Hindawi
Complexity
Volume 2024, Article ID 9791631, 1 page
https://doi.org/10.1155/2024/9791631

Retraction

Retracted: Big Data Analytics for Complex Credit Risk Assessment of Network Lending Based on SMOTE Algorithm

Complexity

Received 23 January 2024; Accepted 23 January 2024; Published 24 January 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] A. Niu, B. Cai, and S. Cai, "Big Data Analytics for Complex Credit Risk Assessment of Network Lending Based on SMOTE Algorithm," *Complexity*, vol. 2020, Article ID 8563030, 9 pages, 2020.

Hindawi
Complexity
Volume 2020, Article ID 8563030, 9 pages
https://doi.org/10.1155/2020/8563030

Research Article

Big Data Analytics for Complex Credit Risk Assessment of Network Lending Based on SMOTE Algorithm

Aiwen Niu,[1] Bingqing Cai,[2] and Shousong Cai [3]

[1]*Glorious Sun School of Business Management, Donghua University, Shanghai 200051, China*
[2]*School of Humanities, Shanghai University of Finance and Economics, Shanghai 200433, China*
[3]*School of Business Administration, Shanghai Lixin University of Accounting and Finance, Shanghai 201209, China*

Correspondence should be addressed to Shousong Cai; caishousong@126.com

Received 8 July 2020; Accepted 4 August 2020; Published 26 September 2020

Guest Editor: Zhihan Lv

With the continuous development of big data technology, the data of online lending platform witness explosive development. How to give full play to the advantages of data, establish a credit risk assessment model, and realize the effective control of platform credit risk have become the focus of online lending platform. In view of the fact that the network loan data are mainly unbalanced data, the smote algorithm is helpful to optimize the model and improve the evaluation performance of the model. Relevant research shows that stochastic forest model has higher applicability in credit risk assessment, and cart, ANN, C4.5, and other algorithms are also widely used. In the influencing factors of credit evaluation, the weight of the applicant's enterprise scale, working years, historical records, credit score, and other indicators is relatively high, while the index weight of marriage and housing/car production (loan) is relatively low.

1. Introduction

In recent years, with the rapid development of Internet finance, P2P platform develops rapidly in this context and gradually forms a new financial platform with great influence. P2P platform relies on cloud computing, social networking, and other channels to collect, organize, and record data, which greatly enhances the ability of financial risk prevention and control based on data mining technology. Through comparative analysis of user information, combining with specific historical data, it can effectively improve the information flow efficiency of both sides of capital supply and demand and provide necessary support for both sides to establish the relationship between supply and demand, and based on this, the financial risk caused by information asymmetry is kept at the lowest level [1]. However, the domestic P2P platform research time is relatively short, has not yet established a sound credit system, coupled with the relevant legal system is not perfect, and is easy to induce credit risk, which is a serious threat to the safety of users' funds. In addition, with the advent of the era of big data, the data of online lending platform are constantly increasing, the data types are diverse, and the data are updated quickly [2]. How to give full play to the data advantages, obtain the required information, and enhance the platform's ability to monitor capital risk factors become the key to the development of the platform. In this case, the platform needs to rely on big data, combined with data mining technology, to build a scientific and reasonable credit risk assessment model to provide necessary basis for platform risk supervision and user investment. Therefore, this study is of great significance on the practical level.

Relatively speaking, the development time of foreign P2P platform is relatively long, and the level of related research results is high, which has more reference value in the aspects of concept discussion and risk assessment [3]. In the aspect of loan success rate, the relationship between personal information and loan success rate has been studied, and then the borrowing strategies are comprehensively discussed with the help of quantitative analysis tools. In terms of credit risk,

domestic scholars use empirical analysis and specific cases to analyze the influencing factors of credit risk and summarize the influencing factors related to default behavior [4]. Based on this, a classification method with random forest as the core is constructed, which greatly improves the effectiveness of credit risk assessment [5]. Affected by the Internet, computer, information technology, etc., building a smart city has become a key task of socialist construction work [6]. The empirical analysis shows that compared with FICO or LC, the evaluation method based on random forest has more advantages in identifying high reputation borrowers [7]. The research results in recent years show that the role of social network cannot be ignored in the development of online lending. The higher the richness of social resources, the lower the cost of obtaining loans [8]. There is a significant negative correlation between the two [9]. Empirical analysis shows that through the analysis of applicants' social networks, we can deeply understand the soft information related to credit risk so as to evaluate the applicants' credit risk more comprehensively. In the dual-channel supply chain system, channel optimization is influenced by channel attitude toward risk, in which risk is classified as general risk and interruption risk [10]. For individuals, P2P platform provides convenience for its financing or capital problems but also produces a series of risk problems, such as imperfect credit system, high moral hazard, and serious adverse selection [11]. At present, credit risk is the key content of risk research, and the research direction includes default characteristics analysis and platform reputation [12]. As the number of selected lines increases, the current same price for all passengers in different riding paths could make the bus industry development a step further [13]. In terms of default probability of applicants, incomplete market-oriented interest rate has more significant prediction effect, but the use of personal public information can also reflect the default risk to a certain extent [14]. Taking Renrendai as an example, its credit certification mechanism has certain advantages in reflecting credit risk, but the index system has certain limitations, so it is necessary to supplement and improve its evaluation index system.

