Financial Engineering and Risk Management (2021) 4: 60-68
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2021.040104
ISSN 2523-2576

Research on Default Rate of Financing Projects of Online Lending Platform Based on XGBoost Model

Xigan Sun

School of Aeronautical Science and Engineering, Beihang University, Beijing 100000, China

Keywords: Default Rate, Financing Projects, Online Lending, XGBoost Model

Abstract: In recent years, with the rapid development of the online credit industry and the wide application of big data technology, using an integrated learning model to evaluate loan risk quickly and accurately has been a concern by academics and practitioners. In order to predict the default rate of the financing projects of the online loan platform with high accuracy and efficiency, this paper adopts the XGBoost model based on the importance of certain features to process loan application data of an online loan platform and establishes the default rate prediction model of online loan projects. Ten years' loan application data of American online lending platforms were selected to verify the model, and the prediction results were compared with those of Random Forest (RF) and LightGBM. The results show that the XGBoost model based on the optimization derivation and the second-order Taylor expansion has higher accuracy in the evaluation.

1. Introduction

With the rapid development of the economy and the industrial model's continuous innovation, the emerging P2P financing platforms have shown explosive development. Consequently, small enterprises, individual entrepreneurs, and start-up companies have more convenient and rapid online financing channels [1]. Efficient and accurate project evaluation has become an important guarantee to promote the rapid landing of high-quality projects and ensure the stable profitability of lending companies [2].

However, in the actual operation of financing platforms and commercial banks, the default prediction and risk identification of applied lending projects have always been a core issue [3]. In the process of continuous development of the Internet personal credit platform, personal credit default events also occur frequently, which brings hidden dangers and inconvenience to the platform's regular operation and other projects' application in the future. Simultaneously, the complexity of the application project itself and the advent of the era of big data have brought significant challenges to the traditional default prediction [4]. Therefore, it is of great significance to study how to deal with project data with complex characteristics and identify default risk to ensure the platform income and the stable development of the lending industry.

In the early stage, Xiao Wenbing et al. [5] used the support vector machine's evaluation model to cross-verify credit evaluation problems and explored the optimal kernel function. However, it was challenging to deal with complex data. Afterward, G B Fernandes [6] adopted the traditional logistic regression to assess the projects' credit risk in the data set. However, due to the research results' lack

of accuracy, the prediction result is not satisfactory. Fan Yanqin [7] conducted two Bayesian classification models to predict personal credit rating, and the research showed that this method had a less misclassification. However, when processing loan data with large dimensions and uneven distribution, this model's prediction effect needs to be improved. Li Jin [8] used the random forest model to conduct an example analysis on a listed company's user credit data and achieved a relatively ideal accuracy rate, but the processing effect for the data containing much noise was not good.

Compared with the logistic regression and random forest models used in the above studies, the XGBoost algorithm can automatically utilize multi-threads of CPU to carry out distributed learning and multi-core computation and improve the computational efficiency of guaranteeing classification accuracy, which is suitable for large-scale processing data. On the other hand, the XGBoost algorithm has achieved good results in text data processing in recent years and has shown a strong predictive ability for multivariate data. Given this, this paper adopts the XGBoost model based on specific characteristic importance to process the loan application data variables of an online lending platform and predict the default rate. On the basis of extracting the importance of model features, the accuracy of credit prediction came from three models, including Random Forest, LightGBM, and XGBoost were compared. It is proved that the XGBoost model has better practical significance in processing multivariable data and predicting the default rate of the projects applied by online lending platform.

2. Methodology

XGBoost is a Gradient Boosting Decision Tree (GBDT) model, a kind of ensemble learning algorithm belonging to the Boosting algorithm category of the three commonly used ensemble methods (Bagging, Boosting, and Stacking). XGBoost model has more parameters and better overall robustness. In general, the ideal results are obtained by adjusting the depth of the tree, the minimum leaf node sample weight, and the L2 regularization coefficient. At the same time, bagging thought is adopted to fully train the normalized features, discrete features, and combined features to get the predicted results. After weighted fusion, the result of model training can be obtained.

