P2P Online Loan Willingness Prediction and Influencing Factors Analysis Based on Factor Analysis and XGBoost

Dongyang Zhang¹*, Yicheng Gong¹, Li Yu¹ and Xiaojie Wang¹

¹Wuhan University of Science and Technology School of Science, Wuhan, China

*Corresponding author email: 1044600738@qq.com

Abstract. This paper selects the open dataset of the auction, combines the factor analysis with XGBoost, analyzes the dimension of the pre-processed high-dimensional data with factor analysis, and then uses XGBoost to predict the investment level of P2P online loan investors, and analyzes and quantifies the willingness to invest. The relationship with specific influencing factors. The main factors affecting the willingness to contribute through the ranking of the characteristic factors of the factor analysis and the reduction of the common factors include: the borrowing rate and the initial rating, the real name certification on the borrower's mobile phone, and the repaid principal and interest paid. The model prediction results show that the XGBoost method combined with factor analysis predicts that the error rate of the investment level on the training set is train-merror: 0.03829, and the error rate on the test set is test-merror: 0.050386. Compared with the method using XGBoost alone, the prediction error rate based on factor analysis and XGBoost on the training set and the test set is reduced by 0.020706 and 0.0152, respectively, which significantly reduces the error rate; and shortens the operation time by 18 seconds. The conclusions obtained in this paper provide reference for the stable development of the P2P industry and the decision-making of investors.

Keywords: Willingness to invest; Incomplete information; Influencing factors; Factor analysis; XGBoost.

1. Introduction

Under the Internet financial system with incomplete information, as the main body of online loan funding, the investor's rationality of funding behavior and the factors that affect the investor's willingness to invest are related to whether the online loan industry can develop smoothly and affect online The future development direction of loans[1]. Since the amount of funds contributed by investors will be affected by various factors to varying degrees, scientific research on the willingness of investors to invest has attracted widespread attention in the academic community[2]. Li Shujin and Chen Da[3] (2018) selected 22 indicators affecting borrowers' default from the macro and micro levels through theoretical analysis, used survival analysis to predict the probability of default of P2P platform borrowers, and verified the P2P platform borrowing The assumption that human default risk will change with macroeconomic fluctuations. Zhang Jie and Zhang Yuansheng [4] (2019) established a set of factor systems for risk assessment of P2P online loan platforms by using a factor analysis method, and conducted a risk evaluation ranking of 80 mainstream P2P online loan platforms in China, selecting safe and reliable investors The platform provides a reference for investment. Zhang Shun, Zhao Cuicui, Yang Li[5] (2019) The research results based on the transaction data of "everyone loan" show that the characteristics of borrowers "urgent" and "poor" will reduce the success rate of borrowing and increase their borrowing costs. Qi [6] (2020) analyzed the information of the P2P network lending platform operator can decide which information is visible in the loan request based on the
information availability theory, which may affect the investor's investment behavior. Hu Jinzhang and Song Weishi[7] (2017) explored whether online lending investors have rational consciousness to weigh the benefits and risks through the expected utility theory. Zheng Yingfei and Chen Xiaojing[8] (2018) mainly analyzed the existence of herd behavior in the group of investors, and concluded that the platform's commitment to guarantee the principal and interest will have a significant impact on the investor's investment decision. Li Yi[9] (2018) combined the regression model and principal component analysis to analyze the loan list data of the online loan platform, and obtained the influencing factors of the investor's willingness to contribute. Deng Dongsheng and Chen Zhao[10] (2019) combined the market risk theory to analyze the influence of individual platform risks and the overall market risk of the industry on the willingness of investors to invest. Liao Li, Xiang Jia, and Wang Zhengwei[11] (2018) used the transaction data of a P2P lending platform in China, and used a herd behavior indicator to show that group wisdom is beneficial for the prediction of the default rate.

In order to better study the main body of P2P online lending under incomplete information, this paper selects the investor's willingness to invest as a research perspective, combines factor analysis with XGBoost[12](2016), predicts the willingness of online loan investors to invest, and analyzes the impact of willingness The main factor.

2. Introduction to Factor Analysis and XGBoost

2.1. Basic Principles of Factor Analysis

The concept of factor analysis originated from the statistical analysis of intelligence tests by Karl Pearson and Charles Spearman in the early twentieth century. The name "factor analysis" was first proposed by Thurstones in 1931 and is a classic method of data dimensionality reduction. The calculation process of factor analysis can be carried out according to the following 9 steps.