Overall, the depth of research on P2P network lending needs to be further expanded. Compared with western developed countries, China's P2P platform development time is relatively short and mainly concentrated after the rise of Internet finance in 2012. The empirical research data are insufficient, mainly referring to the data provided by foreign platforms. However, both the research methods and the research conclusions are difficult to fully meet the domestic research demand. In view of this situation, this paper introduces R language and python to write web crawler program in data crawling of online credit platform, introduces smote algorithm in unbalanced data processing, and constructs credit risk assessment model combined with six data mining algorithms, which is more consistent with the development of domestic P2P platform. It is a kind of research on network credit risk based on big data background, and the new ideas play a positive role in improving the level of theoretical research in China.

2. Model Description

2.1. Commonly Used Data Mining Classification Model. In the field of classification technology, decision tree presents the classification process in the form of directed acyclic tree, which is intuitive and simple, so it has a high application rate [15]. For classified data, the greedy algorithm is used as the core of decision tree to determine the nodes, and then the local optimal decision strategy is used to construct the decision tree. In the dual-channel supply chain system, channel optimization is influenced by channel attitude toward risk, in which risk is classified as general risk and interruption risk [16]. There are significant differences in decision tree types with different classification criteria. For example, taking information theory as the standard, it can be divided into ID3, C4.5, and cart, and SLIQ and sprint can be obtained from Gini index. Among the above methods, only ID3 can be used for discrete variables. On the basis of comprehensive analysis, cart and C4.5 algorithms are selected in this paper.

AdaBoost algorithm is a kind of lifting algorithm which can adjust the distribution of training samples by itself. It has high adaptive ability to ensure that the base classifier fits the samples with higher difficulty in classification. Through the AdaBoost algorithm, the weights of training samples can be combined, the parameters can be updated, and then the corresponding weighting can be completed:

$$\omega_i^{(j+1)} = \frac{\omega_i^{(j)}}{Z_j} \times \begin{cases} e^{-\alpha_j}, C_j(X_i) = y_i, \\ e^{\alpha_j}, C_j(X_i) \neq y_i, \end{cases} \tag{1}$$

where $\omega_i^{(j)}$ refers to the weight of the sample (X_i, y_i) in the round of j iteration. Using this weight can enhance the weight of the wrong classification samples to a certain extent, which is not conducive to highlighting the weight of the correct classification samples [17]. Therefore, for unbalanced data sets, this algorithm can improve the accuracy of minority prediction to the greatest extent, and its defect is that the fitting problem is more prominent.

Support vector machine (SVM) is a realization method based on statistical learning theory. This method relies on the Mercer theorem and combines with nonlinear mapping method to realize the effective mapping of feature space in the Hilbert space and realize the accurate division of samples according to the linear decision boundary [18]. The application fields of this method include nonlinear regression model, high-level data analysis, and sample classification.

Artificial neural network (ANN) is a method to analyze the law of things by imitating the organizational structure of biological neural network. It is based on a large number of nodes with connection relationship, which can realize continuous iteration by connecting different nodes. The online-to-offline (O2O) business model is the new online shopping model in which consumers purchase products or services online and get the products or services in offline physical stores [19]. In this process, we need to determine the weight of the previous iteration, then calculate the weight of the node, and update the weight with the error value. Through the repetition of the above process, the error is reduced to the allowable range. Practice

shows that neural network is suitable for sample classification and variable regression and has good application effect. However, due to the high sensitivity of this method to noise, it is prone to local minimum problem, which has a certain negative impact on the accuracy of the final results.

2.2. Random Forests. Random forest is a combined classifier algorithm with decision tree as its core. In this method, the cart algorithm is used to construct the decision tree, the decision tree is used as the metaclassifier of sample classification, and the corresponding training set is obtained. In the construction of a single decision tree, the corresponding variables can be determined randomly, and node splitting can be completed based on the vector. According to the characteristics of this method, random forest has high robustness to noise, but low sensitivity to multiple reproducibility, so it can be relatively robust to deal with nonequilibrium data and get reasonable results.

The core of random forest is tree classifier, which is composed of various types of classifiers $\{h(x, \theta_k),\ k = 1, 2, \ldots, n\}$. Among them, the nonconstructed classification decision tree obtained by using the cart algorithm is the metaclassifier. According to the simple arithmetic average of single decision tree and majority voting output results, the accurate result data can be achieved, and the steps are as follows.

Firstly, the training sample set is constructed. In general, self-help resampling technology can be used to generate independent sample sets; that is, based on n sample sets, k new organizational sample sets are obtained by random return, and then the corresponding decision tree is formed, while the unselected samples constitute out-of-bag data, namely, OOB.

Secondly, the decision tree node is split. According to the overall situation of decision tree characteristic variables, assume m and then randomly determine m characteristic variables from them to split the corresponding nodes. Among them, the number of characteristic variables randomly obtained by each node is less than the number of assumed characteristic variables, and the corresponding splitting is carried out according to the principle of node impure minimization. It should be emphasized that all decision trees have no pruning operation.

