Risk Analysis of Online Lending Platform Based on Bigdata Technology

Guang-xuan CHEN\textsuperscript{1,2,a}, Li-ping DING\textsuperscript{2,b*}, Guang-xiao CHEN\textsuperscript{3}, Pan-ke QIN\textsuperscript{4},
Guo-ming ZHOU\textsuperscript{1} and Qiang LIU\textsuperscript{1}

\textsuperscript{1}Zhejiang Police College, Hangzhou, Zhejiang, China
\textsuperscript{2}Digital Forensics Lab, Institute of Software Application Technology, Guangzhou & Chinese Academy of Sciences, Guangzhou, Guangdong, China
\textsuperscript{3}Universidad Carlos III de Madrid, Madrid, Spain
\textsuperscript{4}Henan Polytechnic University, Jiaozuo, Henan, China
\textsuperscript{a}ericcgx@163.com, \textsuperscript{b}liping@nfs.iscas.ac.cn
*Corresponding author

Keywords: Online Lending, Prediction Model, Risk Warning, Bigdata.

Abstract. In recent years, online lending platforms have emerged a lot of insecurity events. How to judge the security state and health degree of the online lending platforms are becoming the most concerned problems for the investors and supervision departments. In this paper, we’ve adopted the massive data acquisition technology, Spark distributed computing, machine learning and other bigdata technologies to analyze the risks of the online lending platform. Then, a risk early warning model based on factor analysis method was proposed. Practice has proved that the model can help to improve the accuracy of the risk prediction to online lending platforms.

Introduction

In recent years, the Internet finance in China is developing rapidly. On the one hand, the development of Internet finance can well meet the investment and financing needs of small and medium enterprises, innovative enterprises and the individuals with low and middle income, and can also create a favorable capital environment for public innovation. On the other hand, due to the lack of supervision, Internet finance in the process of innovation and development also exposed a lot of problems and hidden dangers. For example, being an important part of Internet finance, the P2P online lending industry is exposing even more pressing issues that threatened the national economic security. According to the Net Loan Home, the amount of P2P financing reached 982.5 billion yuan in 2015, increased by 389%. And by the end of June 2015, the number of P2P online lending platform reached to 2814, include 786 problematic platforms which accounted for 27.93%.

With the set of a series of rules, the ferociously growing online lending industry without supervision will reshuffle rapidly and a lot of lending platforms that can’t meet the rules will be eliminated, which will lead to the emergence of illegal fund-raising, bankrupt, runaway and other serious security events. In this context, it is of great significant and urgent for the majority of the investors and the supervision departments of the government to effectively monitor the online lending platform and make early warning when problems were detected.

Related Works

At present, the research on the risk and early warning of online lending platform is mainly focused on the following two parts: (1) Theoretical research on risk and supervision and early warning of P2P online lending platform. Huang and Qi [1] introduced that the P2P online lending is facing the risks caused by imperfect legal system and the platform’s own operation condition and network technology. Hu and Meng [2] further analyzed the risks faced by the P2P online lending platforms based on
system science theory. They thought that the environment has pressure to the system, i.e., the lending platform will be influenced by the risks as financial crisis, lack of industry laws, imperfect self-regulation, system security vulnerabilities, and imperfect guarantee institution and credit mechanism. Then, Wang and Liu thought that the online lending industry mainly exist the risks of unsound credit assessment and business regulation and insecure system. Based on this, they put forward that the introduction of big-data technology into P2P risk control will not only broaden the types of data available to financial institutions, but also provide a new solution to the credit risk control of P2P online lending [3].

(2) Empirical research on risk early warning model of P2P online lending platform. Based on the sample data of 2013-2014, the logistic regression model was used to study the financial distress of the platform from the three aspects of enterprise nature, profit rate and risk management guarantee model [4]. The results showed that it is positively correlated with the short-term yield. Ma [5][6] analyzed the risk of P2P online lending platform, constructed an index system including credit rating, mobility, information transparency, technical service, brand and leverage rate and established a risk early warning model based on principal component analysis and improved KLR signal analysis method. The scoring and ranking to 20 online lending platforms verified the feasibility and accuracy of the model. However, the modeling process is mainly based on small sample data, so the accuracy and scientific of the conclusion is still to be further tested [7][8].

