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

Weimin Yang and Lili Gao

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

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, incorporating 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 subdivides 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 $x_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ is used to describe the $i$th e-commerce enterprise sample, where $i = 1, 2, \ldots, m$.

An ideal sequence $x_i^0 = \{x_{i1}^0, x_{i2}^0, \ldots, x_{in}^0\}$ is established, where $x_{ij}^1 = \max \{x_{ij}\}$ represents a positive index and $x_{ij}^2 = \min \{x_{ij}\}$ 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:

$$x_{ij}' = \frac{x_{ij} - x_{\min}}{x_{\max} - x_{\min}},$$

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 $\xi_{ij}$ is set, and its calculation formula is as follows:

$$\xi_{ij} = \frac{\min \{x_{ij} - x_{ij}^0\}}{|x_{ij}^1 - x_{ij}^2| + \partial \min \{x_{ij} - x_{ij}^0\}} + \frac{\partial \min \{x_{ij} - x_{ij}^0\}}{|x_{ij}^1 - x_{ij}^2|},$$

where $\partial$ represents the resolution coefficient.

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

$$r_j = \frac{1}{m} \sum_{i=1}^{m} \xi_{ij},$$

(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 $\tilde{y}$ is an output:

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

where $\Gamma = \{f(x) = \omega_q(x) \mid q : R^m \rightarrow T, \omega_q \in R^m\}$ represents the regression tree space, $\omega_q$ 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 $\vartheta$ is used for training:

$$\vartheta = \sum_{i=1}^{m} l(\tilde{y}_i, y_i) + \sum_{k} \Omega(f_k),$$

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

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

where $\gamma T$ describes the regular term and $\gamma$ 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:
\[
\begin{align*}
\hat{y}^{(0)}_i &= 0, \\
\hat{y}^{(1)}_i &= f_1(x_i) = \hat{y}^{(0)}_i + f_1(x_i), \\
\hat{y}^{(2)}_i &= f_1(x_i) + f_2(x_i) = \hat{y}^{(1)}_i + f_2(x_i), \\
\hat{y}^{(t)}_i &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}^{(t-1)}_i + f_t(x_i),
\end{align*}
\]

where \(\hat{y}^{(t)}_i\) 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:

\[
\theta = \sum_{i=1}^{n} l(y_i, \hat{y}^{(t-1)}_i + f_t(x_i)) + \Omega(f_t) + \text{constant},
\]

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

\[
L = \sum_{i=1}^{n} l(y_i, \hat{y}^{(t-1)}_i + f_t(x_i)) + \Omega(f_t) + \text{constant},
\]

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

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

\[
f(x + \Delta x) = f(x) + f'(x)\Delta x + \frac{1}{2}f''(x)(\Delta x)^2.
\]

\[
g_i = \partial y_i \cdot l(y_i, y^{(t-1)}) \quad \text{and} \quad h_i = \partial^2 y_i \cdot l(y_i, y^{(t-1)}) \quad \text{are set};
\]

then, we obtain the following formula:
The objective function is substituted into the above loss function to obtain

$$\vartheta = \sum_{i=1}^{n} \left( I\left(y_i, y_i^{(t-1)} + f_i(x_i)\right) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right) + \Omega(f) + \text{constant},$$

(11)

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

$$\vartheta = \sum_{i=1}^{n} \left[ I\left(y_i, y_i^{(t-1)} + f_i(x_i)\right) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \Omega(f) + \text{constant}.$$  

(12)

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

$$\begin{cases} g_i = \partial_{y_i^{(t-1)}} I\left(y_i, y_i^{(t-1)} + f_i(x_i)\right) = 2 \left( y_i^{(t-1)} - y_i \right), \\ h_i = \partial_{y_i^{(t-1)}}^2 I\left(y_i, y_i^{(t-1)} + f_i(x_i)\right) = \partial_{y_i^{(t-1)}}^2 \left( y_i^{(t-1)} - y_i \right)^2 = 2. \end{cases}$$

(13)

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

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

(14)

where $y_i^{(t-1)}$ 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:

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

(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:

$$f_i(x) = \omega_{q(x)}, \omega \in R^T, q : R^d = \{1, 2, \cdots, T\}.$$  

(16)
where $\omega$ 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 $q : R^d = \{1, 2, \ldots, T\}$ 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]:

$$\mu = \sum_{i=1}^{n} \left[ g_i f_i(x_i) + \frac{1}{2} h_i f_i(x_i)^2 \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2$$

$$= \sum_{j=1}^{T} \left[ \left( \sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left( \sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T. \quad (18)$$

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-XGBoot algorithm-based e-commerce credit risk assessment model, it is necessary to carry out a test. The proposed method, literature [3] method, and literature [4] method are used for comparative experiments. The imbalance rate $\tau$ is used as the experimental index to test the data balance degree of different methods. The calculation formula of imbalance rate $\tau$ is as follows:

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

where $N_{\text{max}}$ and $N_{\text{min}}$ represent the maximum and minimum values of the sample data in the set. The larger the imbalance rate $\tau$, the more unbalanced the data. The imbalance rate $\tau$ 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
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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100 s</td>
<td>0.951</td>
<td>0.745</td>
<td>0.695</td>
</tr>
<tr>
<td>200 s</td>
<td>0.964</td>
<td>0.715</td>
<td>0.648</td>
</tr>
<tr>
<td>300 s</td>
<td>0.978</td>
<td>0.764</td>
<td>0.668</td>
</tr>
<tr>
<td>400 s</td>
<td>0.971</td>
<td>0.770</td>
<td>0.678</td>
</tr>
<tr>
<td>500 s</td>
<td>0.983</td>
<td>0.721</td>
<td>0.814</td>
</tr>
</tbody>
</table>
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