Ensemble learning is about generating better and different individual learners. Higher accuracy and diversity of individual learners bring about a better integration effect. Compared with stable classification models such as LR and NB, the tree model is an unstable classification model that is more sensitive to sample disturbance. Decision trees are often used as individual learners of ensemble learning because of their simplicity, intuitiveness, and strong interpretability [9].

Based on the advantages of ensemble learning, the XGBoost algorithm build on the CART tree was adopted to establish a score prediction model at the bottom of the model. Meanwhile, to increase the diversity of individual learners in ensemble learning and improve the model's generalization ability, a bagging idea is adopted to add data sample perturbation and attribute perturbation in the top layer to establish several good and different XGBoost models. Meanwhile, the voting method is adopted to integrate the models.

Assume that a data set of n with m features is $D = \{(x_i, y_i)\}(i = 1,2, \dots, n)$. The set containing all Cart trees is $F = \{f(x) = w_{q(x)}, q: R^m \to T, w \in R^T\}$. Where Q represents the decision rule that the sample is mapped to the corresponding leaf nodes, T represents the number of leaf nodes of a tree, and W represents the score of leaf nodes. The predicted value of y_i based on the XGBoost algorithm can be expressed as follows [10].

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

Where K is the number of CART trees and $f_k \in F$.

In each model training, the XGBoost algorithm keeps the prediction of the previous $t-1$ round unchanged and adds a new function f_t to the model. $\hat{y}_i^t = \hat{y}_i^{(t-1)} + f_t(x_i)$ is the prediction result

of the ith sample in the tth model training. Assuming that the errors of the base learners are independent of each other, the learning goal of the XGBoost algorithm is to find the f_t minimization objective function, whose calculations are shown in Equations (2) and (3) [11].

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{2}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

Where, $l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right)$ is the loss function is the loss caused by the difference between the predicted value and the real value. $\Omega(f_t)$ is the regular optimization term of model complexity, which is used to reduce the complexity of the model. γ is the complexity parameter, and λ is a fixed coefficient.

XGBoost algorithm conducts a greedy algorithm to recursively select the tree structure's optimal features starting from the root node and segment the training data according to the features [8]. Assume that I_L and I_R are sample sets to the left and right of the split point, respectively when $I = I_L \cup I_R$. The information gain of each segmentation scheme is calculated. The segmentation with the most considerable information gain is the node's optimal segmentation, and its calculation is shown in equation (4) [12].

$$L_{\text{split}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i+\lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i+\lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i+\lambda}\right] - \gamma \tag{4}$$

Where, $I_j = \{i \mid q(x_i) = j\}$ is the sample set on node j. $g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right), h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$ are the first order and second-order gradient statistics of the training error, respectively. $\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i+\lambda}, \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i+\lambda}, \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i+\lambda}$ are left subtree fraction, right subtree fraction, and undivided fraction, respectively. γ is the complexity cost of adding new leaf nodes, and the segmentation is abandoned when $L_{\text{split}} < 0$. The second-order Taylor expansion of the loss function is performed at the $\hat{y}_i^{(t-1)}$ position to speed up the optimization process, as shown in Equations (5) [13].

$$L^{(t)} = \sum_{j=1}^{T}\left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{5}$$

3. Model

3.1 Data preprocessing

The data selected in this experiment came from the loan data from 2005 to 2014 of Prosper, the first P2P online loan platform in the United States. This data set has several characteristics that reflect the creditworthiness of loan users. Prosper Rating is a parameter set by Prosper according to its model, which is the primary basis for determining the loan interest rate. Official Credit rating agencies provide credit scores. Loan state is classified as Cancelled, Charged off, Completed, Current, Defaulted, Final Payment in Progress, and Past Due. Data variables are shown in the following table 1.

