


Research Article

A Study on RB-XGBoost Algorithm-Based e-Commerce Credit Risk Assessment Model

Weimin Yang¹ and Lili Gao² 

¹Jiangsu Vocational Institute of Commerce, Nanjing, Jiangsu 211168, China

²Southeast University, Nanjing, 211189 Jiangsu, China

Correspondence should be addressed to Lili Gao; gaolili0919@seu.edu.cn

Received 6 August 2021; Revised 30 August 2021; Accepted 27 September 2021; Published 25 October 2021

Academic Editor: Mu Zhou

Copyright © 2021 Weimin Yang and Lili Gao. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The current method's e-commerce credit risk assessment is prone to poor data balance and low evaluation accuracy. An RB-XGBoost algorithm-based e-commerce credit risk assessment model is proposed in this study. The adaptive random balance (RB) method is used to sample and process the obtained data to improve the balance degree of the data. An assessment index system is constructed based on the processed data. Based on the risk evaluation index system and the XGBoost algorithm, this paper constructed an e-commerce risk assessment model and assessed the e-commerce credit risk using this model. The experimental results show that the proposed method has good data balance, a high kappa coefficient, and a large receiver operating characteristic (ROC) curve area, which can effectively improve e-commerce credit risk assessment accuracy.

1. Introduction

At present, e-commerce has entered society, and informatization has become an inevitable trend and core content of e-commerce, which has a significant impact on the fields of culture, society, and politics [1, 2]. In network economic activities, this technology effectively improves resource allocation and enhances China's economic competitiveness. Therefore, the progress of e-commerce technology is of great significance in economic growth, industrial structure optimization, and economic operation quality and efficiency in China. However, the problem of the credit crisis will lead to great risks in the practical application of e-commerce and seriously restrict the steady development of e-commerce. Therefore, it is necessary to analyze and study the e-commerce credit risk assessment methods to avoid the risks in e-commerce transactions.

Wu et al. minimizes e-commerce credit assessment indicators by a rough set method to obtain important influencing factors of assessment in [3]. A C-XGBoost model is first established to forecast for each cluster of the resulting clusters based on a two-step clustering algorithm, incorpo-

rating sales features into the C-XGBoost model as influencing factors of forecasting in [4]. Aiming at the customer characteristics of social network e-commerce, Zhuang builds a customer value model that integrates the value of social network to help companies subdivide the customer accurately in [5]. To improve and enhance the predictive ability of consumer purchasing behaviours on e-commerce platforms, a new method of predicting purchasing behaviour on e-commerce platforms is created in [6]. In the support vector regression method, a particle swarm optimization algorithm is introduced to optimize the model parameters, and the optimized model is used to complete the assessment of e-commerce credit risk. This method has good effectiveness, but the data imbalance rate obtained by this method is high, leading to a poor data balance degree. Chang et al. determines the risk assessment indicators based on the actual transaction situation and relevant literature and constructs a two-layer hybrid model to evaluate the credit risk of e-commerce combined with the back propagation (BP) neural network and naive Bayesian algorithm [7]. This method has relatively high assessment stability but does not process the data set before assessment, resulting in the

unsatisfactory effect of the ROC curve obtained by this method and the problem of low assessment accuracy. An e-commerce credit risk assessment model based on the RB-XGBoost algorithm is proposed to solve the issues in the above methods.

2. System and Model Description

2.1. e-Commerce Credit Risk Assessment Index System

(I) Data balance processing

The e-commerce credit risk assessment model based on the RB-XGBoost algorithm is used to sample and process e-commerce risk data through the adaptive random balance RB method to reduce the imbalance of data [8–10]. The specific process is shown in Figure 1.

(II) Grey correlation analysis of data

We set that m stands for the number of e-commerce enterprises, n stands for the number of risk assessment indicators, and $x_i = \{x_{i(1)}, x_{i(2)}, \dots, x_{i(n)}\}$ is used to describe the i th e-commerce enterprise sample, where $i = 1, 2, \dots, m$.

An ideal sequence $x_j^0 = \{x_1^0, x_2^0, \dots, x_n^0\}$ is established, where $x_j^1 = \max_i \{x_{ij}\}$ represents a positive index and $x_j^2 = \min_i \{x_{ij}\}$ represents a negative index.