Thirdly, the decision tree completes the corresponding combination. Based on the decision trees obtained in the above steps, the output results are determined by averaging all decision trees by majority voting, and then the error analysis stage is entered.

For the data of nontraining set, the possibility of error classification by a specific classifier is the generalization error. Theoretical research shows that if the number of decision trees reaches a certain degree, the upper bound of random forest generalization error will converge according to the law of large numbers. Under the premise of the given sample, the interval function provided by using the random forest is as follows:

$$\mathrm{mr}(x, y) = P_{\Theta}\left(h\left(x, \theta_k\right) = y\right) - \max_{j \neq y} P_{\Theta}\left(h\left(x, \theta_k\right) = j\right). \tag{2}$$

The strength of classifier set $\{h(x, \theta)\}$ can be expressed as follows:

$$s = E_{X,Y} mr(x, y). \tag{3}$$

According to the above expression, there is a positive correlation between the strength of the classifier set and the value of the interval function, that is, the strength of the classifier set increases with the increase of the value of the interval function, and the prediction accuracy will also be improved accordingly:

$$\mathrm{PE}^{*} \leq \frac{\overline{\rho}\left(1 - s^2\right)}{s^2}. \tag{4}$$

According to the above expression, the upper bound of generalization error is negatively correlated with the strength of the combined classifier, but positively correlated with the decision tree. Therefore, by weakening the correlation or enhancing the strength of a single decision tree, the generalization error performance can be improved.

The first is the OOB estimation. The bagging method can be used in self-service sampling. If the data are selected in the future, it will be used to predict the classification accuracy, that is, the OOB estimation of classification error rate. After averaging, the random forest generalization error estimation can be obtained.

The second is the characteristic importance value. The application of the random forest method can determine the specific degree of the importance of a single eigenvalue. At the same time, the performance of each decision tree can be evaluated by using the data outside the bag, that is, the accuracy rate of OOB. By combining with the noise interference test, the performance of the decision tree can be tested more accurately, that is, the new OOB accuracy rate. The important value of feature V in the decision tree can be expressed as the difference between the accuracy of new and old OOB, and its important value is determined after averaging. If there are more features in the basic samples, the best model can be determined by sorting the important values. Figure 1 shows the parameter selection of the random forest algorithm.

3. Data Collection and Processing

3.1. Data Sources. According to the relevant data, there are more than 1700 domestic P2P platforms, which complete the lending process with the help of third-party platforms. At present, Renrendai is the largest and longest established P2P platform in China. Therefore, this paper selects Renrendai loan as the research object, combines R language and python to write a web crawler program, obtains its relevant data, and gets about 50 variables, including amount and interest rate.

3.2. Data Preprocessing

Step 1. Eliminate the variables that do not meet the conditions. Specifically, it includes the variables with the same values, the variables with repeated specific contents, the variables not related to the research topic, and the variables with serious missing data.

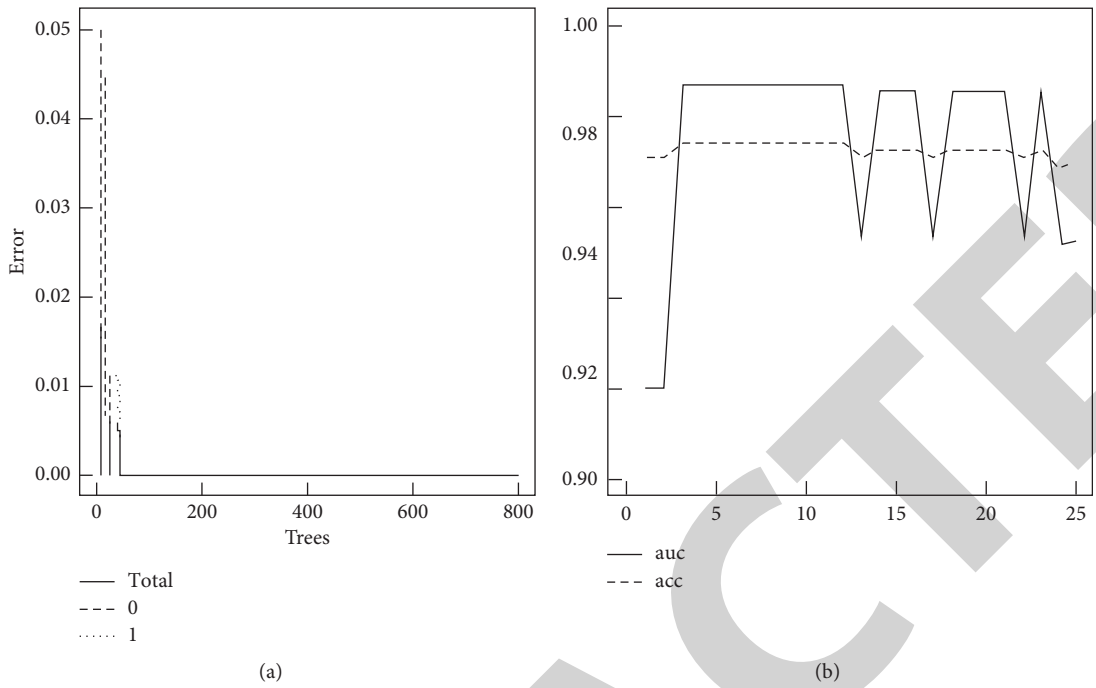

Figure 1: Parameter selection of the random forest algorithm. (a) ntree. (b) mtry.