Step1. Standardize the original data to eliminate differences in magnitude and dimension between variables.

Step2. Find the correlation matrix of the standardized data.

Step3. Find the eigenvalues and eigenvectors of the correlation matrix.

Step4. Calculate the variance contribution rate and cumulative variance contribution rate.

Step5. Determining factors: Let \( F_1, F_2,..., F_p \) be \( p \) factors, where the previous factor contains a total of data information.

Step6. Factor rotation: If the obtained factors cannot be determined or their actual meaning is not obvious, then the factors need to be rotated to obtain more obvious practical meaning.

Step7. Use the linear combination of the original indicators to obtain the scores of each factor.

Step8. Comprehensive score: The variance contribution rate of each factor is taken as a weight, and a comprehensive evaluation index function is obtained from the linear combination of each factor.

Step9. Sorting of scores: Use the comprehensive score analysis to get the score ranking.

2.2. Basic Principles of XGBoost

The full name of XGBoost is eXtreme Gradient Boosting. It was proposed by Dr. Tianqi Chen of the University of Washington in 2014. It is used in the Higgs sub-signal recognition competition in Kaggle, and has attracted wide attention due to its superior efficiency and high prediction accuracy. XGBoost's reasoning process can be summarized as the following 4 steps:

Step1. The definition of the objective function of XGBoost is shown in equation (1).

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

(1)

Step2. Use Taylor expansion to define an approximate objective function as shown in equation (2).

\[
\text{Obj}(t) \approx \sum_{i=1}^{n} \left[ l(y_i, \hat{y}_i(t-1)) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \Omega(f_t) + c
\]  

(2)

Step3. Refine the definition of \( f_t \) and split the tree into structural parts and leaf weight parts.

Step4. The objective function contains a single independent quadratic function to find the smallest
This article combines factor analysis in multivariate statistical analysis methods and XGBoost in machine learning methods to analyze relevant information issued by online loan platforms held by investors under the p2p online loan system under incomplete information. Willingness to contribute and its influencing factors.

3. Dimensionality Reduction of Investor's Characteristic Variables Based on Factor Analysis

In this paper, factor analysis is used to reduce the dimensions, and the pre-processed data is processed by SPSS. The results are shown in Table 1.

Table 1. Explanation of variance.

| Variable | Initial eigenvalue | Cumulative Sum Loading |
|----------|--------------------|-----------------------|
|          | total              | cumulative%           |
| 1        | 4.460              | 16.518                |
| 2        | 2.782              | 10.305                |
| ...      | ...                | ...                   |
| 15       | 0.644              | 2.386                 |

From the variance interpretation table in Table 1, it can be seen that the sum of the contributions of the first fifteen common factors to the cumulative variance of the sample is 85.133%, so the first fifteen common factors are selected to establish the factor loading matrix.

In order to facilitate subsequent processing, the fifteen common factors in Table 2 are renamed FACT1, ..., FACT15. Observe Table 2. With the load value greater than 0.85 as the standard, the public factor FACT1 is mainly determined by (x1) Browrowing. interest. rate, (x2) Initial. rating; the public factor FACT2 is at (x6) Pending. principal, (x7 ) Interest.to.be.paid is determined by two indicators; ... ; The public factor FACT15 is determined by the (x27) academic. certificate indicator.

4. Prediction of Investor's Investment Grade Based on XGBoost

4.1. Learning Results and Conclusions of XGBoost Combined with Factor Analysis

This paper uses python to reduce the dimensionality of factor analysis. The data obtained based on the XGBoost model is used to predict the funding level of investors. When dividing the training set and the test set, random numbers were set, and the training set was selected by python: the ratio of the test set was 8: 2. The training set had 234031 samples and the test set had 58508 samples. Table 2 shows the partial learning results based on XGBoost on the test set.

Table 2. Partial learning results based on XGBoost on the test set.

| Sample number | Predicted investor funding level | Investor's actual funding level |
|---------------|---------------------------------|--------------------------------|
| 1             | D                               | D                              |
| 2             | A                               | A                              |
| ...           | ...                             | ...                            |
| 58508         | E                               | E                              |

After training the training set, test the results on the test set. Comparing the second and third columns in Table 2, the error rate of the training set is train-meanerror: 0.03829 The error rate of the test set is test-meanerror: 0.050386.