There are differences between the dimensions corresponding to different risk assessment indices, so it is necessary to eliminate the data dimensions before data comparison [11, 12]. The negative index is replaced with the positive index and normalizes by the following formula:

$$x'_{ij} = \frac{x_{ij} - x_{\min}}{x_{\max} - x_{\min}}, \quad (1)$$

where x_{\min} and x_{\max} , respectively, represent the minimum and maximum values of the j th risk assessment index and x_{ij} represents the corresponding value of the j th indicator in the i th e-commerce enterprise.

A correlation coefficient ξ_{ij} is set, and its calculation formula is as follows:

$$\xi_{ij} = \frac{\min_i \min_j |x_{ij} - x_j^1| + \partial |x_{ij} - x_j^2|}{|x_{ij} - x_j^1| + \partial \min_i \min_j |x_{ij} - x_j^2|}, \quad (2)$$

where ∂ represents the resolution coefficient.

The correlation degree r_j is calculated according to the correlation coefficient:

$$r_j = \frac{1}{m} \sum_{i=1}^m \xi_{ij}. \quad (3)$$

(III) Risk assessment index system

The risk assessment indices are sorted according to their relevance. In the assessment process, the assessment indices of $r_j > r_0$ are selected to build the risk assessment index system [13, 14], as shown in Figure 2.

2.2. e-Commerce Credit Risk Assessment Model. The establishment of the e-commerce credit risk assessment model based on the RB-XGBoost algorithm uses the XGBoost algorithm.

The basic elements for XGBoost model establishment are the tree set. The binary tree structure in the classification regression tree can reflect the actual results of the decision tree. In the decision tree structure, there are two branches of “no” and “yes,” which correspond to the branches on the right and left, respectively. Each feature variable is divided by a binary tree, and the feature space is divided to obtain several leaf nodes.

A set $D = \{(x_i, y_i)\}$ is set, in which there are m variables and n samples. The prediction model is obtained based on the regression tree integration model through K functions, and \hat{y} is an output:

$$\hat{y} = \sum_{k=1}^T f_k(x_i), f_k \in \Gamma, \quad (4)$$

where $\Gamma = \{f(x) = \omega_{q(x)}\} (q: R^m \rightarrow T, \omega_i \in R^m)$ represents the regression tree space, ω_i represents the score corresponding to the i th leaf, T represents the number of leaf nodes in the tree structure, q stands for the tree structure, f_k stands for tree, and x_i represents the independent variable corresponding to the i th sample.

For the tree model, objective function ϑ is used for training:

$$\vartheta = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad (5)$$

where l is the convex loss function to measure the difference between the real value y_i and the predicted value \hat{y}_i and Ω represents the penalty term, and its expression is as follows:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2, \quad (6)$$

where $(1/2)\lambda\|\omega\|^2$ describes the regular term and γ represents leaf node penalty, which is mainly used to avoid overfitting problems.

In the process of e-commerce credit risk assessment, European space cannot be directly used to optimize the objective function [15, 16]. Therefore, the RB-XGBoost algorithm-based e-commerce credit risk assessment model trained the model through boosting learning strategy. The specific process is as follows:

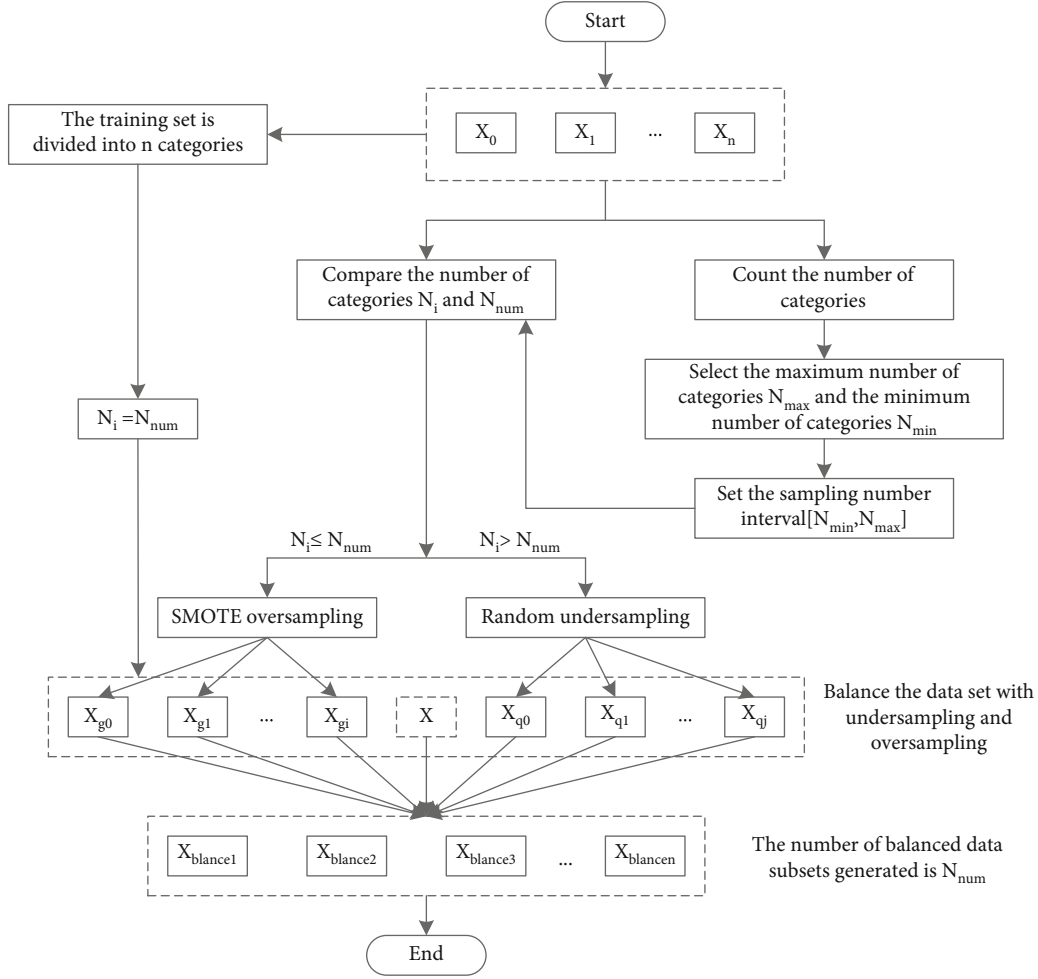


FIGURE 1: Data balance sampling processing flow.

$$\begin{cases} \hat{y}_i^{(0)} = 0, \\ \hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i), \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i), \\ \vdots \\ \hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i), \end{cases} \quad (7)$$

where $\hat{y}_i^{(t)}$ represents the output corresponding to the accumulation model in the t th round of training and $f_t(x_i)$ represents the function newly added to the t th round training.

According to the above process, the objective function is transformed into the following formula:

$$\vartheta = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{constant}, \quad (8)$$

where constant is a constant term.

The fitting results of the model and training data in the assessment process can be measured by the loss function

$L = \sum_{i=1}^n l(\hat{y}_i, y_i)$, in which the logical loss function $l(\hat{y}_i, y_i) = y_i \ln(1 + e^{-y\wedge_i})$ and the square loss function $l(\hat{y}_i, y_i) = (y_i - y\wedge_i)^2$ are widely used in the assessment process [17, 18]. The RB-XGBoost algorithm-based e-commerce credit risk assessment model brings the square loss function into the target function to obtain the following formula:

$$\begin{aligned} \vartheta &= \sum_{i=1}^n l(y_i, (y\wedge_i^{(t-1)} + f_t(x_i)))^2 + \Omega(f_t) + \text{constant} \\ &= \sum_{i=1}^n [2(\hat{y}_i^{(t-1)} - y_i)f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + \text{constant}, \end{aligned} \quad (9)$$

where $\hat{y}_i^{(t-1)} - y_i$ represents the residual.