Step 2. Missing value processing. It is found that the variables of some loan items are incomplete, such as the lack of industry, enterprise scale, and position. According to the specific situation, its industry can be defined as e-commerce; the enterprise scale is expressed as 0; and the position is expressed as individual shopkeeper.

Step 3. Data normalization processing. The output variable selects the number of overdue times, in which if it exceeds 0, it is marked as 0; otherwise, it is marked as 1; 0 and 1 are used to represent binary variable values; integers are used to represent education level, subject type, etc.; working hours are represented by the median value; and the amount of loan can be expressed as $x' = (X - \min(X))/(\max(X) - \min(X)) \times 10$. The basic information of data after preprocessing is shown in Table 1.

4. Credit Risk Assessment Model Based on Data Mining Algorithm

4.1. Unbalanced Data Processing. In the data sample obtained in this paper, there are 30 default items, accounting for 2.935%, and the rest are nondefault items, namely, unbalanced data set. Traditional data mining algorithms have some limitations in dealing with unbalanced distribution classes, and it is difficult to effectively focus on a few classes. Therefore, its classification performance is difficult to meet the requirements.

The data sampling method can be selected, that is, up or down sampling; on the contrary, data mining algorithms can be optimized and improved, such as cost sensitive learning. Through comparative analysis, it can be found that the possibility of incomplete data in down sampling is high. Therefore, the application of upward sampling is more extensive. The basic up sampling method achieves the balance of data sets by randomly copying a few samples, but it is difficult to avoid the fitting problem.

The smote algorithm uses a small number of samples to construct artificial samples, thus achieving the balance of data sets, which is conducive to avoid the over fitting phenomenon. In this algorithm, the artificial samples are inserted into the adjacent samples in the feature space to increase the number of samples. For $X_i \in S_{\min}$, k nearest neighbor points are searched, and the nearest neighbor points are selected by combining the parameters such as correlation coefficient or Euclidean distance. After determining the nearest neighbor points, the corresponding sample points Y_i are determined. On the basis of determining the difference between X_i and Y_j with the corresponding eigenvector, the random number δ is determined, and then the artificial samples X_{new} are determined as follows:

$$X_{\text{new}} = X_i + \left(X_i - Y_j\right) \times \delta, \tag{5}$$

where j is the number of sample points $j = 1, 2, \ldots, n$. Repeat the above steps, and stop after all minority samples are processed.

The smote algorithm uses the smote function to complete the confirmation of majority class sample m and minority class n. On the basis of determining the up and down sampling rates, the final majority class sample and minority class sample number $N + nN$ and nNm are obtained.

The first is sample classification, which is divided into the test set sample and training set sample by random sampling; the second training set data balance is that minority class

Table 1: Basic information of data.

No.	Variables	Actual meaning	Data processing
1	ID	User ID	
2	BorrowType	Type of standards	1: credit certification standard; 2: institutional guarantee standard; 3: field certification standard
3	Title	Loan title	1: capital turnover; 2: expansion of production/operation; 3: decoration; 4: purchase of goods/raw materials/equipment; 5: daily life consumption; 6: purchase of cars; 7: investment in entrepreneurship; 8: purchase of house; 9: online store stocking/business expansion/capital turnover; 10: education and training expenditures; 11: marriage expenses; 12: others
4	Amount	Loan amount	0~10
5	Interest	Interest rate	
6	Term	Repayment period	
7	LoanType	Way of guarantee	1: principal; 2: principal + interest
8	Prepayment Rate	Prepayment Rate	
9	Gender	Gender	0: female; 1: male
10	Age	Age	
11	Education	Education	1: high school or below; 2: junior college; 3: undergraduate; 4: graduate or above
12	Marriage	Marriage	1: married; 2: unmarried; 3: divorced; 4: widowed
13	Field	Company industry	0: one; 1: retail/wholesale; 2: manufacturing; 3: public utilities; 4: catering/hotel industry; 5: construction engineering; 6: nonprofit organization; 7: education/training; 8: transportation industry; 9: IT; 10: finance/law; 11: medical/hygiene/health; 12: government agencies; 13: others
14	Size	Company scale	0: one; 1: <10 persons; 2: 10–100 persons; 3: 100–500 persons; 4: >500 persons
15	EmpLength	Years of working	1: 1 year or less; 2: 1–3 years (including 3 years); 3: 3–5 years (including 5 years); 4: more than 5 years
16	Income	Income	12000~5000 RMB; 25000~10000 RMB; 310000~20000 RMB; 420000~50000 RMB; more than 550000 RMB
17	House	House	0: no; 1: yes
18	Mortgage	Mortgage	0: no; 1: yes
19	Car	Car	0: no; 1: yes
20	Autoloan	Car loan	0: no; 1: yes
21	Score	Credit score	
22	Grade	Credit rating	1:A; 2: B; 3: C; 4: D; 5: E; 6: IIR
23	Application	Borrowing application	
24	Succeed	Successful borrowing	
25	Paid	Number of paying off	
26	Total	Total amount of borrowing	0~10
27	Overdue	Overdue times	
28	Status	Repayment situation	0: overdue time >0; 1: overdue time = 0