Table 1 Data variables and descriptions

Variable	Description
ListingKey	Unique key for each list. Using the same value as the key used in the list object in the API
ListingNumber	A public number displayed on a website that uniquely identifies the list
ListingCreationDate	Trading start time
CreditGrade	Reflects credit ratings of customers prior to July 1, 2009. The higher the credit rating, the stronger the solvency
Term	The loan cycle of loan applying
LoanStatus	The state of the loan. Including Completed, Current, Defaulted, Chargedoff
BorrowerAPR	Lender's annual interest rate
BorrowerRate	Standard rate on borrowing. As a proxy variable for the price of borrowing funds from P2P platforms, BorrowerRate is the payment that the financier pays to the investors and is the most direct and important cost of financing. It reflects the capital supply and demand of the two sides in the comprehensive consideration of a variety of factors recognized by the use of capital costs

This experiment's primary purpose is to build a loan default prediction model to ensure the loan platform's stable profit and the smooth progress of the loan application. Therefore, some variables of data need to be preprocessed in this paper. Firstly, remove extraneous variables, including redundant numbered columns (Listing Key, Listing Number, Loan Number) and investor-only columns (LP_Customer Payments, LP_Service Fees, etc.,). Then, remove variables with duplicated meanings. In the data set, Bank APR and Bank Rate are directly related. As a result, only Bank APR will be analyzed. Prosper Rating (numeric) and Prosper Rating (Alpha) are different expressions of the same meaning. Therefore, only the variable Prosper Rating (Alpha) is selected. Credit Scorer Angel Power is related to Credit Scorer Angel Power.

Consequently, only the variable Credit Scorer Angel Power is selected. Meanwhile, Prosper has adjusted the way it evaluates its customers since July 2009. As a result, this research will only analyze loans application after July 1, 2009.

Because the original data has some missing values, it is necessary to deal with the packets' missing values. The missing variable data of the original data is shown in Table 2. First of all, variables with too high a missing rate, including Closed Date, Group Key, Loan First Defaulted Cycle Number, are deleted in this paper. Then, replace the 1307 missing entries in the Occupation column with 'Other'. Next, the 15 missing items in the Employment Status Duration column are simply deleted. Finally, for the Debt Toin Comer Ratio column, the values are assigned randomly between 0 and 0.5 because there are many missing data, and most of the values are below 0.5.

Table 2 Data missing value

	Whether the missing	Quantity	Number of missing	Missing rate
ClosedDate	True	16507	47302	0.741306
Occupation	True	62502	1307	0.020483
EmploymentStatusDuration	True	63794	15	0.000235
GroupKey	True	302	63507	0.995267
DebtToIncomeRatio	True	58724	5085	0.079691
LoanFirstDefaultedCycleNumber	True	4321	59488	0.932282

To make the data set more comfortable to process in this experiment, the variables Prosper score, Credit Scorerangelower. Employment Status Duration is converted to int format. Whether the transaction is still in progress and whether the investor has lost money in the closed transaction, this paper divides all data into three groups: Current, Completed, and Completed, and sets the variable

"Completed" to 1 and the variable "Defaulted" to 0.

3.2 XGBoost model design

In this paper, a prediction model is established by studying the selected data sets combining the XGBoost ensemble learning model. Firstly, K regression trees are built by the XGBoost model. The target function with as few leaf nodes as possible is used to train the model for higher accuracy and better generalization. The XGBoost model is used to resume K regression trees, and the target function with as few leaf nodes as possible is used to train the model with high accuracy, better generalization, and small prediction error. Greedy strategy and quadratic optimization were used to determine the optimal node and the minimum loss function, based on which tree splitting was carried out continuously. The optimal tree model is built at each split, and the iteration stops when MAX_DEPTH is reached.

Finally, this paper adopts Python based on Numpy and Pandas environment to construct the XGBoost loan platform application project default rate prediction model. The algorithm flow of the model is shown in Figure 1 below.