The loss function can be approximated by the Taylor expansion to obtain the following formula:

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)(\Delta x)^2. \quad (10)$$

$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, y\wedge_i^{(t-1)})$ and $h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, y\wedge_i^{(t-1)})$ are set; then, we obtain the following formula:

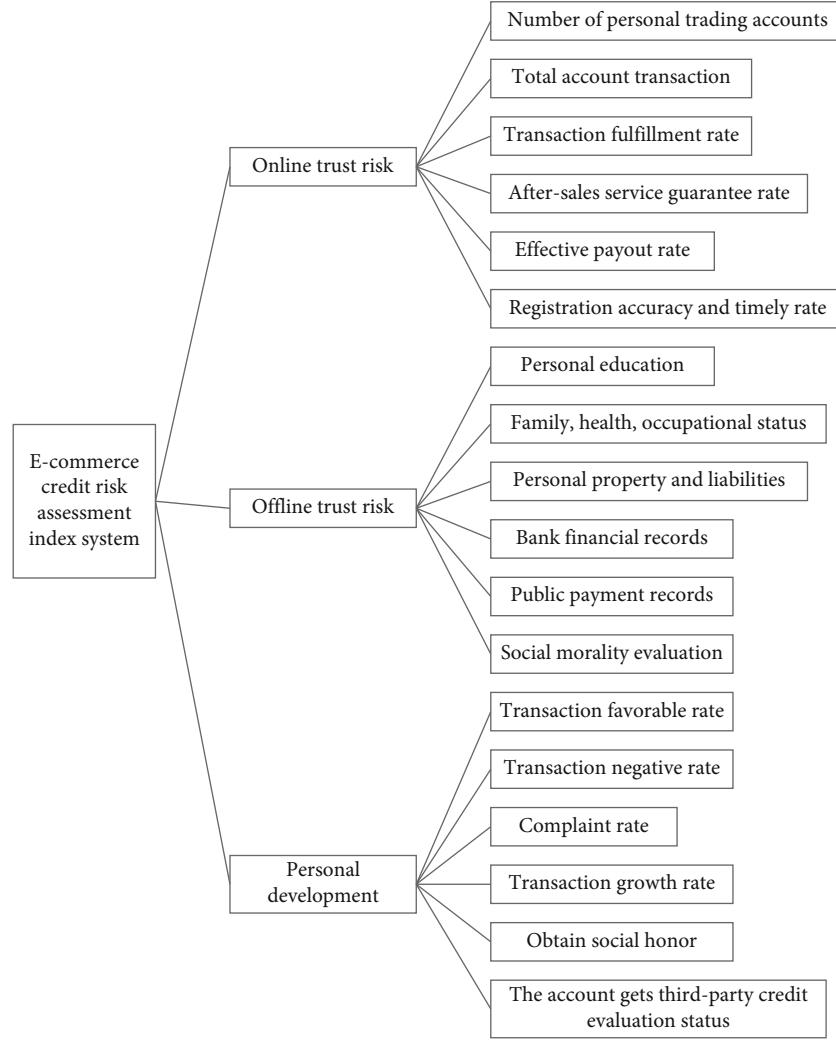


FIGURE 2: E-commerce credit risk assessment index system.

$$l(y_i, y^{\wedge(t-1)} + f_t(x_i)) \approx l(y_i, y^{\wedge(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i). \quad (11)$$

The objective function is substituted into the above loss function to obtain

$$\vartheta = \sum_{i=1}^n \left[l(y_i, y^{\wedge(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + \text{constant}. \quad (12)$$

When the loss function belongs to square loss in the training process, there is the following formula:

$$\begin{cases} g_i = \partial_{y_i^{(t-1)}} l(y_i, y^{\wedge(t-1)}) = 2(\hat{y}_i^{(t-1)} - y_i), \\ h_i = \partial_{y^{\wedge(t-1)}}^2 l(y_i, y^{\wedge(t-1)}) = \partial_{y^{\wedge(t-1)}}^2 (y^{\wedge(t-1)} - y_i)^2 = 2. \end{cases} \quad (13)$$

The parameters g_i and h_i are substituted into the objective function to obtain the following formula:

$$\vartheta = \sum_{i=1}^n \left[y_i - (y^{\wedge(t-1)} + f_t(x_i)) \right]^2 + \Omega(f) + \text{constant}, \quad (14)$$

where $\hat{y}_i^{(t-1)}$ describes the output result of the model during the $t-1$ th round training and y_i describes the dependent variable existing in the objective function. If the dependent variable y_i is known, the above objective function can be simplified to obtain the following formula:

$$\vartheta = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f) + \text{constant}. \quad (15)$$

In the formula, g_i and h_i are the parameters existing in the loss function. The values of the above parameters are different in different loss functions, so the values of parameters g_i and h_i can be determined in the form of the loss function.