$N = 15$, majority class $M = 496$, taking $n = 500\%$, $m = 200\%$, $k = 5$, and keeping the proportion of 3 : 5, so as to improve the model performance. Table 2 shows the data composition.

4.2. Model Empirical Analysis. In this paper, the classification variable is repayment, and then the scientific selection of model parameters is performed, in order to obtain the analysis results of different data mining models, to lay the foundation for the subsequent empirical analysis. Table 3 lists the results of parameter selection and important variables of each model.

(1) Determine the model parameters and output the corresponding results. In this paper, the random forest algorithm is selected to determine the number of decision trees and the number of variables of node branches (mtry). Then the model is built according to the new training set. If the number of decision trees is less than 40, the error rate fluctuation is not more than 0.05; if the number of decision trees is more than 40, the prediction error rate is reduced to 0; determine the selected variables of 3–13 nodes to achieve the maximum AUC and accuracy under stable state. To sum up, ntree = 800 and mtry = 3 can be selected to complete the model construction, and each category can be accurately predicted. Figure 2 shows the Friedman average ranking.

On the whole, the variables with higher importance were paid, succeeded, application, score, field, etc., while the variables with lower importance were house and marriage. The importance of some variables is 0. Therefore, in the process of credit risk assessment, personal work information, credit rating, and historical records are the main variables. Relatively speaking, the importance of personal life information is lower than the above variables. Taking Renrendai as an example, the platform is based on the credit rating mechanism, combined with the

Table 2: Data composition.

Data processing	Data set	0	1	Total	0 (%)	1 (%)
Random sampling	Original data	30	992	1022	2.935	97.065
	Test set (test)	15	496	511	2.935	97.065
	Training set (train)	15	496	511	2.935	97.065
SMOTE	New training set (ntrain)	90	150	240	37.5	62.5

Table 3: Parameter selection and important variables of each model.

Model	Original training set		New training set	
	Parameter selection	Important variables	Parameter selection	Important variables
CART	—	Succeed, field	—	Succeed
C4.5	—	Succeed, field, size	—	Autoloan, score
AdaBoost	—	Succeed, field, tile, application, size, score	—	Succeed, empLength, paid, size, grade, borrowType
SVM	C: classification mode, polynomial kernel function	—	C: classification mode, polynomial kernel function, weight is 2 : 1.4	—
ANN	Number of hidden nodes = 6, maximum iteration times = 200	—	Number of hidden nodes = 11, maximum iteration times = 207	—
RF	Number of spanning trees = 800, number of variables selected by node branches = 25	Paid, succeed, score, field, application, grade	Number of spanning trees = 800, number of variables selected by node branches = 3	Paid, succeed, application, score, size, grade

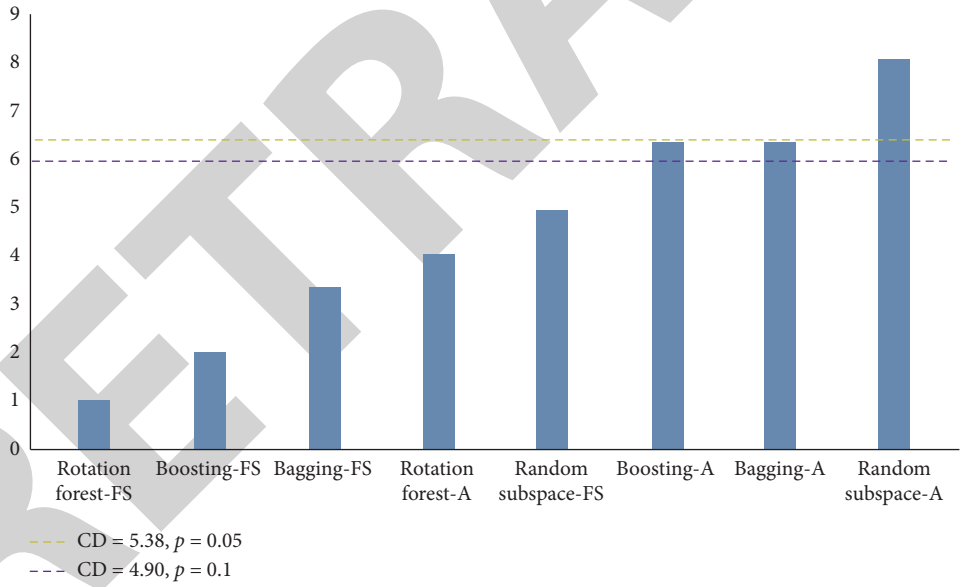

Figure 2: Friedman average ranking.

materials provided by the applicant, and serves as a reference for investors. In the main variables of credit risk evaluation, historical loan information can reflect the use of customer loans, while personal work information can reflect the stability of applicants' work, which is an important reference for evaluating their repayment ability. For the platform, we must further strengthen the collection, collation, and storage of data, provide more powerful information support for credit risk assessment and qualification review, and help investors obtain more income on the premise of ensuring the safety of investors' funds to the greatest extent. Table 4 shows classification results summary of each model.

