Research on Credit Rating Model of P2P Project Based on Light GBM Algorithms

Pengcheng Sun

Economics and Management School, Jilin Agricultural Science and Technology University, Jilin, 132101, China

Keywords: Light GBM Algorithm; P2P; Credit Rating

Abstract: Under the background of the development of big data and Internet finance, according to personal credit, we can effectively control the default rate of P2P projects to ensure the good operation of related financial projects or platforms. From the point of view of credit risk of P2P platform, taking the risk control of borrowers as the research objective, this paper constructs the evaluation index system of borrowers' credit, and establishes the evaluation model of borrowers' credit risk based on Light GBM algorithm using the data of P2P network lending platform. It has been proved by practice that the model can predict the credit risk of P2P network credit borrowers well and has high-precision classification ability. At the same time, based on the results of the Light GBM algorithm to determine the factors affecting the default rate, the improvement content of the P2P platform and the development direction of the countries in this field can be clarified.

1. Introduction

With the development of Internet technology and the arrival of the era of big data, Internet finance has been developed rapidly and attached great importance to by the state. To do a good job of credit rating is the key to ensure the healthy development of Internet finance [1] As a new lending mode, P2P credit has its own advantages. Firstly, its loan threshold is low. Borrowers can issue loan targets directly on the P2P credit platform by providing part of their own information and identity information. Secondly, its income is several times that of bank term deposits [2]. The emergence of Internet finance is in line with the requirements of social informationization, and also provides favorable conditions for the development of P2P network lending. The rapid development of the industry and the risks are also exposed. The default of the borrower has caused serious losses to the P2P network lending platform and lenders. Personal credit risk has become the main risk faced by the P2P network lending platform [3]. The essence of the credit risk caused by the P2P network lending platform is the credit risk caused by the information that the investor knows and the real information asymmetry of the borrower, which means that the borrower is unwilling or unable to perform the contract treaty and causes the default, resulting in the investor suffering. Loss, the P2P platform cannot bear the risk of default. This paper is an attempt of Light GBM machine learning method in the field of P2P risk management control. Such an attempt has certain academic significance and practical value, and can provide strong technical support for the development of China's P2P industry, and thus promote China's Internet finance industry. Great development.

2. Methodology

Light GBM is an open source, fast and efficient decision tree algorithm based promotion framework published by Microsoft Research Asia. It is used for sorting, classification, regression and other machine learning tasks to support efficient parallel training. Decision trees are a method of classification and regression, and most of the actual research is used for classification. The structure of the decision tree is a tree structure. Most of them use the second eucalyptus tree. On each leaf node, according to a certain judgment condition, the two categories of "eligible condition" and "non-conformity" are output, and the output is repeated repeatedly. With the continuous progress of financial market and the establishment and improvement of personal credit information system, more

  DOI: 10.25236/mmmce.2019.070

and more investors will recognize the mode of P2P network lending and invest in P2P network lending. In order to increase the stability of the model and simplify the calculation and application, this study classifies the alternative indicators [4]. Credit rating, group membership, borrowing purposes, borrowing period, and whether there are real estate can be grouped directly according to the eigenvalues. These behavioral information have the potential to objectively reveal users' own characteristics and predict their behavior. By analyzing them with Light GBM, we can objectively evaluate personal credit level.

Figure 1 below. Decision tree can be understood as a set of many if-then rules, or as a conditional probability distribution defined in a particular space and class space. The decision tree creation includes three main steps: feature selection, decision tree generation and decision tree pruning. This method has the advantages of high readability and fast classification.