Each tree is redefined by the following formula:

$$f_t(x) = \omega_{q(x)}, \omega \in R^T, q : R^d = \{1, 2, \dots, T\}, \quad (16)$$

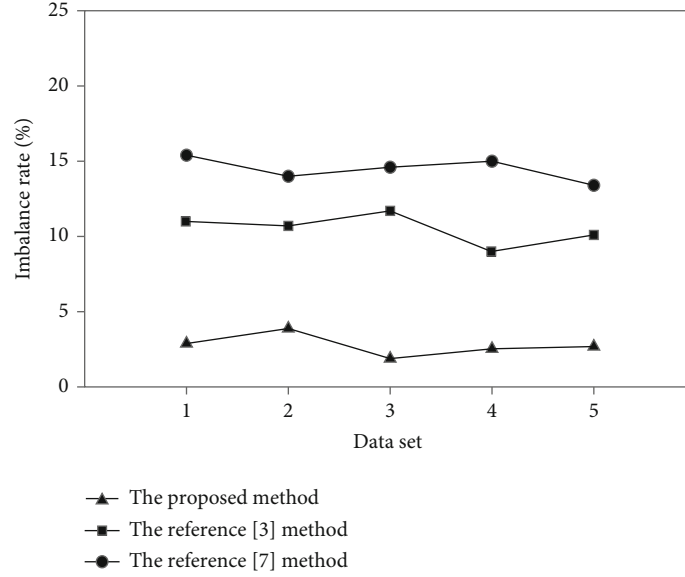


FIGURE 3: Imbalance rate of different methods.

where ω describes the weight corresponding to the leaf node in the tree structure, $\omega_{q(x)}$ describes the predicted value obtained by the tree model, and $q : R^d = \{1, 2, \dots, T\}$ represents the structure of the tree.

Model complexity includes L2 regularization of leaf node score and the total number of leaf nodes T [19, 20]. Model complexity $\Omega(f_t)$ can be obtained through tree definition:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2. \quad (17)$$

The smoothness of leaf nodes can be improved by L2 regularization to solve the overfitting problem [21, 22]. In the objective function, when the complexity of the model increases, there are two different types of accumulation, one of which is $I_j = \{i | q(x_i) = j\}$, where I_j represents the set of samples in the leaf node j . After adding complexity to the objective function, the final objective function is obtained, that is, the e-commerce credit risk assessment model [23, 24]:

$$\begin{aligned} \mu &= \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \\ &= \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T. \end{aligned} \quad (18)$$

Based on the selected risk assessment indices, the risk assessment is performed using the e-commerce credit risk assessment model.

3. Experiments and Results

To verify the effectiveness of the RB-XGBoost algorithm-based e-commerce credit risk assessment model, it is neces-

sary to carry out a test. The proposed method, literature [3] method, and literature [4] method are used for comparative experiments. The imbalance rate τ is used as the experimental index to test the data balance degree of different methods. The calculation formula of imbalance rate τ is as follows:

$$\tau = \frac{N_{\max}}{N_{\min}}, \quad (19)$$

where N_{\max} and N_{\min} represent the maximum and minimum values of the sample data in the set. The larger the imbalance rate τ , the more unbalanced the data. The imbalance rate τ of the proposed method, the reference [3] method, and the reference [7] method are shown in Figure 3.

Based on the data in Figure 3, the data imbalance rate obtained by the proposed method is less than 5% when testing different data sets, while the imbalance rate obtained by the methods of literature [3] and literature [7] fluctuates around 10% and 15%, respectively. It can be seen that the imbalance rate obtained by the proposed method is low, indicating that the data obtained by the proposed method is well balanced. This is due to the data sampling and processing by the adaptive random balance RB method before constructing the e-commerce credit risk assessment model, which ensures the balance of the data.

