Research Article

Research on Credit Risk Identification of Internet Financial Enterprises Based on Big Data

Hua Peng

1Wuyi University, Wuyishan, Nanping 354300, China
2National Changhua University of Education, Changhua 50007, China

Correspondence should be addressed to Hua Peng; penghua@wuyiu.edu.cn

Received 26 August 2021; Accepted 23 October 2021; Published 13 November 2021

1. Introduction

Internet finance refers to a new financial form that, based on traditional finance, realizes functions such as payment, financing, and credit intermediaries under a series of Internet technologies such as big data technology and cloud computing technology [1]. The forms of Internet finance are generally third-party payment, financial e-commerce, credit evaluation, Internet money funds, big data index funds, and other models. Internet finance is the assimilation of the Internet field and the financial field. However, it is not the simple integration of the Internet and finance but the innovative transformation of financial business by using a safe and effective Internet network under certain market conditions. It is the fusion of traditional financial business and the Internet spirit. Internet finance is a new thing, its development speed and scale far exceed that of other industries, and its development prospect is great; therefore, the Internet finance industry is a sunrise industry. The risk of Internet finance is a big difficulty to the development of the industry. The investigation on the risk management of Internet finance has become a hot topic in society and academia [2].

Jiang and Yuan [3] proposed a credit evaluation method in e-commerce transactions based on decision analysis. The credit risk influencing factors of business transactions are divided into several evaluation factors by introducing the Kanno doff integral evaluation method, and the credit risk F-integral evaluation model is constructed to divide the transaction credit grade. The interval distribution of commodity prices is used to add points to successful transactions and deduct points from failed transactions at different levels, to solve the problem of cycle deception in transactions. Based on the credit scoring results, the risk calculation method of e-commerce transactions is designed to evaluate the credit degree of the current transaction. Yang et al. [4] used the enterprise transaction data of an Internet financial platform as the object, analyzed the propagation behavior of overdue loan default, and proposed to build a model to identify the high-risk enterprises of the Internet financial...
2. Credit Risk Detection of Internet Financial Enterprises Based on Big Data

When designing the enterprise credit risk assessment index system, it is essential to consider the defects of the traditional enterprise credit risk assessment index systems [15]. In the development of a credit risk assessment index system, the integration of dynamic data and static data is adopted to complete data mining, analysis, and modeling from big data. Moreover, the enterprise identity data, behavior data, and external data are used to construct an enterprise credit risk assessment index system [16]. It is crucial that the enterprise credit risk evaluation index system shall be mainly designed from the perspectives of income information, loan information, account information, repayment, and overdue information, and third-party information, such as multi-end loan information, black and gray list, credit information, and so on. The enterprise credit risk evaluation index system based on big data has the characteristics of rich evaluation data items, a combination of static data and dynamic data, wide data sources, and timeliness. Although the traditional enterprise credit risk assessment index system in the crowd coverage is not ideal, the data are mostly static, and the data authenticity cannot be verified. The data of the big data platform for Internet enterprises are mostly used from multiple channels, such as APP product data of companies, credit information data of the PBC, credit information data of Internet credit information companies, online shopping data of e-commerce companies, data of third-party cooperative enterprises, public data crawled by crawlers, and data published by the Public Security and Inspection Law. The data used in this study are desensitized from many aspects to ensure the safety of the data.

2.1. Credit Evaluation Index Weight Calculation. The improved analytic hierarchy process (AHP) method [17] is used to compute the weight of evaluation indicators. This method uses the concept of optimal transfer matrix to improve AHP, make it naturally meet the consistency requirements, and directly calculate the weight value. The major steps are as follows:

(i) A credit risk evaluation index system is established based on market transactions and the evaluation index is set according to the customer credit evaluation index system.

