Research on Credit Card Overdue Risk Prediction based on CatBoost Model

Zhihui Zhuo*

{zhuo2023012@163.com* }

School of Information And Safety Engineering, Zhongnan University of Economics and Law ,Wuhan,China

Abstract. At present, credit card business has become an indispensable financial service for commercial banks and individuals. With the development of credit card business in contemporary transcations, the overdue risk has extremely influenced the utilization of credit cards and affect the economical environment. Existing machine learning models have achieved promising performance toward complex data and become one of most widely used prediction method especifically based on Optuna optimization, which provides valuable insights for improving the level of credit card risk control in commercial banks. In this work, we initially explores transaction flow data that best reflects customer behavior and conducts in-depth analysis of credit card delinquency-related behaviors using theories including behavioral economics and information asymmetry. Furthermore, we construct a CatBoost model based on Optuna hyperparameter optimization to predict credit card delinquency risk. The experiment collects credit card transaction data from various commercial banks and conducts further predictive analysis, providing improvement recommendations that can be referenced for credit card risk management in banks. From our extensive simulations, we can conclude that our model can achieve acceptable prediction accuracy with reasonable responsing costs.

Keywords: Credit card; Expected risk; CatBoost model; Optuna; Risk Prediction

1 Introduction

With credit cards occupying a prominent position, millions of consumers utilize credit cards due to the convenience, purchasing power, and financial flexibility. However, with the benefits of credit cards comes the responsibility of managing credit effectively, and one critical aspect of this responsibility is avoiding overdue payments [1]. Although credit card business brings substantial profits to banks, the lax management of credit cards has led to high default rates among credit card customers, posing significant risks to banks. Therefore, whether to issue credit cards to users and which users to target has become a hot topic in the banking industry. This work utilizes machine learning models to assist banks in making novel decisions and developing new strategies for credit card issuance, contributing to the development of the banking industry in our country [2]. Specifically, overdue risk is known as the likelihood that a cardholder will fail to make the required minimum payment on their credit card by the due date. This risk is inherent in any credit-based system and is influenced by various individual and systemic factors [3].

MSIEID 2023, December 08-10, Guangzhou, People's Republic of China

DOI 10.4108/eai.8-12-2023.2344364

In this work, we construct a predictive model for identifying suitable users for credit card issuance using the CatBoost model based on Optuna optimization. The model's prediction performance is evaluated using accuracy evaluation metrics including Receiver Operating Characteristic (ROC), average precision, and confusion matrix, comparing the prediction effectiveness before and after optimization under different feature selection methods. The experiment in this article utilizes relevant data of credit card holders from a certain bank.

2 Related Works

Initially, Support Vector Machine (SVM) models are proposed to predict credit card default risk, and researchers have achieved good results when using this model [4]. SVM is a machine learning model based on binary classification and has the largest linear classifier in the feature space. By using kernel functions, the model can be transformed into a nonlinear classifier. SVM models are highly applicable to small-sample data, overcome the curse of dimensionality and nonlinear separability problems, and have fast computation speed [5]. SVM models are widely used in credit card centers for predicting credit card default risk and have shown good performance in practice. This article will compare SVM models with CatBoost models to verify if CatBoost can better predict credit card default risk [6].

Subsequently, Backpropagation (BP) neural network is a traditional neural network model that uses the Backpropagation algorithm and is not a popular deep learning model currently [7]. The key to machine learning lies in overcoming the problem of BP algorithm's inability to train deep networks. The training of BP neural networks involves repeatedly adjusting neuron weights to accumulate errors and generate an artificial neural network system that can simulate the original problem [8]. In credit card centers, BP neural networks are commonly used models for predicting default risk.

3 Model Framework

3.1 Feasibility Analysis

The original intention behind the design of the CatBoost algorithm was to better handle categorical features in GBDT (Gradient Boosting Decision Trees). In GBDT, the simplest method for handling categorical features is to use Greedy Target-based Statistics (Greedy TS) to replace them with the average value of the corresponding labels. This method also serves as the criterion for node splitting in decision trees. Overall, CatBoost effectively addresses the challenges GBDT faces when dealing with categorical features, thereby improving the accuracy and efficiency of the algorithm, which is expressed in following Equation 1.

$$\hat{X}_k^i = \frac{\sum_{j=1}^{n} [X_{j,k}=X_{i,k}] \cdot Y_i}{\sum_{j=1}^{n} [X_{j,k}=X_{i,k}]} \tag{1}$$

Indeed, the method of replacing features with the average value of labels has a drawback when dealing with categorical features, which contain more information than labels. If the average value of labels is forcibly used as a replacement for features, the model may encounter the

problem of conditional shift when dealing with unseen feature values, especially when the data structure and distribution of the training and testing datasets differ.