Comparing the second and third columns of Table 2, we can be calculated

\[
\text{Error rate} = \frac{\text{Number of samples with different prediction categories}}{\text{Total number of test set samples}} = 0.050386
\]

After factor analysis, the training time of XBGoost is 6 minutes and 52 seconds. Visualize the results of XBGoost's prediction of the investor's funding level. A part of it was intercepted for specific analysis, as shown in Figure 1.
4.2 Learning Results Using XGBoost Alone

In order to compare with the learning results of XGBoost combined with factor analysis, this paper also made the learning results using XGBoost alone.

After training the training set, the test result on the test set, the error rate of the training set is train-erreur: 0.058996. The error rate of the test set is test-erreur: 0.065586. Similarly, based on the learning results, we can calculate

\[
\text{Error rate} = \frac{\text{Number of samples with different prediction categories}}{\text{Total number of test set samples}} = 0.065586
\]

The training time using XGBoost alone is 7 minutes and 10 seconds.

Compared with the method using XGBoost alone, the method using factor analysis plus XGBoost significantly reduces the error rate of the prediction results, and the error rate of the prediction results of the training set and the test set is reduced by 0.020706 and 0.0152 respectively; in addition, the model running time is shortened It took 18 seconds.

5. Conclusion and Outlook

This paper uses factor analysis plus XGBoost to predict the willingness to invest and the factors that affect the investment on a p2p online loan platform with incomplete information. The error rate of the training set predicted by the model established in this paper is only train-erreur: 0.03829. The error rate of the test set is test-erreur: 0.050386, and the main factors affecting the funding are the borrowing rate and initial rating, Real-name authentication on the borrower's mobile phone, as well as the principal and interest repaid.

This article only combines the two learning methods to get the results. Future research can consider combining other statistical learning methods with machine learning methods to form a comparison to find the optimal combination. In addition, the investment behavior of capital contribution is a kind of individual decision-making behavior. Future research can consider combining the investor's capital contribution decision-making with game theory to guide the decision-making from the perspective of the game.

Acknowledgments

This work was supported by two funds: 1. Hubei Province College Students' Innovative Entrepreneurial Training Plan Program (201810488097) 2. Hubei Province Key Laboratory of Systems Science in Metallurgical Process (Y201906).
References

[1] Yu Song, Yuyan Wang, Xin Ye, Dujuan Wang, Yunqiang Yin, Yanzhang Wang. Multi-view ensemble learning based on distance-to-model and adaptive clustering for imbalanced credit risk assessment in P2P lending[J]. Information Sciences, 2020, 525.

[2] Golnoosh Babaei, Shahrooz Bamdad. A multi-objective instance-based decision support system for investment recommendation in peer-to-peer lending[J]. Expert Systems With Applications, 2020, 150.

[3] Li Shujin, Chen Da. Credit risk assessment of borrowers on China's P2P lending platform [J]. Journal of Hangzhou Dianzi University (Social Science Edition), 2018, 14 (03): 1-9.

[4] Zhang Jie, Zhang Yuansheng. Research on Risk Evaluation of China's P2P Online Loan Platform Based on Factor Analysis [J]. Friend of Accounting, 2019 (07): 23-27.

[5] Zhang Shun, Zhao Cuicui, Yang Li. P2P online lending is urgent to save the poor? — Empirical evidence based on "everyone loan" [J / OL]. Southern Finance: 1-9

[6] Qi Wang, Xiong Xiong, Zunxin Zheng. Platform Characteristics and Online Peer-to-Peer Lending: Evidence from China[J]. Finance Research Letters, 2020.

[7] Hu Jinzhang, Song Weishi. Investors' Rational Consciousness and Weighing Behavior in P2P Lending: An Empirical Analysis Based on "Everyone's Loan" Data [J]. Financial Research, 2017 (07): 86-104.

[8] Zheng Yingfei, Chen Xiaojing. P2P online loan platform principal and interest protection and investor response [J]. Finance Forum, 2018, 23 (01): 66-80.

[9] Li Yi. Empirical research on influencing factors of loan willingness of P2P online loan platform [D].

[10] Deng Dongsheng, Chen Zhao. Internet financial risk and investor risk awareness: Evidence from transaction data of online loan platform [J]. Finance and Economics, 2019, 40 (02): 101-113.

[11] Liao Li, Xiang Jia, Wang Zhengwei. Group Wisdom of P2P Lending Investors [J]. China Management Science, 2018, 26 (10): 30-40.

[12] Chen T Q, Guestrin C. XGBoost: A Scalable Tree Boosting System, 2016 [C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.