(2) Comparative analysis of model performance is done before and after data balance processing. In the aspect of classifier performance evaluation, the accuracy index is usually selected; however, for the classification of unbalanced data, it is not appropriate to select the accuracy only. Therefore, the original model can be optimized by combining the

Table 4: Classification results summary of each model.

Model	Original training set						New training set					
		0	1	True positive rate	True negative rate	Accuracy rate	AUC	0	1	True positive rate	True negative rate	Accuracy rate	AUC
CART	0	13	2	0.7	0.977	0.972	0.885	0 13	1	0.932	0.962	0.962	0.948
	1	12	481					1 17	476				
C4.5	0	8	5	0.5	0.975	0.964	0.784	0 12	1	0.866	0.971	0.972	0.921
	1	13	483					1 10	481				
AdaBoost	0	8	5	0.5	0.985	0.972	0.792	0 11	3	0.7	0.971	0.963	0.882
	1	7	488					1 13	480				
SVM	0	7	6	0.531	0.992	0.97	0.761	0 12	4	0.732	0.975	0.961	0.854
	1	2	492					1 11	479				
ANN	0	9	4	0.666	0.981	0.972	0.825	0 11	3	0.7	0.983	0.981	0.891
	1	8	485					1 4	489				
RF	0	9	4	0.665	0.997	0.987	0.833	0 14	0	2	0.971	0.972	0.984

Table 5: Results of 3-fold cross validation.

Model	3-fold cross validation	TP	FP	FN	TN	True positive rate	True negative rate	Accuracy rate	AUC
CART	1	9	10	0	321	9	0966	0965	0982
	2	6	4	2	325	08	0984	0972	0841
	3	8	14	0	314	08	0952	0954	0928
	Mean value					0866	0965	0961	0917
C4.5	1	8	8	2	321	08	0972	0972	0935
	2	6	2	2	327	06	0994	0983	0841
	3	8	11	0	317	08	0962	0963	0933
	Mean value					0832	0975	0973	0906
AdaBoost	1	8	7	2	321	08	0975	0975	0939
	2	6	5	4	324	06	0981	0975	0842
	3	6	8	4	320	06	0972	0965	0835
	Mean value					0762	0975	0972	0871
SVM	1	6	11	4	318	06	0963	0952	0835
	2	6	10	4	321	06	0966	0956	0835
	3	7	18	3	310	07	0941	0934	0875
	Mean value					0732	0955	0952	0841
ANN	1	7	11	3	318	07	0961	0953	0883
	2	8	14	2	317	08	0954	0954	0928
	3	8	23	2	307	08	0926	0925	0915
	Mean value					0866	0942	0942	0911
RF	1	9	6	1	325	9	0971	0981	0983
	2	6	2	4	327	06	0992	0981	0841
	3	9	11	1	317	9	0963	0964	0983
	Mean value					0911	0977	0972	0935

specific index and sensitivity index. We can compare the ROC curve of each model in Figure 3.

The first is accuracy. The model built by the new training set can basically achieve the accuracy rate of 0.963–0.982, among which ANN, RF, and C4.5 rank in the top three. Even though the accuracy of cart, AdaBoost, SVM, RF has declined, it can more accurately predict a few items. Among them, C4.5 and ANN models have greatly improved the prediction accuracy based on the original training set model.

The second is ROC curve and AUC. The closer to the upper left corner, the higher the accuracy of the model prediction. Comparatively speaking, the ROC curve of the model constructed based on the new training set is more concentrated in the upper left corner, which indicates that the classifier has better performance. In particular, after the original sample is balanced by using the smote algorithm, the model constructed based on this has significantly higher AUC, which is more than 0.85. RF, cart, and C4.5 rank in the top three, and RF, cart, and C4.5 rank the best. The AUC of random forest method is very close to 1, reaching 0.987. Compared with other models, its advantages are very significant.