In CatBoost, to address this issue and improve the accuracy and stability of the algorithm, a common approach is to enhance the Greedy Target-based Statistics by adding a prior distribution term. This method reduces the influence of noise and low-frequency category data on the data distribution, enabling the model to better handle unseen feature values and maintain consistency with the prediction target, as shown in Equation 2.

$$\hat{X}_k^i = \frac{\sum_{j=1}^{p-1} \left[x_{\sigma_{J,k}} = x_{\sigma_{p,k}}\right] \cdot Y_{\sigma_j} + a \cdot p}{\sum_{j=1}^{p-1} \left[x_{\sigma_{J,k}} = x_{\sigma_{p,k}}\right] + a} \tag{2}$$

Where p is the added prior term, a is usually a weight coefficient greater than 0. Adding a prior term is a commonly used method that reduces noisy data for features with fewer categories and can improve the accuracy and stability of the model in specific cases. In regression problems, a commonly used value for the prior term can be the mean of the labels in the dataset. However, CatBoost differs from other models in that it adopts a novel method for calculating leaf node values, avoiding the overfitting issues caused by direct computation. With this approach, CatBoost can efficiently build models with high accuracy and strong robustness when dealing with large-scale and high-dimensional data.

3.2 Optuna Optimization

Optuna is a state-of-the-art automatic hyperparameter tuning framework entirely installed in Python. Optuna utilizes a runtime-defined API, allowing users to write highly modular code and dynamically construct the search space for hyperparameters. It employs various samplers, including grid search, random search, Bayesian search, and evolutionary algorithms, to find the optimal values for hyperparameters. The workflow of Optuna is illustrated in Figure 1.

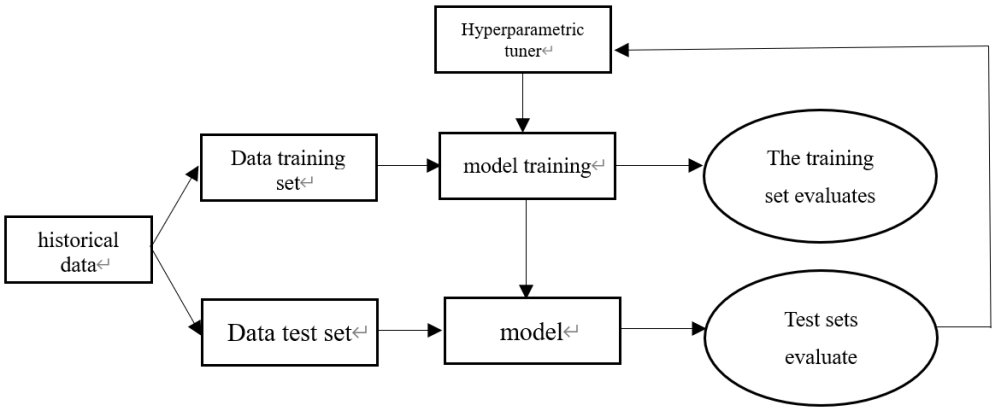


Fig. 1. Hyperparameter optimizer algorithm learning process

Optuna utilizes its own trial history to determine the next values to be tested within the defined search space. It employs a Bayesian optimization algorithm called "Tree-structured

Parzen Estimator." When operating, we introduce Bayes into the process, where *p(x|y)* represents the conditional probability of hyperparameters being x given that the model loss is *y* and following Equation 3 describes the detail calculation process.

$$p(x|y) = \begin{cases} l(x) & if y < y^* \\ g(x) & if y \geq y^* \end{cases} \tag{3}$$

4 Experimental Analysis

4.1 Experimental Setups

We use credit card data from a certain bank spanning from January 2015 to December 2022 as the sample for analysis. Throughout this process, the study encountered issues with data imbalance and insufficient error rates. Initially, in order to adjust the imbalance in the sample data, a combination of upsampling and downsampling was employed. Data was randomly selected from the sets of credit card samples with normal repayment and default repayment. This approach aimed to achieve a ratio of 2:1 between normal repayment and default samples, thereby mitigating the impact of data imbalance. This handling method resulted in high accuracy training of the model.