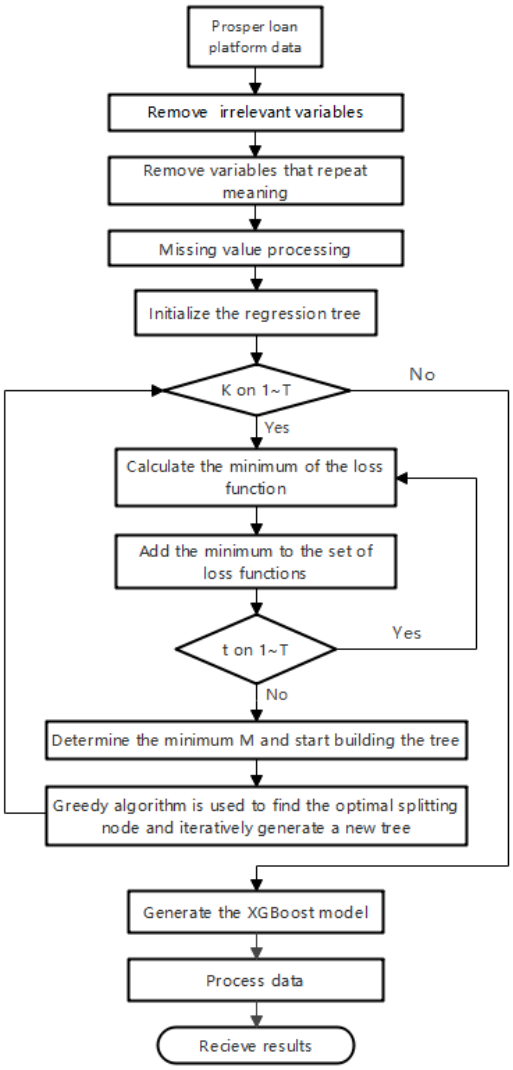


Figure 1. Model framework

4. Experimental results and analyze

4.1 Experimental results

Through many experimental data and eliminating the irrelevant variables in the experimental process, 70% of them were selected as the training sample set and 30% as the testing sample set. After the training based on the sample set, the final feature importance chart is extracted, as shown in figure 2. The top four features are ProsperRating, Incomerange, StatemonthlyIncome, and CurrentIntention.

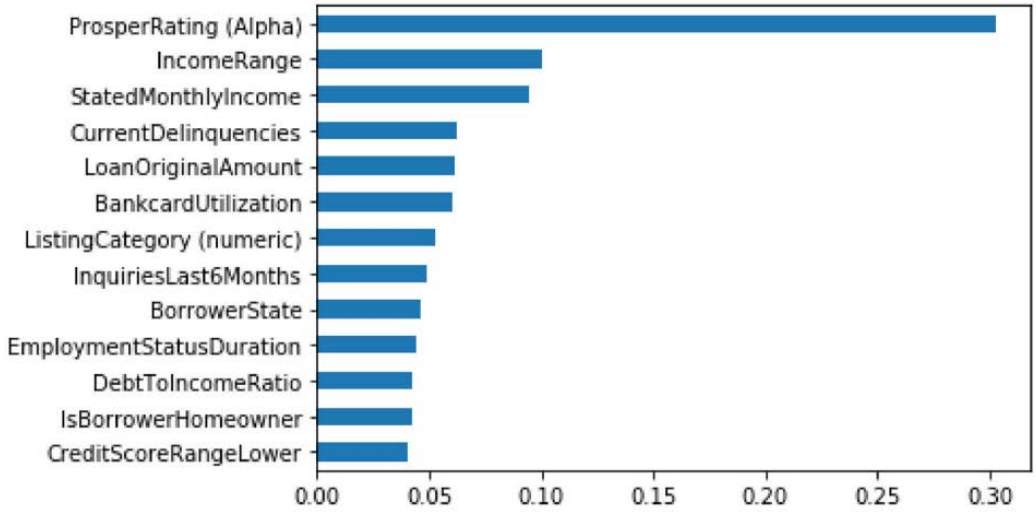


Figure 2. Feature importance ranking

These four characteristics have the most significant impact on whether lenders ultimately default. Therefore, the platform should focus on these characteristics when processing loan applications.