Generally speaking, in the related research of credit risk evaluation, it is of great significance to

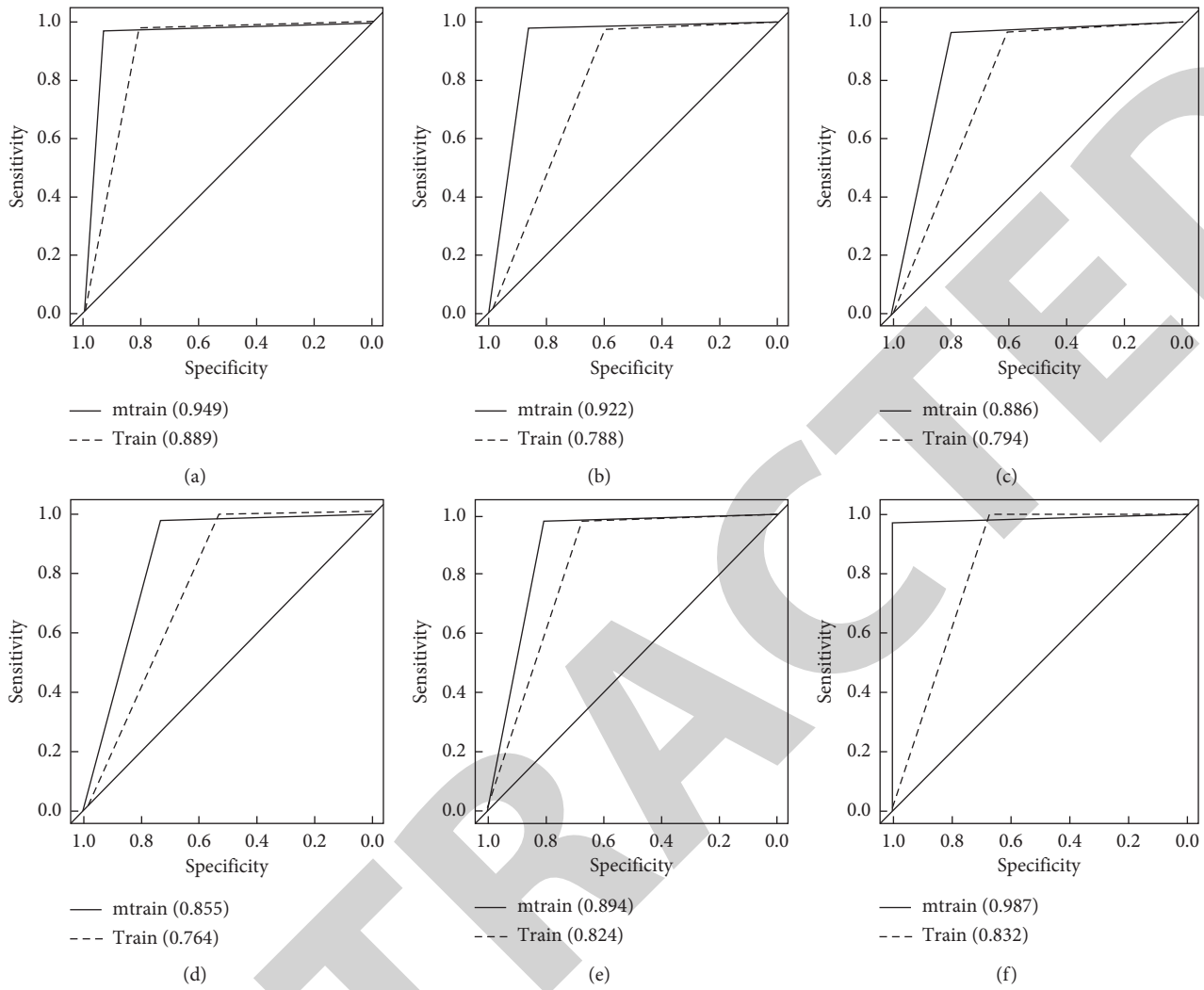

Figure 3: ROC curve of each model. (a) CART. (b) C45. (c) AdaBoost. (d) SVM. (e) RF. (f) ANN.

strengthen the research on the prediction of a small number of samples, which can provide more information support for relevant investors, help investors choose investment projects more scientifically, so as to minimize the credit risk, and improve the security of funds, which has good practical value. In this case, according to the characteristics of the original training set, this paper introduces the smote algorithm to deal with it, which greatly improves the performance of credit risk assessment model, and improves the accuracy of default project prediction.

(3) The prediction performance of different models was compared and analyzed. Through the analysis of Table 4, it can be found that the true rate of the random forest model in the model built based on the new training set is about 1, and its AUC is as high as 0.987, which has relatively high accuracy and has good identification ability for relevant default samples. In summary, this paper preliminarily determines that the random forest model has higher prediction accuracy and the best performance.

In order to verify the conclusion of this paper and determine the best model, this paper selects a 3-fold cross validation method. According to the standard of this paper, dependent variables include default variables and nondefault variables. In order to balance the two categories in the original data, we can divide them into three parts randomly, that is, three data sets including default variables and nondefault variables, and run them as test sets. The data sets are processed by using the smote algorithm, and then the corresponding models are established and their classification performance is evaluated to carry out the targeted test. It can be seen from Table 5 that the mean values of true positive rate are larger, the difference is larger.

Among them, the models with more than 0.85 include RF, cart, and ANN, which are in the forefront. Therefore, the above three models have high recognition ability for default items; the true negative rate of RF, AdaBoost, and C4.5 is in the top three, and the accuracy rate of RF, C4.5, and

AdaBoost is in the top three. Considering that the accuracy rate is difficult to distinguish the minority class from the majority class, the accuracy can only be used as a reference to determine the best model, rather than the main factor. The AUC of RF, cart, and ANN ranked in the top three. To sum up, the best performance is the random forest model, which has broad application prospects in the evaluation of network lending credit.