4.2 Statistical Description

Following Table 1 describe the statistical analysis of the continuity data. Additionally, WOE (Weight of Evidence) and IV (Information Value) are commonly used feature selection methods that help evaluate the impact of each variable on credit card default risk prediction. WOE is a method for assessing the strength of the relationship between each variable and the dependent variable. It calculates the distribution of each variable across different categories of the dependent variable. IV is a metric used to assess the predictive power of each variable, reflecting the degree of influence of each variable on the dependent variable. Through the analysis of WOE and IV, we can select the variables that have the greatest impact on credit card default risk prediction, thereby constructing the optimal predictive model. The specific formulas for calculating WOE and IV are shown in Equation 4 and 5, respectively.

Table 1. Statistics of sample continuity variables.

Index	Mean value	Median
Number of card transactions in the last 7 days (single)	1.22	1.64
Share of credit card transactions in the past 7 days ($ten thousand)	0.062	0.077
Number of card transactions in the last 30 days (single)	3.11	1.78
Share of credit card transactions in the past 30 days ($ten thousand)	0.56	0.43
Number of card transactions in the last 1 year (single)	2.84	3.62
Share of credit card transactions in the past 1 year ($ten thousand)	0.22	0.41
Number of installments in the past year	2.12	1.59
Credit line ($ten thousand)	0.71	0.55
Interest free period	16.23	19.45

rate of interest	0.03	0.03
Last repayment date	12.21	14.72
Age	25.77	30.24
Number of history overdue	0.18	0.22
Financial assets ($ten thousand)	0.77	0.89
Data integrity	0.71	0.75

$$WOE_i = ln\frac{G_i/G}{B_i/B}, \quad i = 1,2,3 \dots \dots n \tag{4}$$

$$IV = \sum_{1=1}^{\eta} \left[\frac{G_i}{G} - \frac{B_i}{B}\right] WOE_i \tag{5}$$

The ROC curve of the model is illustrated in Figure 2. From Figure 2, we can observe that the CatBoost model exhibits a favorable shape. By utilizing the functions available in the matplotlib, the AUC value of the CatBoost model can be calculated as 0.909.

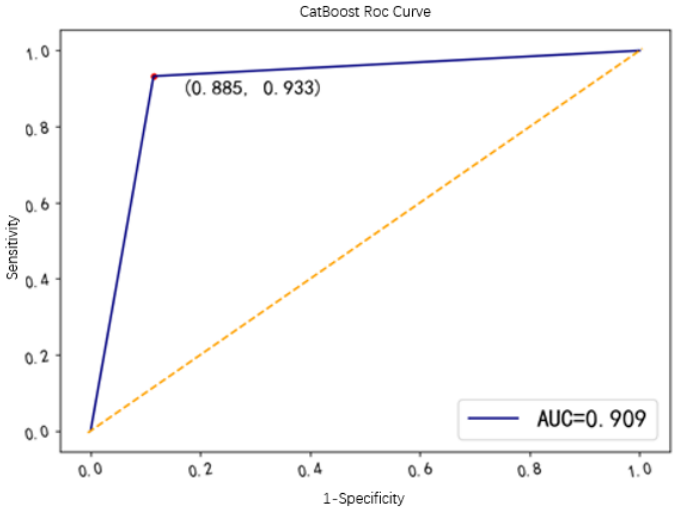


Fig. 2. Hyperparameter optimizer algorithm learning process

4.3 Comparison Analysis

Following Table 2 describes the prediction results of CatBoost model.

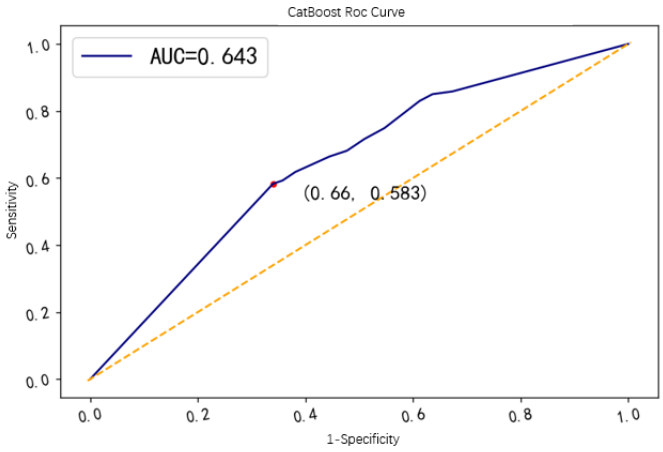
Table 2. CatBoost sample test comparison results