(ii) A judgment matrix is built. After establishing the matrix according to the credit risk evaluation index system, the weight of each level index in the customer credit risk evaluation index system is determined by AHP. By comparing the two elements, the relative importance of each element in the hierarchy relative to a certain factor in the upper hierarchy is determined, and a judgment matrix is created. A comparison matrix of two factors for a given criterion is calculated as follows:

\[ Y = (y_{ij})_{n \times n}, \quad i, j = 1, 2, \ldots, n \]  (1)
where \( y_{ij} \) is the scale of the importance of factor \( Y_i \) and factor \( Y_j \) relative to the criterion index.

(iii) The weight of evaluation indexes at all levels is calculated based on the improved AHP method. There is no need to do a consistency check after calculating the weights of indexes by the improved AHP method. Firstly, the judgment matrix is modified to get the quasi-optimal matrix \( Y^* \), followed by the square root method is used to solve the \( Y^* \) eigenvector. Next, the elements of the judgment moment \( Y^* \) are multiplied by the lines to get the following expression:

\[
M_i = \prod_{j=1}^{n} y_{ij}, \quad i = 1, 2, 3, \ldots, n.
\]  

By dividing the product into \( n \) power roots, an equation \( P_i = N_i^{1/n} \) is obtained, and the root vector \( P = [P_1, P_2, \ldots, P_n]^T \) is normalized, i.e., the sorting weight vector \( P \) can be calculated as

\[
P_j = \frac{P_j}{\sum_{j=1}^{n} P_j}.
\]  

The improved AHP method is represented in Figure 1.

2.2. Credit Evaluation Based on Linear Weighted Synthesis Method. The linear weighted synthesis method for customer credit evaluation is a kind of comprehensive method to obtain the comprehensive evaluation value by weighted summation of each index value.

The evaluation value of the credit risk of the first \( j \) customer can be calculated as

\[
F = \frac{\sum_{i=1}^{n} Y_{ji} \times M_i}{Q_{ji}},
\]  

where the index value of the \( i \) index of the \( j \) customer is \( Q_{ji}, j = 1, 2, \ldots, m \).

The static client credit risk rating is classified into four levels, with grade 1 denoting 89–100 points, grade 2 denoting 75–88 points, grade 3 denoting 59–74 points, and grade 4 denoting ≤58 points. The weighted credit scoring method is used in static client credit risk rating. The first step is to score some credit survey indicators as a whole, the second step is to uniformly weighted average, and the last step is to get the method of credit risk score. This can be computed using the following equation:

\[
X = \sum_{i=1}^{n} (b_i \cdot Y_i),
\]  

where the credit score of a customer is represented by \( X \), The weight of the proposed \( i \)th credit investigation index is represented by \( b_i \) and set to \( \sum_{i=1}^{n} (b_i \cdot Y_i) = 1 \). The evaluation score of the \( i \)th evaluation index is shown by \( Y_i \). This method is applied to quantitative analysis and research of customer credit. The theoretical basis of financial enterprise decision making is mainly to obtain the credit status of the enterprise based on objective data, while the static client credit rating method is a simple and easy method to understand and operate.

2.3. Credit Risk Detection Based on Big Data. According to the regional division characteristics of big data credit risk [18], the credit risk is detected by rule-based matching method. The centralized detection process is shown in Figure 2.

\[
\text{Max} f(x) = \frac{1}{\sqrt{\sum_{i=1}^{n} (b_i(i) - y_i(i))}}
\]  

The data are divided into different packets and matched. If the matching process is successful, the output is generated. The credit risk of big data is divided into five areas and the centralized detection problem will be converted to the target maximum value, and read the data of the fitness function before the processing step, where \( y_k(x) \) is the actual output of data, \( y^*_k(x) \) represents the expected output of data, and \( k \) shows the total amount of data obtained.

The matching process of packets given in Figure 2 is as follows.

2.3.1. High Credit Risk Data Matching Rules. In the process of matching from right to left, if the character in the string does not match its corresponding string, it shall be handled as follows.

If \( x[i] \) data are in a string \( S[1, \ldots, n] \), the string can be moved to \( x[i + 1] \) character alignment, as shown in Figure 3.