5. Conclusion

This paper comprehensively and systematically studies the credit risk factors in P2P network lending and constructs a data mining model in risk assessment, which lays the foundation for the follow-up research. The smote algorithm is used to process the unbalanced data, and then the corresponding model is established, which can reduce the volatility of prediction accuracy and improve the risk identification ability of AUC index and default items. The future research focuses on the following: first, strengthen the analysis of user behavior; second, judge the correlation between user behavior and credit risk; and third, build a user credit risk assessment system to provide real-time search function for the platform.

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also forms part of an ongoing study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The findings were sponsored by the National Social Science Fund of China (Grant no. 18CGL015).

References

[1] H. Liang, J. Zou, K. Zou, and M. J. Khan, "An improved genetic algorithm optimization fuzzy controller applied to the wellhead back pressure control system," *Mechanical Systems and Signal Processing*, vol. 142, no. 8, Article ID 106708, 2020.

[2] J. Hu, Y. Sun, G. Li, G. Jiang, and B. Tao, "Probability analysis for grasp planning facing the field of medical robotics," *Measurement*, vol. 141, no. 7, pp. 227–234, 2019.

[3] F. Hu and G. Wu, "Distributed error correction of EKF algorithm in multi-sensor fusion localization model," *IEEE Access*, vol. 8, no. 5, pp. 93211–93218, 2020.

[4] D. Jiang, G. Li, Y. Sun, J. Kong, and B. Tao, "Gesture recognition based on skeletonization algorithm and CNN with ASL database," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 29953–29970, 2019.

[5] H. Liang, J. Zou, Z. Li, M. J. Khan, and Y. Lu, "Dynamic evaluation of drilling leakage risk based on fuzzy theory and PSO-SVR algorithm," *Future Generation Computer Systems*, vol. 95, no. 6, pp. 454–466, 2019.

[6] Z. Liu, B. Hu, B. Huang, L. Lang, H. Guo, and Y. Zhao, "Decision optimization of low-carbon dual-channel supply chain of auto parts based on smart city architecture," *Complexity*, vol. 2020, no. 5, 14 pages, Article ID 2145951, 2020.

[7] M. Buchholz and L. Tonzer, "Sovereign credit risk comovements in the eurozone: simple interdependence or contagion?" *International Finance*, vol. 19, no. 3, pp. 246–268, 2016.

[8] W. Sun, Y. Zhao, and L. Sun, "Big data analytics for venture capital application: towards innovation performance improvement," *International Journal of Information Management*, vol. 50, no. 2, pp. 557–565, 2020.

[9] M. Xie, H. Li, and Y. Zhao, "Blockchain financial investment based on deep learning network algorithm," *Journal of Computational and Applied Mathematics*, vol. 372, no. 7, p. 112723, 2020.

[10] Z. Liu, B. Hu, Y. Zhao et al., "Research on intelligent decision of low carbon supply chain based on carbon tax constraints in human-driven edge computing," *IEEE Access*, vol. 8, no. 3, pp. 48264–48273, 2020.

[11] L. Sun, Y. Zhao, and W. Sun, "Study on supply chain strategy based on cost income model and multi-access edge computing under the background of the internet of things," *Neural Computing and Applications*, 2019.

[12] S. Morris and H. S. Shin, "Illiquidity component of credit risk—the 2015 Lawrence R. Klein lecture," *International Economic Review*, vol. 57, no. 4, pp. 1135–1148, 2016.

[13] Z. Liu, S. Chen, B. Hu, M. Zhou, and Y. Zhao, "Research on staged pricing model and simulation of intelligent urban transportation," *IEEE Access*, vol. 7, no. 9, pp. 141404–141413, 2019.

[14] F. Wang, L. Ding, H. Yu, and Y. Zhao, "big data analytics on enterprise credit risk evaluation of e-business platform," *Information Systems and e-Business Management*, 2019.

[15] J. Wang, X. Wang, F. Meng, R. Yang, and Y. Zhao, "Massive information management system of digital library based on deep learning algorithm in the background of big data," *Behaviour & Information Technology*, pp. 1–12, 2020.

[16] L. Zheng, X. Qi, and K. Yang, "Optimal independent pricing strategies of dual-channel supply chain based on risk-aversion attitudes," *Asia-Pacific Journal of Operational Research*, vol. 35, no. 2, pp. 1–17, 2018.

[17] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory," *Knowledge and Information Systems*, vol. 33, no. 2, pp. 245–265, 2012.

[18] R. Yang, L. Yu, Y. Zhao et al., "Big data analytics for financial market volatility forecast based on support vector machine," *International Journal of Information Management*, vol. 50, no. 2, pp. 452–462, 2020.

[19] X. Qi, W. Wang, and L. Zheng, "The influence of online subsidies service on online-to-offline supply chain," *Asia-Pacific Journal of Operational Research*, vol. 35, no. 2, pp. 1–14, Article ID 1840007, 2018.