The assessment accuracy of the proposed method, literature [3] method, and literature [7] method is verified by the kappa coefficient and ROC curve. The kappa coefficient can weigh the difference between the assessment results and the real results. The calculation formula of kappa coefficient K is as follows:

$$K = \frac{p_o - p_e}{1 - p_e}, \quad (20)$$

where p_o represents the proportion of correctly evaluated samples in the total number of samples and p_e represents

TABLE 1: Kappa coefficients for different methods.

Number of iterations (time)	Kappa coefficient		
	The proposed method	The reference [3] method	The reference [7] method
100 s	0.951	0.745	0.695
200 s	0.964	0.715	0.648
300 s	0.978	0.764	0.668
400 s	0.971	0.770	0.678
500 s	0.983	0.721	0.814

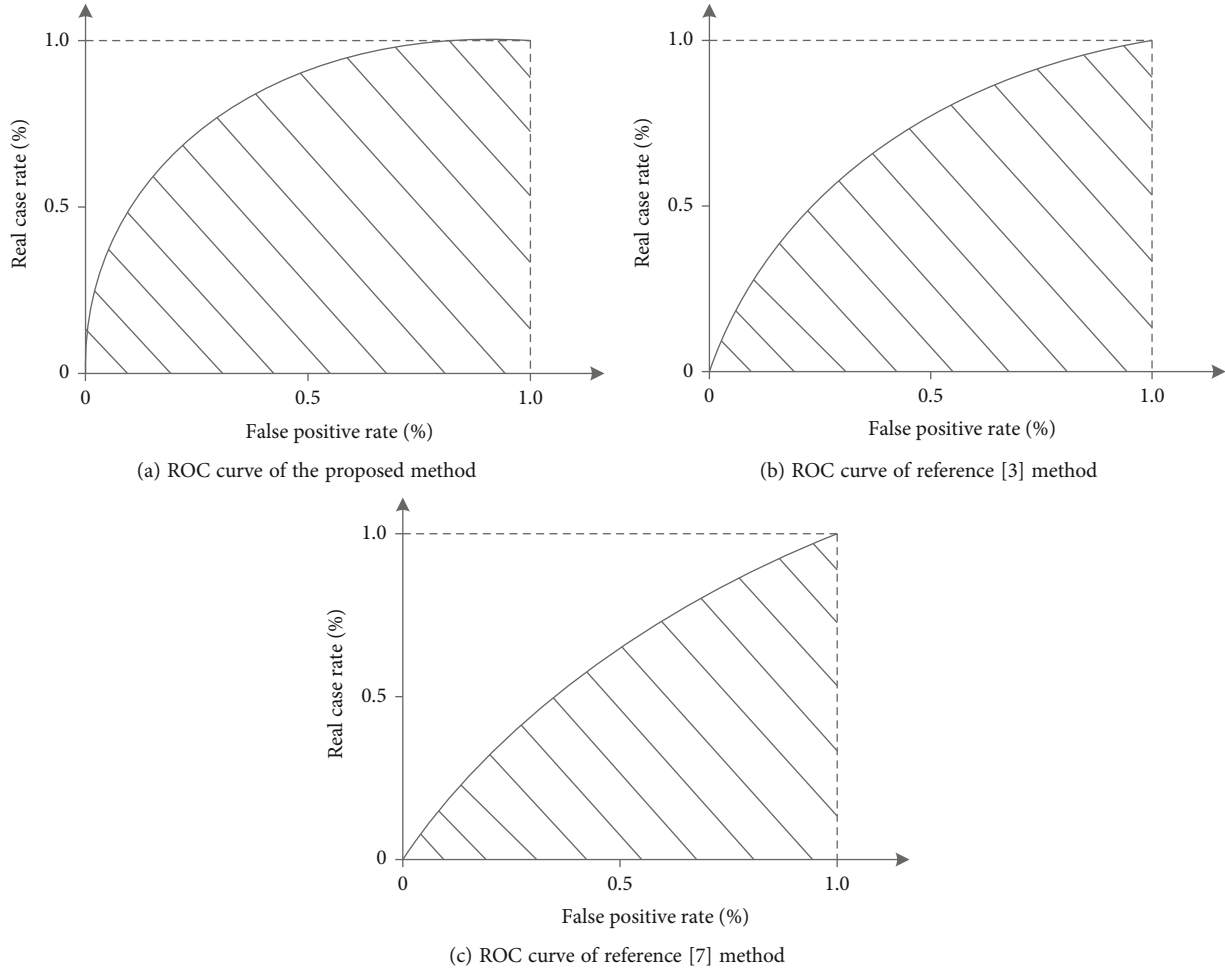


FIGURE 4: ROC curves of different methods.

