

# *Research on Credit Scoring Mechanism of P2P Lending Platform*

Yonghong Zhang

Tianjin Light Industry Vocational Technical College  
Tianjin, 300350, China

Miao Xu, Cheng Chen

Tianjin University of Science & Technology  
Tianjin, 300000, China

**Abstract**—P2P lending is a financing method suitable for small and micro enterprises. With the development of information technology, P2P lending has become an important aspect of traditional financing research. It is a great challenge for P2P companies to effectively check the borrowers and detect those with bad credit history. Analyze each borrower's basic information based on the information of the lending club website through the multiple linear regression (MLR). Retain the information which has a great influence on the credit score according to the analysis results and detect the cluster-based outlier of the information to find out the abnormal value. Then find the corresponding borrower through abnormal values, namely, the bad credit score borrower. The internal rate of return, IRR, is a further optimized credit scoring mechanism, which is characterized by the ability to measure the benefits of borrowing and lending.

**Keywords**—P2P; credit scoring; outlier detection; internal rate of return IRR

## I. INTRODUCTION

P2P loan is a financing method for small and micro enterprises, which has become an important research aspect of traditional financing. However, it plagues the online P2P company on how to effectively audit the lender. As for online P2P companies, it is difficult to examine the lender in detail through the relevant documents submitted by the lender in the application of the loan, like the traditional bank, so it is difficult to judge the credit of the lender. Therefore, it is a huge challenge for P2P company on how to effectively detect bad credit lenders. At present, in view of this problem, S.Arya combines FICO credit and peer test data to compare the relationship between credit score and lenders' impulsivity, time preference, attitude towards risk and credibility, and concludes that the lender's credit is related to the impulsiveness, time preference and credibility of the lender. Starting with the establishment of a scientific and reasonable personal credit scoring model, Chen Xinli first studied the different credit scoring methods, compared and analyzed the similarities and differences of the methods and their advantages and disadvantages. Finally, the Fuzzy Analytic Hierarchy Process (FAHP) was used as a modeling method. Then he analyzed the individual credit evaluation model and evaluation index of P2P network lending platform at home and abroad, combined the evaluation index of traditional financial institutions with the emerging internet index, and determined the index system of this model. Then he established a personal credit scoring model with Fuzzy Analytic Hierarchy Process (FAHP), and analyzed

the level of the model. Finally, the model was transformed into a simple and easy personal credit scoring form. Based on statistical technology and neuron fuzzy, A.Soner proposed a three level hybrid adaptive neuron fuzzy inference system credit scoring model. By using the credit card data of the Turkey international bank to test the model through ten overlapping validation, the results showed that the score model was superior to linear discriminant analysis, logistic regression analysis and artificial neural network (ANN), in terms of the average correct classification rate and the estimation of the wrong classification cost. The integrated trust model, which was proposed by D.Chen, analyzed using data from 785 online banks from PaiPaiDai, the first and largest online P2P platform in China. The results showed that trust in lenders and intermediaries were an important factor that affects lenders' willingness to borrow money. Emerkter and others analyzed P2P loan features, evaluated credit risks, and found that charging higher interest rates for high-risk lenders was not sufficient to compensate for default losses by using data from Lending Club. Therefore the loan club must find effective ways to attract high FICO scores and high income lenders to maintain their business. Chen Qing made an exploratory study on the impact of the lender's overdue rate by using random data captured on the website, and made a prediction model for the overdue rate.

## II. THEORETICAL ANALYSIS AND RESEARCH HYPOTHESES

### A. *The theoretical framework of credit scoring mechanism*

Credit score refers to the comprehensive evaluation of the willingness and ability of the lender to repay the debt by a qualitative and quantitative method following a certain principle by the network lending platform. For the credit score of the lender, we can fully use the data set of the housing mortgage loan (we use the data of Lending Club, the largest online lending platform in the United States) to detect the credit status of the user after the loan, and evaluate it accordingly. The result can be used to judge the potential breach of contract and whether users need to provide the basis for prepayment. The credit score is calculated according to the corresponding results, and then the user is divided into different grades from high to low by the credit score, which will provide an important basis for investors and service agencies. We will use the multiple linear regression model to analyze the information of the lender, select the attribute of the higher correlation of credit score, and then find out the bad credit lender by the abnormal value detection method.

*B. Credit scores are important for lending*

By exploring the characteristics of P2P lending, assessing the credit risk and measuring loan performance, we found that credit rating and debt income played an important role in loan defaults. There is a high correlation between credit rating and non-performing loans. Loan clubs must find ways to attract high FICO scores and high income borrowers to maintain their business. We need to compare CSVM with other non-linear SVM technologies and find a way to achieve the same level of classification performance while maintaining relatively cheap calculation costs.

