The role artificial intelligence in modern banking: an exploration of AI-driven approaches for enhanced fraud prevention, risk management, and regulatory compliance. Reviews of Contemporary Business Analytics, 6(1), 110-132.
[7] Dzhaparov, P. (2020). Application of blockchain and artificial intelligence in bank risk management. Икономика и управление, 17(1), 43-57.
[8] Singh, J., Singh, G., Gahlawat, M., & Prabha, C. (2022). Big data as a service and application for indian banking sector. Procedia Computer Science, 215, 878-887.
[9] Cornwell, N., Bilson, C., Gepp, A., Stern, S., & Vanstone, B. J. (2023). The role of data analytics within operational risk management: A systematic review from the financial services and energy sectors. Journal of the Operational Research Society, 74(1), 374-402.
[10] Arsic, V. B. (2021). Challenges of financial risk management: AI applications. Management: Journal of Sustainable Business and Management Solutions in Emerging Economies, 26(3), 27-34.
[11] Al-Dmour, H., Saad, N., Basheer Amin, E., Al-Dmour, R., & Al-Dmour, A. (2023). The influence of the practices of big data analytics applications on bank performance: filed study. VINE Journal of Information and Knowledge Management Systems, 53(1), 119-141.
[12] Wang, L., & Wang, Y. (2022). Supply chain financial service management system based on block chain IoT data sharing and edge computing. Alexandria engineering journal, 61(1), 147-158.