the randomness ratio. The higher the kappa coefficient K , the more accurate the evaluation results of the method are. The kappa coefficients of the proposed method, the literature [3] method, and the literature [7] method are shown in Table 1.

From the data in Table 1, we can see that the kappa coefficients of the proposed method in multiple iterations are higher than those obtained by the methods in literature [3] and literature [7], indicating that the proposed method can accurately complete the assessment of e-commerce credit risk. This is because this method constructs a risk assessment index system based on the data with a high balance

and completes the assessment of the e-commerce credit risk based on the high-precision risk assessment indices.

The abscissa is the true positive rate in the ROC curve, and the ordinate is the false positive case rate. The larger the area enclosed by the ROC curve and the abscissa, the higher the accuracy of the assessment results of the method. The proposed method, literature [3] method, and literature [7] method are, respectively, used to evaluate the credit risk of different e-commerce enterprises, and the obtained ROC curves are shown in Figure 4.

By analyzing Figure 4, it can be seen that the area enclosed by the ROC curve of the proposed method and

abscissa is larger than that enclosed by the ROC curve of the methods of literature [3] or literature [7] and abscissa, indicating that the proposed method has higher assessment results accuracy and can complete credit risk assessment accurately in e-commerce enterprises.

4. Conclusion

Aiming at the problems of high data imbalance rate and low accuracy of assessment results in the current e-commerce credit risk evaluation methods, an e-commerce credit risk evaluation model based on the RB-XGBoost algorithm is proposed. The risk assessment index system is first constructed by using the data with a high balance rate, and then, the risk assessment model is established by the XGBoost algorithm. This model realizes the assessment of e-commerce credit risk, solves the problems existing in the current methods, ensures the degree of data balance, and improves the accuracy of risk assessment. Future work includes improving the risk assessment model and further enhances the accuracy of risk assessment.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare no competing interests.

Acknowledgments

This study is sponsored by Qing Lan Project.