To verify the prediction performance of the XGBoost model used in this paper under unbalanced data, the ROC curve is introduced. The full name of ROC is the Receiver Operating Characteristic curve. Firstly, each sample is predicted as a positive example. Secondly, the threshold is changed from 0 to maximum based on the XGBoost model's results. As the threshold increases, the number of positive samples predicted by the model gets smaller and smaller until there is no positive sample. In this process, two important values are calculated each time. The ROC curve of the XGBoost model is established respectively on the vertical axis with the True Positive Rate (TPR) and the horizontal axis with the False Positive Rate (FPR). This is shown in Figure 3. From this figure, the influence of arbitrary thresholds on the generalization performance of the XGBoost model can be visually detected. At the same time, the AUC value of the XGBoost model is close to 0.7, showing good generalization performance. Therefore, it shows that the XGBoost model has good practical significance in predicting loan default rates.

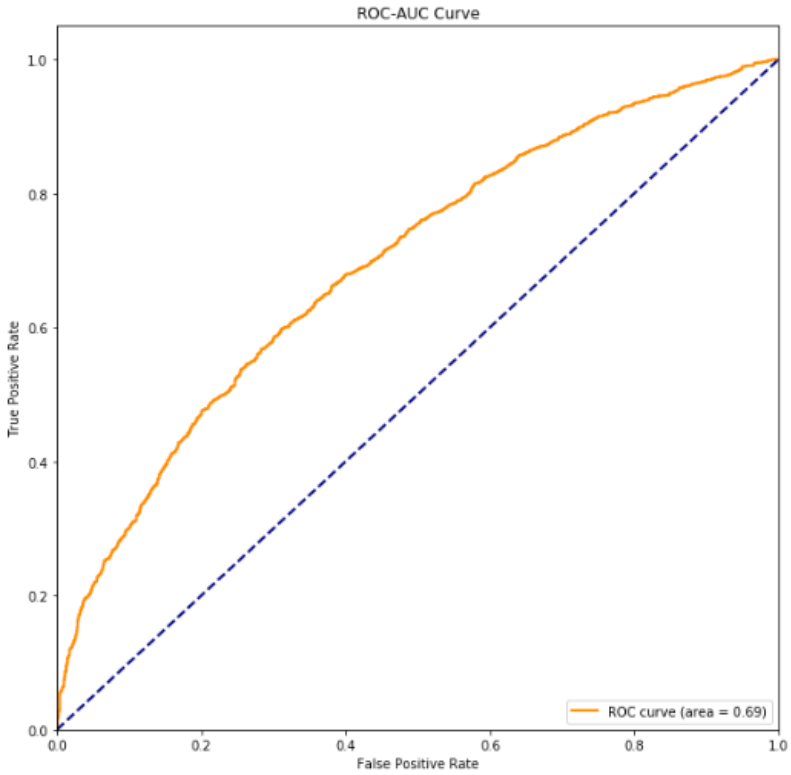


Figure 3. The ROC-AUC curve of XGBoost

In the case of category imbalance, the PR curve is widely considered superior to the ROC curve because the data of positive cases is mainly concerned. Since the selected data is of a particular imbalance, this paper draws a PR curve to study further the relationship between the accuracy and the recall rate of the Xgboost model, as shown in Figure 4. PR curve shows the relationship between Precision and Recall, representing the proportion of samples that are predicted to be positive examples and are positive examples. The similarity between the PR curve and ROC curve is that TPR (Recall) is adopted for both of them, and AUC can be used to measure the classifier's effectiveness. The difference is that the ROC curve uses FPR, while the PR curve uses precision, so both PR curve indicators focus on the positive examples. The PR curve's horizontal axis is the Recall rate, and the vertical axis is the precision rate.