\( H(x) \) is set to indicate the height at which the string \( S \) moves upward, \( I \) indicates the length of the string, \( R(x) \) indicates that the string \( S \) appears at the position closest to the right of the data \( x \), and \( i \) indicates that the position below the position does not match (the length from the far left), as shown below:
2.3.2. Matching Rules for Low Credit Risk Data. When different strings are aligned, the moving distance is determined by matching rules of low credit risk data during the matching process from right to left. The specific matching process is as follows.

The string \( S \) is entered and the moving distance is initialized. Next, the string is traversed from right to left, and the traversal position \( i \) is analyzed as given in the following equation:

\[
H(x) = \begin{cases} 
    l - i, & \text{the data } X \text{ are not in the string,} \\
    i, & \text{the data } X \text{ are in the string.}
\end{cases}
\] (7)

All the strings are aligned and matched one by one from right to left. If it is matched with the leftmost end of the string, it means that the matching is successful. According to the matching process, it can ensure that any distance is a safe match and no omission will occur. Furthermore, it can achieve centralized and precise detection of big data credit risk areas at the fastest matching speed.

3. Credit Risk Identification Model of Internet Financial Enterprises

Based on the XGBoost algorithm [19], the credit risk identification model of Internet financial enterprises is constructed. It is a common and effective open-source implementation of the gradient boosted trees algorithm. The XGBoost algorithm provides better performance because of its vigorous handling of different data types, distributions, relationships, and the variety of hyperparameters that can be fine-tuned [20]. The XGBoost algorithm can be used for regression, binary and multiclass classification, and ranking problems [21]. The basic element of constructing the XGBoost model is a tree set, and the binary tree structure is in the classification and regression tree which can reflect the actual result of the decision tree [22]. There are two branches in the structure of the decision tree, namely, "Yes" and "No," which correspond to the right and left branches. Each feature variable is partitioned by the binary tree, and the feature space is partitioned to obtain several leaf nodes.

Suppose a set \( B = \{(x_i, y_i)\} \), in which there are \( m \) variables and \( n \) samples. Through \( K \) functions, the output \( \hat{B} \) of the prediction model can be obtained based on the regression tree integration model:

\[
\hat{B} = \sum_{k=1}^{T} f_k(x_i), \quad f_k \in \Gamma,
\] (9)

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

The tree model is trained with the objective function \( u \) as given in the following equation:

\[
u = \sum_i o(\hat{y}_i, y_i) + \sum_k \Omega(f_k),
\] (10)

where \( o \) is the convex loss function to measure the difference between the real value \( y_i \) and the predicted value \( \hat{y}_i \) and \( \Omega \) stands for penalty term, and its expression is as follows:

\[
\Omega(\cdot) = y_i + \frac{1}{2} \| \phi \|.
\] (11)
where $\|\phi\|$ represents the regular term and $\gamma$ stands for leaf node penalty, which is mainly used to avoid overfitting problems.

In the process of credit risk identification of Internet financial enterprises, the European space cannot be directly used to optimize the objective function. Therefore, the credit risk identification model is trained through boosting learning strategy model, and the specific process is as follows:

\[
B_i^{(0)} = 0, \\
B_i^{(1)} = f_1(x_i) = B_i^{(0)} + f_1(x_i), \\
B_i^{(2)} = f_1(x_i) + f_2(x_i), \\
\vdots \\
B_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = B_i^{(t-1)} + f_1(x_i),
\]

where $B_i^{(t)}$ represents the function newly added to Round $t$.

Based on the above process, the objective function can be converted as

\[
\nu = \sum_{i=1}^{n} \left( B_i B_i^{(t)} + f_i(x_i) \right) + \Omega(f_i) + \alpha,
\]

where $\alpha$ is a constant.