In short, the risk assessment of the online lending platform is a complex problem of cross-domain and multi-data source. Multidimensional data acquisition and multi-angle feature analysis are the important guarantees that can ensure the final model complete accurate early warning. Based on the features of bigdata as large data volume, various types, high speed and high efficiency, it can greatly broaden the feature fields of the historical data for the final model training. So in this paper, we introduced bigdata and related technology to research monitoring and early warning model for the online lending platform.

**Risk Assumptions and Analysis**

Through a comprehensive analysis of the reasons of the risks of the online lending platforms, this paper proposed the following four assumptions:

**Positive correlation between abnormity of operational data and risk**

The operating data of the online lending platform mainly include number of borrowers, amount of loans, expected rate of return, platform object, and so on. These operating data usually have a reasonable value range in the industry. The platform will lead to fault when some of the data is far from the value range. For example, when the expected rate of return far higher than the average of the industry while the number of the loan object showed in the platform is small, thus the platform is likely to appear the phenomenon of “cash pool”. Platform operating data is the most direct indicator associated with its risk, the more the data deviates from the reasonable range, the greater the risk it faces.

**Positive correlation between negative public opinion and risk**

Online public opinion refers to a collection of the cognitive, attitudinal, emotional and behavioral tendencies that are generated by the stimulation of various events and spread online. Online public opinion comes from the real world and at the same time it has positive or negative reaction to the real world. Especially, the Internet will quickly make some sensitive events into public hot topic. Therefore, based on the characteristics as timeliness and generality of the online public opinion, it is effective to put the ratio of the negative online public opinion of the online lending platform as its risk early warning indicator.

In this paper, the relative information of text format was collected from the industry related websites, forums and microblogs and then the negative information (illegal, false propaganda,
complaints) contained was sorted out through text classification. The negative information can timely and comprehensively reveal the current problems of the platform. The more problems, the greater risks the platform faced with.

**Negative correlation between the credit of the platform and risk**

Improper operation and malicious fund-raising fraud are the reasons to the bankruptcy of the online lending platform. In reality, many platforms use false propaganda as risk-free or high-yield to attract customers while their real aim is to establish the fund pool for their own. Search the credit information and complaints of the platform affiliated enterprises and related legal person in the searching platform can help judge the risk of the platform. The higher credit and less complaint, the lower the risk of the platform is.

**Negative correlation between the background strength and risk**

The background strength of the online lending platform mainly includes the amount of registration and the actual amount paid, cooperative guarantee institution, fund trustee, the background of the affiliated enterprises, and so on. Some platforms with background of state-owned listed companies usually do not exist illegal fund-raising, false propaganda and other vicious incidents. In addition, their financial and management team has a certain advantage that can better cope with the overdue and bad debts. So, the stronger the background of the platform, the lower the risk it will face.

In the following modeling, in the next modeling process, the corresponding indicator field will be established to verify each hypothesis introduced above.

**Strategy for Risk Early Warning Modeling**

Here, we adopt Factor analysis to establish the risk early warning model. Based on the idea of dimension reduction, divide the variables into different group according to the size of the correlation of the variables, making the variables in the same group with high correlation and low correlation in different groups. This also converted the variables with complex correlations into a few comprehensive factors possessing statistical methods with strong fitting ability. The advantages of factor analysis method include: (1) simplified the complex set of primitive variables into a few comprehensive factors and preserved the interpretability of the original variables as much as possible. (2) The factors are become more interpretable through oblique rotation and the commonality of the principal factor become even more clearly. The advantages of factor analysis above made the method widely used. Here, we adopted the factor analysis method to construct risk forecasting mechanism for the online lending platform.