*C. The effect of credit score on lending behavior*

We can make full use of housing mortgage loan data sets to observe credit score after loan and make corresponding assessment with the lender’s credit score. The result can be used to be the basis of potential default judgment and advance payment. The results show the credit score is divided into different grades, and the corresponding information commercial paper can be used as an important basis to measure the default potential of credit rating. According to the data analysis of lending Club website, credit score has the greatest impact on the lending behavior, and the higher the score is, the easier it is to obtain the loan. In addition, the more personal financial information the borrower provides, the easier it is to obtain a loan. Besides, the borrower’s age, gender, race and appearance will have an impact on lending behavior. In addition, the higher credit score of the borrower indicates that the lower credit risk and the higher corresponding lending rate.

III. RESEARCH DESIGN OF CREDIT SCORE

*A. Data and samples*

The data comes from lending club, and we randomly select data from lending club, but not all user attributes are valid. We

get 9 attributes of the data set, after preprocessing. The 9 attributes are: credit score, interest, housing ownership, registered monthly income, debt income ratio, loan amount, open credit line, total credit limit, and circular credit limit. The 9 attributes obtained after being preprocessed are not all effective, because some attributes have little effect on the results of credit scores, and there is no substantial significance for the results of the test. We will eliminate that to reduce the complexity of the test and improve its efficiency towards these attributes. In this case, we will use multiple linear regressions to analyze 9 attributes and eliminate those with less influence.

*B. Correlation analysis*

As for the experimental properties, we choose the heavy ones to analyze, and the properties a little associated with the test results can be filtered out. A multivariate linear regression method is used to filter properties.

Multivariate linear regression will have one dependent variable and multiple independent variables, and the function of multiple linear regression is to analyze the correlation between dependent variable and independent variables. We can derive the correlation by association. As the credit score is the final indicator of the lender’s credit, it is used to be the dependent variable and the remaining eight variables are regarded as independent variables. Then we use SPSS software to conduct multiple linear regression operations on the lender data randomly selected from lending Club website and obtain relevant analysis results. TABLE I is the descriptive statistics of attributes and TABLE II is the correlation analysis of attributes.

TABLE I. A STATISTICAL DESCRIPTION OF THE TARGET ATTRIBUTE

variable	Minimum	Maximum	Mean value		Standard Deviation	Variance
	Statistic	Statistic	Statistic	Standard Error	Statistic	Statistic
Credit score	662.00	967.00	700.0667	3.11193	38.11316	1452.613
Interest	532.00	11274.00	1514.5067	83.45492	1022.10984	1044708.520
House-ownership	1000.00	4000.00	2593.3333	75.14143	920.29078	846935.123
Registered monthly income	1167.00	66670.00	8303.3133	632.76560	7749.76419	60058845.04
Debt-to-income ratio	246.00	19884.00	2102.6200	142.59598	1746.43692	3050041.902
Loan Amounts	200.00	5200.00	2021.9000	89.26404	1093.25674	1195210.292
Open credit line	400.00	3300.00	1282.0000	48.66973	596.08004	355311.409
Total credit line	600.00	6600.00	2624.0000	103.55074	1268.23240	1608413.423
Revolving credit line	565.00	212065.00	21238.1200	2143.74015	26255.34752	6.893E8

Note: Data is from lending club

TABLE II. CORRELATION

	Interest	House-ownership	Registered income	monthly	Debt-to-income ratio	Loan Amounts	Open credit line	Total credit line	Revolving credit line
Credit score	0.001	0.493	0.019		0.387	0.003	0.026	0.144	0.166

Note: Data is from lending club

As can be seen from TABLE II, the effect of interest and loan amounts on credit score is very small. So these two properties can be filtered out. The following six attributes will be highlighted: House-ownership, registered monthly income, debt-to-income ratio, opens credit line, total credit line and revolving credit line.

C. Outlier detection method

The outlier detection method clustering can be divided into four types: based on statistics, based on proximity, based on density and based on clustering. Statistical-based methods require information about data distribution and therefore can't be used in our data sets. In addition, the approach based on proximity and density requires a large amount of data, which is inefficient, so we finally choose the cluster-based approach, which is described as below. First of all, we collect all the data into group K and calculate the center. Secondly, we calculate all data objects that are closest to the center. Thirdly, we calculate the relative distance  $\beta$ , and is  $\beta = D(d, center) / M(d, center)$ , among which the distance between the data object and the nearest center is  $D(d, center)$ ,  $M(d, center)$  is the median of the distance between all data objects and their nearest center. Finally, we compare the gap between the relative distance and the specified threshold to get the results. This outlier detection method can be used as a method to find the bad credit score and finally find the abnormal lender in the P2P loan.