The fitting results of the model and training data in the identification process can be measured by the loss function $L = \sum_{i=1}^{n} l(B_i, B)$, in which the logical loss function $l(B_i, B) = y_i \ln(1 + e^{-B})$ and the square loss function $l(B_i, B) = (B_i - B)^2$ are widely used in the identification process. The credit risk identification model of Internet financial enterprises based on the RB-XGBoost algorithm is brought into the square loss function in the objective function and the following equation is obtained.

\[
\nu = \sum_{i=1}^{n} \left( B_i B_i^{(t-1)} + f_i(x_i) \right)^2 + \Omega(f_i) + \alpha,
\]

where $B_i^{(t-1)}$ stands for residual.

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

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

When the loss function is a square loss during training, the following equation can be computed.

\[
\begin{cases}
g_i = \partial_{B_i} l(B_i, B_i^{(t-1)}) = 2(B_i^{(t-1)} - B_i), \\
h_i = \partial_{B_i} \partial_{B_i} l(B_i, B_i^{(t-1)}) = \partial_{B_i} \partial_{B_i} (B_i^{(t-1)} - B_i)^2.
\end{cases}
\]

Substitution of parameters $g_i$ and $h_i$ in the objective function yields the following expression:

\[
\nu = \sum_{i=1}^{n} \left( B_i - \left( B_i^{(t-1)} + f_i(x_i) \right) \right)^2 + \Omega(f) + \alpha,
\]

where $B_i^{(t-1)}$ represents the output of the model in Round $t - 1$ training and $B_i$ represents the dependent variable in the objective function, and if $B_i$ is known, the above objective function [20] can be simplified to obtain the following expression:

\[
\nu = \sum_{i=1}^{n} \left( g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right) + \Omega(f) + \alpha,
\]

where $g_i$ and $h_i$ are the parameters in the loss function. In different loss functions, the values of the above parameters are different; therefore, the values of $g_i$, $h_i$ are determined by the loss function. Hence, each tree is redefined by the following equation.

\[
f_i(x) = \omega q(x), \quad \omega \in \mathbb{R}^T, \quad q: R^d = \{1, 2, \ldots, T\},
\]

where $\omega$ represents the weights of leaf nodes in the tree structure, $\omega q(x)$ denotes the predicted values obtained by the tree model, and $q: R^d = \{1, 2, \ldots, T\}$ shows the tree structure.

**Model complexity includes two parts:** L2 regularization of leaf node score and a total number $T$ of leaf nodes. Model complexity $\Omega(f_i)$ can be obtained by tree definition:

\[
\Omega(f_i) = \gamma T + \frac{1}{2} \lambda \sum_{x=1}^{T} \omega_j^2.
\]

The smoothness of leaf nodes can be improved by L2 regularization to solve the overfitting problem. When the complexity of the model increases, there are two different types of accumulation, one of which is $I_j = \{lq(x_i) = j\}$, where $I_j$ represents the set of samples in a leaf node $j$. After adding complexity to the objective function, the final objective function, that is, the credit risk identification model of Internet financial enterprises, is obtained:

\[
\mu = \sum_{j=1}^{n} \left[ g_j f_i(x_i) + h_j f_i^2(x_i) \right] - \lambda \sum_{j=1}^{T} \omega_j^2.
\]

The credit risk identification model of Internet financial enterprises constructed above is used to complete risk identification.

**4. Experimental Results and Analysis**

To examine the credit risk evaluation ability of the proposed credit risk identification model of Internet financial enterprises, help the enterprises to avoid the risk of electricity arrears, and provide the basis for urging the payment of electricity fees, five Internet financial enterprises in a certain city are selected as experimental objects, and the annual reports of these five companies in the recent three years are
selected as experimental data samples to carry out the experimental analysis.

4.1. Accuracy Test of Credit Risk Assessment with Different Methods. To evaluate the credit risk of five selected Internet financial enterprises, the proposed method is used along with the methods given in [3, 4], and the results of these methods are compared with the actual credit risk of the 5 companies. The experimental results are shown in Figure 4.