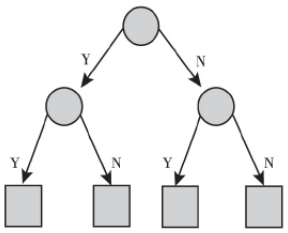

Fig.1. Decision tree structure

The idea of Gradient Boosting is: one-time iteration of variables, the sub-model is added one by one in the iterative process, and the loss function is guaranteed to decrease. Assuming fi(x) is a submodel, the composite model is:

$$F_m(X) = \partial_0 f_0(X) + \partial_1 f_1(X) + \cdots + \partial_m f_m(X) \tag{1}$$

The loss function is L [Fm (X), Y]. Each time a new sub-model is added, the gradient of the loss function towards the variable with the second highest information content decreases.

$$L[F_m(X), Y] < L[F_{m-1}(X), Y] \tag{2}$$

The credit scoring method is a major leap from the subjective judgment to the quantitative analysis, which greatly improves the objectivity and standard of the rating and reduces the rating cost. It not only has many advantages of multi-parameter and non-parametric statistics, but also makes full use of the existing prior information to effectively process the non-homogeneous relationship between data. It is easy to operate, can handle many problems in many fields, and has many Robustness and other characteristics. No matter how cautious the data is collected, the errors in the data cannot be completely eliminated, so we must check the quality of the data before conducting an empirical analysis. The stronger the repayment ability of the borrower, the lower the risk of default [5]. The longer the borrower works, the more familiar he is with his own industry or work. Similar to linear regression analysis, the basic principle of logistic regression analysis is to use a set of data to fit a logistic regression model, and then use this model to reveal the relationship between several independent variables in the population and the probability of a dependent variable taking a certain value. The advantage of this method is that it can make the algorithm faster and more effective; the disadvantage is that the information on abandoned leaves will be ignored, resulting in the splitting results are not refined enough. Figure 2 depicts the process of leaf division.

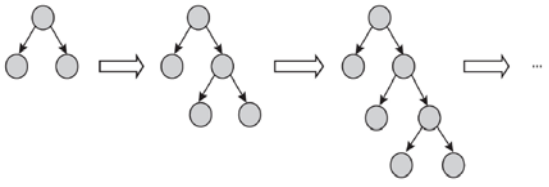

Fig.2. Decision tree learning process by leaf splitting

In order to improve the evaluation ability of the model, this study has done an appropriate

treatment of the value of the independent variable, and replaces the original value with the WOE value. This model is a method for comprehensively considering the credit risk of borrowers and the risk of credit instruments to determine the risk of loans and making credit decisions. It is one of the credit risk assessment models commonly used by banks in China. In addition, it is more important that the decision tree has the "black box" feature that is different from most other data mining methods. It can display the analysis process and the results of the analysis in a graphical form, which is convenient for human beings. Understanding [6]. It does not need to select the node that maximizes the revenue, and each node in each layer is split, that is to say, every iteration traverses all the data of the whole training data. The advantage is that each layer of leaves can be completed in parallel, with natural parallelism. From the model, it can be seen that whether there is real estate, the duration of loan creation and the interest rate of borrowing have a greater impact on the borrower's credit risk, while the amount of borrowing, credit rating, borrowing period and borrowing purpose have no particularly significant impact on the borrower's credit risk. Even if the credit rating of the borrower is high, it is not necessarily possible to obtain loans; on the contrary, for the low-risk loan business, even if the credit rating of the borrower is low, it is possible to obtain loans.

3. Result Analysis and Discussion

Of course, there are also some shortcomings in using decision tree to evaluate credit risk, such as: the financial data of credit risk assessment are generally continuous, which brings the contradiction between efficiency and accuracy to the construction of decision tree; before data analysis, we should pay attention to not only missing value and noise data, but also inconsistent data representation. When inconsistencies are observed, they can be modified by programming software. Because in practice, credit rating and default rate are usually interrelated rather than independent, and vulnerable to the impact of macroeconomic environment so that the market value of all credit instruments changes in the same direction. The decision tree sub-model in Light GBM splits the node by splitting the leaf, so its computational cost is relatively small. It is precisely because of this splitting mode that the depth of the tree and the minimum of each leaf node are needed. The amount of data to avoid over-fitting. After that, according to the user's loan repayment and usage, the borrower's repayment ability and default cost are examined to evaluate different credit ratings.

Lending Club is currently the largest P2P platform in the United States. According to relevant statistics, Lending Club has a 75% market share in the US P2P market. Table 1 below shows the distribution of the loan term for the Lending Club loan term with two loan terms.