The specific factor analysis can be expressed as:

\[ X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \epsilon_i \quad (m \leq p) \]  \hspace{1cm} (1)

and then simplified as:

\[ X = AF + \epsilon \]  \hspace{1cm} (2)

Here, \( X_i \) includes \( p \) primitive variables (\( i = 1, 2, \ldots, p \)), \( F_1, F_2, \ldots, F_m \) refers to common factors, \( A \) is factor loading matrix and \( \epsilon \) is specific factors [9][10]. The basic steps of the algorithm are:

- Verify that the selected primitive variables are appropriate for the factor analysis. As for the calculation of correlation coefficient matrix of the primitive variables, correlation coefficients that are less than 0.3 are not suitable for the factor analysis.
- By calculating the factor load \( A \) to construct the factor variable, and then conduct varimax rotation on \( A \).
\[
A = \begin{bmatrix}
    a_{11} & a_{12} & \cdots & a_{1m} \\
    a_{21} & a_{22} & \cdots & a_{2m} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{p1} & a_{p2} & \cdots & a_{pm}
\end{bmatrix} = \begin{bmatrix}
    u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1m}\sqrt{\lambda_p} \\
    u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2m}\sqrt{\lambda_p} \\
    \vdots & \vdots & \ddots & \vdots \\
    u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & u_{pm}\sqrt{\lambda_p}
\end{bmatrix}.
\]

- Calculate the factor variable scores for each sample and conduct comprehensive evaluation to the variable system.

\[F_j = \beta_{j1x1} + \beta_{j2x2} + \cdots + \beta_{jmxm} (j=1,2,\ldots,m),\]  

Comprehensive evaluation function is

\[F = \alpha_1 F_1 + \alpha_2 F_2 + \cdots + \alpha_m F_m.\]

Here, \(\alpha_i\) refers the variance contribution of \(F_i\).

Then, the whole model is constructed as follows: Firstly, a large number of raw data are collected by different methods. Then, a series of data preprocessing, such as missing value repair and outlier detection, are required to unify the original data format to meet the training model. And then the processed data is divided into training samples and test samples in two parts, the training samples into different models for machine learning; Finally, the use of test samples to verify the accuracy of different models, and by adding or subtracting the original field and further preprocessing to continually optimize and improve the final risk warning model. The specific process is shown in Fig. 1:

**Figure 1. Modeling of Risk Early Warning.**

**Data Process and Model Construction**

The poor operation of the platform itself and malicious fraud are the most important factors that contribute the risks of the platform. This can be reflected on the relative data as platform running information, credit record, complains, positive or negative online public opinion, and so on. Therefore, collecting and sorting these data is the main task in the data acquisition phase.

Then, the data preprocess will conducted so as to guarantee the data be in the good condition for the next phase. Data preprocess usually includes text information processing, outlier detection, data consistency processing, data conversion and datasets with missing values. Based on the Spark distributed computing platform, a variety of models were chosen to train the sample set using machine learning method. Then the accuracy will be tested through the sample set. Finally, an optimal risk
early warning model will constructed through the repeated adjustment. Here, artificial neural network model and Bayesian network model are the most adopted models.

Summary

The rapid development of Internet finance in China have brought great conveniency and benefits to the public. However, it also exposed a lot of problems and hidden dangers because of the lack of supervision. In order to help the investors and supervision departments to judge the security state and health degree of the online lending platform, we proposed a risk early warning model based on factor analysis method. Massive data acquisition technology, Spark distributed computing, machine learning and other bigdata technologies were adopted in the modeling and analysis. Practice has showed that the early warning model can effectively help us to prediction the risk of the online lending platforms. In the future, we will to optimize the model so as to improve the accuracy in the filed.

Acknowledgement

This research was financially supported by the Scientific Research Project of Zhejiang Police College (NO. 20150615), Educational Technology Planning Program of Zhejiang Province (NO. JB058), Research Program of University Laboratory of Zhejiang Province (NO. YB201536), NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization (NO. U1509219), and the National High Technology Research and Development Program of China (No. 2015AA016003).

References