Ours-CatBoost		Predicted value	
		Normal reimbursement	Late payment
Actual value	Normal reimbursement	1422	93
	Late payment	78	702
SVM		Predicted value	
		Normal reimbursement	Late payment
Actual value	Normal reimbursement	1825	375
	Late payment	206	894

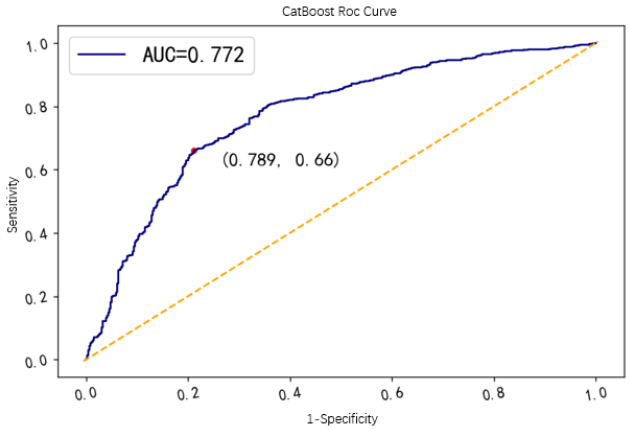
BP		Predicted value	
		Normal reimbursement	Late payment
Actual value	Normal reimbursement	1763	437
	Late payment	235	865

From above Table 2, we can observe that the BP neural network model used in this study achieves an accuracy of 77.33% in predicting default cases and an accuracy of 82.9% in predicting normal repayment cases. The support vector machine (SVM) model achieves an accuracy of 81.2% in predicting default cases and an accuracy of 82.9% in predicting normal repayment cases. The prediction results of CatBoost model are showing the best performance with 92.4% prediction accuracy and 94.0% for default cases.

The ROC curve of the BP neural network model and SVM model on the credit card test set is shown in Figure 3.



(a)SVM



(b)BP

Fig. 3. ROC curves comparison results.

As shown in Figure 3, the results obtained from the test set used in this study reveal an AUC of 0.643 and 0.772. This indicates that the SVM and BP modelS indeed possesses good predictive capabilities for a significant portion of credit card default cases. However, when compared to the previous CatBoost model mentioned earlier, it falls slightly short.

5 Conclusion

In conclusion, the CatBoost model presents a potent tool in the arsenal of credit risk prediction techniques. Its adept handling of categorical variables, resistance to overfitting, and competitive performance metrics position it as a promising model for credit card overdue risk prediction. Future research might further explore its integration with other machine learning models or its application in other financial risk prediction scenarios.

References

[1] Raj, S. Benson Edwin, and A. Annie Portia: Analysis on credit card fraud detection methods. 2011 International Conference on Computer, Communication and Electrical Technology. IEEE, (2011).

[2] Ma, Yuhan: Prediction of default probability of credit-card bills. Open Journal of Business and Management 8.01 (2019): 231.

[3] Lin, Liqiong, et al: Determinants of credit card spending and debt of Chinese consumers. International Journal of Bank Marketing 37.2 (2019): 545-564.

[4] Wang, Haifeng, and Dejin Hu: Comparison of SVM and LS-SVM for regression. 2005 International conference on neural networks and brain. Vol. 1. IEEE, (2005).

[5] Kurani, Akshit, et al: A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science 10.1 (2023): 183-208.

[6] Bo, Yin, et al: Real-time hard-rock tunnel prediction model for rock mass classification using CatBoost integrated with Sequential Model-Based Optimization. Tunnelling and underground space technology 124 (2022): 104448.

[7] Li, Jing, et al: Brief introduction of back propagation (BP) neural network algorithm and its improvement. Advances in Computer Science and Information Engineering: Volume 2. Springer Berlin Heidelberg, (2012).

[8] Zhang, Jing-Ru, et al: A hybrid particle swarm optimization–back-propagation algorithm for feedforward neural network training. Applied mathematics and computation 185.2 (2007): 1026-1037.