Table 1 Proportion of loan term

Loan term	Total loan amount	Proportion of total loans(%)
28	897748	78.64
50	483456	25.33

Data integration is the integration of data from multiple collectors and storage in the same database, which integrates these scattered data for data analysis. After the initial tree is obtained by the decision tree algorithm, the error or noise of the sample data is recorded, which may cause the decision tree to be over- or over-fitting. Therefore, in order to improve its validity, the decision tree needs to be pruned. In terms of estimated data sources, the intensity model abandons the company's capital structure data that is difficult to obtain, and selects observable market data such as the company's credit rating adjustment and bond credit spread for credit risk pricing. By accumulating a large number of borrowers' data, a credit evaluation system for borrowers is established, and pre-loan approval and post-loan risk hints are realized according to the application of mathematical methods. In addition, experiments also show that Light GBM achieves linear acceleration by using multiple machines for specific set-up training.

As a new type of financing mode, P2P network lending has many problems in supervision and development, and the risk of this financing mode has also attracted more and more attention. Although most platforms adopt a diversified approach, each borrowing has many investors, which does spread some of the risk. The method is to introduce independent variables one by one. After

introducing an independent variable, the selected variables are tested one by one. If the original variable is no longer significant due to the introduction of the latter variable, then it is eliminated. Redundancy is an important issue in data integration, so it is necessary to detect redundant data. For quantitative data, correlation coefficients can be used for analysis; for qualitative data, chi-square tests can be used for analysis. When the constructed tree cannot give the correct classification for all objects, add some sample data outside the training set to the training set data. Due to the lack of personal credit information system in China, the P2P platform lacks sufficient third-party credit data to evaluate the credit of borrowers. In addition, some borrowers provide false credit materials in order to obtain loans, which increases the possibility of arrears. The P2P platform should record the borrower's historical credit data and transaction data, as well as the borrower's basic information, and establish a comprehensive borrower database system to improve the P2P online loan credit index system and improve the prediction accuracy of the evaluation model.

4. Conclusions

From the above analysis, it can be found that the default prediction model widely used by most P2P platforms needs to be improved, and the Light GBM algorithm will be one of the alternatives for future improvement methods. Based on the characteristics of P2P network credit platform, this paper establishes a P2P personal credit index system based on Internet behavior information, and builds a model based on Light GBM algorithm to conduct personal credit analysis for P2P borrowers. This paper summarizes the core achievements in the credit risk assessment of P2P network borrowers (including natural persons, legal persons and other organizations), and points out the defects of previous research theories and models. With the continuous development of economic theory and optimization theory, in the future credit evaluation model research, more appropriate evaluation methods can be selected to further improve the prediction ability of P2P network credit evaluation model.

Acknowledgement

In this paper, the research was sponsored by the Business Administration (2018-2021) Subsidy for Key Cultivation Disciplines at School Level of Jilin Agricultural Science and Technology University).

References

[1] Guo Y, Zhou W, Luo C, et al. Instance-based credit risk assessment for investment decisions in P2P lending. European Journal of Operational Research, 2016, 249(2):417-426.

[2] Zhang Y, Jia H, Diao Y, et al. Research on Credit Scoring by Fusing Social Media Information in Online Peer-to-Peer Lending. Procedia Computer Science, 2016, 91:168-174.

[3] Musau F, Guojun Wan. Group formation with neighbor similarity trust in P2P E-commerce. Peer-to-Peer Networking and Applications, 2014, 7(4):295-310.

[4] Melika Mešković, Kos M. Optimization of Chunk Scheduling Algorithm in Hybrid CDN-P2P Live Video Streaming. Iete Journal of Research, 2017(3):1-10.

[5] Kang X, Wu Y. Incentive Mechanism Design for Heterogeneous Peer-to-Peer Networks: A StackelbergGame Approach. IEEE Transactions on Mobile Computing, 2015, 14(5):1018-1030.

[6] Mild A, Waitz M, W ckl, Jürgen. How low can you go? — Overcoming the inability of lenders to set proper interest rates on unsecured peer-to-peer lending markets. Journal of Business Research, 2015, 68(6):1291-1305.