It can be seen that the credit risk grade scores of the five enterprises evaluated by the proposed method are closer to the actual credit risk grades scores of the five enterprises. This is because the proposed method can effectively combine the actual and objective data of the market transaction data of five companies to obtain the credit status of the enterprises. In addition, the proposed method can accurately obtain detailed information of the credit risk of the enterprises and establish an accurate credit risk evaluation index system, which makes the results of credit risk evaluation more scientific and accurate. The comparison results for the credit risk assessment accuracy of the two methods are shown in Figure 5.

By analyzing Figures 4 and 5, we can see that this method can effectively measure and evaluate all credit risk indicators and obtain the credit risk of enterprises. When the number of samples reaches 20000, the credit risk evaluation accuracy of references [3, 4] is 0.50% and 0.30%, respectively, whereas the accuracy of credit risk evaluation of the proposed method reaches 0.92%.

4.2. Kappa Coefficient and ROC Curve Test of Different Methods. To verify the recognition accuracies of all the three methods, the kappa coefficient and ROC curve are used. The kappa coefficient can weigh the difference between the predicted result and the real result. The kappa coefficient $K$ can be computed as

$$K = \frac{P_o - P_e}{1 - P_e},$$

where $P_o$ represents the proportion of correctly identified samples in the total number of samples and $P_e$ is the randomness ratio. The higher the kappa coefficient $K$ is, the more accurate is the recognition result of the method. The kappa coefficients of the proposed method, the method in reference [3], and the method in reference [4] are shown in Table 1.

It is evident that the kappa coefficients of the proposed method are higher than those of reference [3] and reference [4] in multiple iterations, which indicates that the proposed method can accurately identify the credit risks of Internet financial enterprises. This is because the method constructs a risk identification index system based on the data with high
balance and completes the identification of credit risks of Internet financial enterprises based on the high-precision risk identification indexes.

In the ROC curve, the abscissa is the real case rate and the ordinate is the false-positive case rate. The larger the area enclosed by the ROC curve and the abscissa is, the higher is the recognition accuracy of the method. The proposed method and the methods given in reference [3] and reference [4] are used to identify the credit risks of different Internet financial enterprises, and the obtained ROC curve is shown in Figure 6.

Figure 6 shows that the area of ROC curves and abscissa obtained by the proposed method is larger than the area of ROC curves and abscissa obtained by the methods proposed in [3, 4], indicating that the accuracy of the proposed method is higher, and the identification of credit risks can be accurately completed in the Internet financial enterprises.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.99</td>
<td>0.84</td>
<td>0.69</td>
</tr>
<tr>
<td>2000</td>
<td>0.98</td>
<td>0.61</td>
<td>0.74</td>
</tr>
<tr>
<td>3000</td>
<td>0.97</td>
<td>0.66</td>
<td>0.66</td>
</tr>
<tr>
<td>4000</td>
<td>0.97</td>
<td>0.77</td>
<td>0.72</td>
</tr>
<tr>
<td>5000</td>
<td>0.98</td>
<td>0.72</td>
<td>0.81</td>
</tr>
</tbody>
</table>

**Table 1: Kappa coefficients of different methods.**

5. Conclusion

The rapid development of Internet finance provides new financing channels for the development of small and microenterprises and individual entrepreneurship. The conventional methods of credit risk prediction of Internet financial enterprises cannot get the characteristics of credit risk zoning, leading to large errors in the results of credit risk identification. In this study, a new method of credit risk identification based on big data for Internet financial enterprises is proposed. The risk evaluation steps of Internet financial enterprises are studied, and the importance of assessment indicators is measured using the improved AHP method. The linear weighted synthesis method is employed to systematically assess the credit of clients. Based on the unique characteristics of big data credit risk region division, the big data credit risk is predicted with the help of rule-based matching method. The XGBoost supervised machine
learning algorithm is used to develop a credit risk prediction model of Internet financial enterprises. The performance of the model is evaluated with the kappa coefficient and ROC curve. Experimental results show that the proposed method can correctly assess the credit risk of Internet financial enterprises.

**Data Availability**

The data used to support the findings of this study are available from the corresponding author upon request.

**Conflicts of Interest**

The author declares that there are no conflicts of interest.

**References**