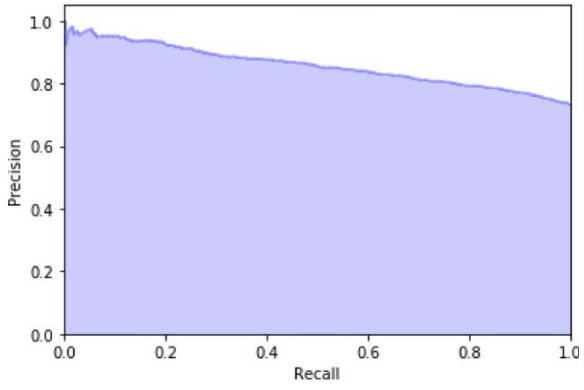
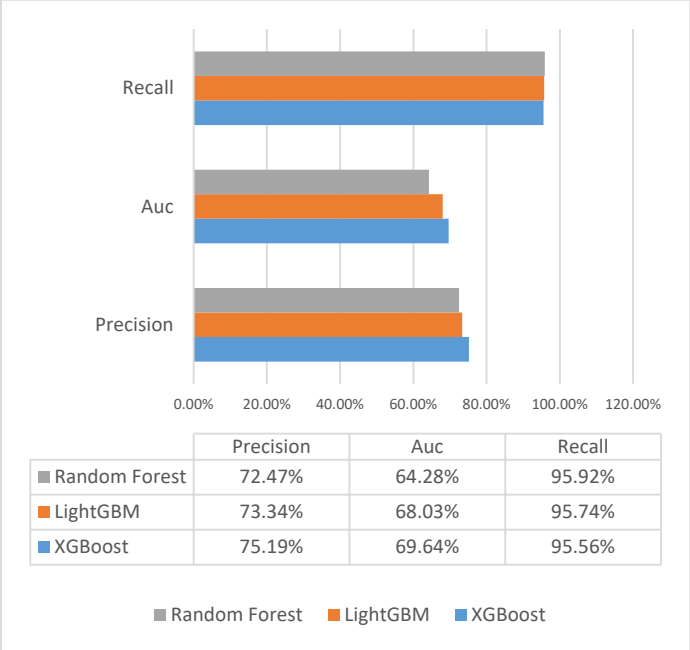


Figure 4. 2-class Precision-Recall curve

Compared with the ROC curve statistics, the PR curve provides another aspect of the XGBoost model's prediction effect in the category imbalance problem. By studying the PR curve of the XGBoost model in this application scenario, it can be found that the accuracy of the model fluctuates between 0.7 and 1.0. Compared with the ROC curve's accuracy, the improved accuracy shown by the PR curve shows that the XGBoost model has a good performance in dealing with the unbalanced data of the loan platform.

4.2 Comparative experiment

To verify the superiority of the method, the random forest algorithm and LightGBM model were selected to predict the project default rate on the selected data set, and the prediction results were compared with those of the XGBoost model. The experimental results of each model are shown in figure 5.



	Precision	Auc	Recall
Random Forest	72.47%	64.28%	95.92%
LightGBM	73.34%	68.03%	95.74%
XGBoost	75.19%	69.64%	95.56%

Figure 5. Comparative result

According to the experimental results, although the three groups of models' recall rates were similar, the XGBoost algorithm's precision reached 75.19%. This value is about 2.7 percentage points higher than the prediction accuracy of the random forest algorithm (72.47%) and 1.85 percentage points higher than the prediction accuracy of the LightGBM model. AUC values of XGBoost were 5.36% and 1.61% higher than those of random forest and LightGBM, respectively. Because XGBoost uses CART as the base classifier, it adds regular terms to control the model's complexity, supports data sampling, and can automatically learn the processing strategy of missing values. Therefore, combining the above analysis and experimental data, the model shows better robustness and higher accuracy in the experimental process.