References

- [1] Y. Liu, Y. Tian, Y. Xu et al., “TPGN: a time-preference gate network for e-commerce purchase intention recognition,” *Knowledge-Based Systems*, vol. 220, no. 2, p. 106920, 2021.
- [2] A. Vera-Baquero, O. Phelan, P. Slowinski, and J. Hannon, “Open source software as the Main driver for evolving software systems toward a distributed and performant E-commerce platform: a Zalando fashion store case study,” *IT Professional*, vol. 23, no. 1, pp. 34–41, 2021.
- [3] J. Wu, “E-commerce credit risk assessment based on rough set and support vector regression,” *Statistics and Decision*, vol. 35, no. 23, pp. 51–54, 2019.
- [4] S. Ji, X. Wang, W. Zhao, and D. Guo, “An application of a three-stage XGBoost-based model to sales forecasting of a cross-border E-commerce enterprise,” *Mathematical Problems in Engineering*, vol. 2019, 15 pages, 2019.
- [5] Y. Zhuang, “Research on E-commerce customer churn prediction based on improved value model and XG-boost algorithm,” *Management Science and Engineering*, vol. 12, no. 3, pp. 51–56, 2018.
- [6] P. Song, “An XGBoost algorithm for predicting purchasing behaviour on E-commerce platforms,” *Technical Gazette*, vol. 27, no. 5, pp. 1467–1471, 2020.
- [7] Z. Chaohui, L. Jiajia, and R. Hui, “Research on E-commerce credit evaluation method based on Bayesian and neural network hybrid algorithm,” *Information Science*, vol. 38, no. 2, pp. 81–87, 2020.
- [8] W. Wei, B. Zhou, D. Polap, and M. Wozniak, “A regional adaptive variational PDE model for computed tomography image reconstruction,” *Pattern Recognition*, vol. 92, pp. 64–81, 2019.
- [9] Y. Zhang, S. Qiao, R. Lu, N. Han, D. Liu, and J. Zhou, “How to balance the bioinformatics data: pseudo-negative sampling,” *BMC Bioinformatics*, vol. 20, no. 25, pp. 695–695, 2019.
- [10] D. Zachariah and P. Stoica, “Effect inference from two-group data with sampling bias,” *IEEE Signal Processing Letters*, vol. 26, no. 8, pp. 1103–1106, 2019.
- [11] U. P. Shukla and S. J. Nanda, “Designing of a risk assessment model for issuing credit card using parallel social spider algorithm,” *Applied Artificial Intelligence*, vol. 33, no. 1-4, pp. 191–207, 2019.
- [12] F. Assef, M. T. Steiner, P. J. Steiner Neto, and D. G. . B. Franco, “Classification algorithms in financial application: credit risk analysis on legal entities,” *IEEE Latin America Transactions*, vol. 17, no. 10, pp. 1733–1740, 2019.
- [13] F. Li, L. Chen, W. Chen et al., “Antibiotics in coastal water and sediments of the East China Sea: distribution, ecological risk assessment and indicators screening,” *Marine pollution bulletin*, vol. 151, pp. 110810–110810. 11, 2020.
- [14] W. Wang, X. Liu, J. Qin et al., “An extended generalized TODIM for risk evaluation and prioritization of failure modes considering risk indicators interaction,” *IIE Transactions*, vol. 51, no. 11, pp. 1236–1250, 2019.
- [15] B. Song, W. Yan, and T. Zhang, “Cross-border e-commerce commodity risk assessment using text mining and fuzzy rule-based reasoning,” *Advanced Engineering Informatics*, vol. 40, pp. 69–80, 2019.
- [16] J. Lv, T. Wang, H. Wang, J. Yu, and Y. Wang, “A SECPG model for purchase behavior analysis in social e-commerce environment,” *International Journal of Communication Systems*, vol. 33, no. 6, article e4149, 2020.
- [17] S. Savoldelli, C. Cattò, F. Villa et al., “Biological risk assessment in the History and Historical Documentation Library of the University of Milan,” *Science of The Total Environment*, vol. 790, no. 5, p. 148204, 2021.
- [18] W. Wei, X. Fan, H. Song, X. Fan, and J. Yang, “Imperfect information dynamic stackelberg game based resource allocation using hidden Markov for cloud computing,” *IEEE Transactions on Services Computing*, vol. 11, no. 1, pp. 78–89, 2016.
- [19] V. Srikrishnan and K. Keller, “Small increases in agent-based model complexity can result in large increases in required calibration data,” *Environmental Modelling and Software*, vol. 138, no. 5, p. 104978, 2021.
- [20] F. Kong, “Development of metric method and framework model of integrated complexity evaluations of production process for ergonomics workstations,” *International Journal of Production Research*, vol. 57, no. 7-8, pp. 2429–2445, 2019.
- [21] F. Zheng, S. Derrode, and W. Pieczynski, “Parameter estimation in switching Markov systems and unsupervised smoothing,” *IEEE Transactions on Automatic Control*, vol. 64, no. 4, pp. 1761–1767, 2019.
- [22] M. Pitek, A. Lisowski, and M. Dbrowska, “The effects of solid lignin on the anaerobic digestion of microcrystalline cellulose and application of smoothing splines for extended data analysis of its inhibitory effects,” *Bioresource Technology*, vol. 320, p. 124262, 2021.

- [23] M. Soui, I. Gasmı, S. Smiti, and K. Ghédira, "Rule-based credit risk assessment model using multi-objective evolutionary algorithms," *Expert Systems with Applications*, vol. 126, pp. 144–157, 2019.
- [24] E. Brons-Piche, G. J. Eckert, and M. Fontana, "Predictive validity of a caries risk assessment model at a dental school," *Journal of Dental Education*, vol. 83, no. 2, pp. 144–150, 2019.