5. Conclusions

This paper proposes a credit risk assessment method for an online lending platform based on the XGBoost model. In order to predict the default rate of financing projects of online loan platforms

with high precision and accuracy, this paper selects the XGBoost algorithm based on CART tree learners to establish a scoring prediction model. The model is verified using the data of the American loan platform Prosper. The experimental results show that the XGBoost model based on the optimal derivation and second-order Taylor expansion has good practical significance in predicting default rate. Compared with the random forest and LightGBM model, XGBoost has higher precision and accuracy. We will continue to search for the parameter optimization and model fusion methods of XGBoost algorithm integration in future research. It is used to further improve the classification accuracy and computational efficiency of XGBoost algorithm integration and improve its application's user rating prediction model.

References

[1]. CHEN Qiu-hua, YANG Hui-rong, CUI Heng-jian. Personal Credit Scoring Model and Statistical Learning after Variable Selection [J]. Journal of Applied Statistics and Management, 2020, 39 (02): 368-380.

[2]. LENG Aolin, XING Guangyuan, FAN Weiguo. Credit Risk Transfer in SME Loan Guarantee Networks [J]. Journal of Systems Science & Complexity, 2017, 30 (05): 1084-1096.

[3]. LIU Xuefeng, ZHANG Wei, XIONG Xiong, SHEN Dehua, ZHANG Yongjie. Credit Rationing and the Simulation of Bank-Small and Medium Sized Firm Artificial Credit Market [J]. Journal of Systems Science & Complexity, 2016, 29(04): 991-1017.

[4]. PRAGER David, ZHANG Qing. Valuation of Stock Loans under a Markov Chain Model [J]. Journal of Systems Science & Complexity, 2016, 29 (01): 171-186.

[5]. Xiao Wenbing, Fei Qi, Wan Hu. Credit scoring models and credit-risk evaluation based on support vector machines [J]. J. Huazhong Univ. of Sci. & Tech. (Nature Science Edition), 2007 (05): 23-26.

[6]. Guilherme Barreto Fernandes, Rinaldo Artes. Spatial dependence in credit risk and its improvement in credit scoring. 2016, 249 (2): 517-524.

[7]. Fan Yanqin, Qin Yangsen, Yuan Yuan. Application of Bayesian network based on principal component analysis in personal credit evaluation [J]. Journal of Guilin University of Aerospace Technology, 2019, 24 (04): 568-575.

[8]. Li Jin. Research on credit risk assessment of green credit based on random forest algorithm [J]. Financial theory and practice, 2015 (11): 14-18.

[9]. Breiman L I, Friedman J H, Olshen R A, et al. Classification and Regression Trees (CART) [J]. Encyclopedia of Ecology, 1984, 40 (3): 582-588.

[10]. Yang Guijun, Xu Xue, Zhao Fuqiang. Predicting User Ratings with XGBoost Algorithm [J]. Data analysis and knowledge discovery, 2019, 3 (01): 118-126.

[11]. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.

[12]. Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine [J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[13]. BAI Pengfei, AN Qi, Nicolaas Fransde ROOIJ, LI Nan, ZHOU Guofu. Internet Credit Personal Credit Assessing Method Based on Multi—Model Ensemble [J]. Journal of South China Normal University (Natural Science Edition), 2017, 49(06): 119-123.

[14]. Breiman L. Random Forests [J]. Machine Learning, 2001, 45(1): 5-32.

[15]. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 785-794.

[16]. Zhang R, Gao Y, Yu W, et al. Review Comment Analysis for Predicting Ratings[A]// Web-Age Information Management [M]. Springer, 2015: 247-259.

[17]. Chen T, Guestrin C., XGBoost: A Scalable Tree Boosting System [C] //Acm Sigkdd International Conference on Knowledge Discovery & Data Mining.2016.

[18]. CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system [C] //Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.

[19]. BREIMAN L. Random forests [J]. Machine Learning, 2001, 45: 5-32.