Through the scientific analysis of loan information of P2P network loan website, we find out the non-performing loan. The empirical analysis shows that the large probability of the borrowers corresponding to abnormal outlier information has a bad credit score record, which provides a good reference for P2P loan companies.

IV. FURTHER OPTIMIZATION OF CREDIT SCORE—THE METHOD OF IRR

A. Internal rates of return (IRR)

IRR loans with internal rates of return can be uses as a measure of profitability, but also as a measure of earnings from borrowing. The IRR can easily calculate the initial cash outflow(loan amount) and the relevant several cash inflow(payment) investments, which may contain abnormal repayment schedule. In the loan market, IRR is the lender's real interest rate, which is different from the borrower's. IRR also includes the cost of the borrower's loan. There are two advantages of using IRR. First is its permission for more changes in binary information. Second is that IRR considers not only loan payments, but also loan rates.

B. Samples and data

The research used the data of Lending Club. It collects varieties of data including the borrower's information with annual income and loan purposes. Lending Club also provides credit records and FICO scores for information borrowers. From this data, Lending Club assigns a score and sets the interest rate. All the borrower information can be viewed on the Lending Club website and each payment time period. We analyzed 40,901 loans. There are 4,800 failures (11.74%) and 36,101 successes (88.26%).

C. Variables

The variables used in the method are the same as those used in the previous method—the variables used in outlier detection are the same, namely, that is, interest, home ownership, registered monthly income, debt income ratio, loan amount, open credit limit, total credit line and revolving credit line.

D. Concrete analyses

Through the IRR histogram of the loan, we can find out whether the IRR obeys the normal distribution, whether or not we obey the statistical law of asymmetric distribution. The IRR histogram lists the number of loans, the purpose of the loan, and the housing situation of each category. This kind of research focuses on profit score, which can be used as an important index to measure credit rating, so it offers us a new thinking model. The relationship between rank and IRR is not linear, but inverted and U-shaped. It can be considered that P2P lending market is very effective, in which lenders do not have a chance to get positive abnormal returns.

V. SOME SUGGESTIONS ON HOW TO BUILD CREDIT SCORING SYSTEM IN CHINA

First of all, the industry self-discipline organization can act as a bridge between the supervision department and the platform enterprise and it plays the role of communication and coordination. The self-regulatory organizations established in various regions are more familiar with the industry situation and conditions and better master them. In addition, they have a deeper understanding of their own development situation and existing problems. Secondly, the industry self-discipline organization plays the role of supervising and guiding the platform to join the organization. Self-discipline organizations can formulate the codes of conduct and rules of industry self-discipline within a certain range. Members of the association are required to abide by and implement them and they can strengthen internal communication and cooperation to realize information sharing and achieve the effect of supervision and guidance. It is an effective and reasonable method to supplement the supervision of the government, which is an orderly and sustainable development of the industry.

The existing platforms of P2P online lending, such as difficulties in withdrawal, failure and absconding with funds have made investors discouraged and have serious negative effects on the development of the industry. In order to effectively avoid platform for illegal operations and reduce the degree of information asymmetry, the market platform should make full and true disclosure of its management and operation information to form an open and transparent market environment.

As an information intermediary, P2P online lending platform has the obligation to disclose the information of loan target truthfully. On the one hand, the disclosure of loan targets is to avoid platform forgery and virtual borrowing to make illegal financing. On the other hand, disclose enough information for investors through the platform, guarantee investors' right to know, indicate the information of borrowers and the risk of borrowing. They are conducive to investors making appropriate judgment with related risks and make investment choice according to their own risk preferences and risk bearing abilities. In addition, the related information database with government should be shared with the database of online lending platform to finally construct an effective credit scoring system. Users can actually obtain authoritative data and make a reasonable assessment with the credit status of the borrowers.

## VI. CONCLUSIONS

P2P network loan is different from the traditional bank loan. It is very difficult for the network loan platform to distinguish the user's credit situation through the existing customer data, so the analysis of the user's credit situation will be a very important part. Although the abnormal value finding method used in this paper can find out the bad credit score, it is still not

perfect and will be improved continuously in the future research. The advantage of the internal rate of return (IRR) method is that it can measure the income of the borrowing behavior and is more optimized than the outlier finding method. This study is based on a single electronic platform, Lending club, and the subjects have specific identity information. But these conclusions can't be extended to other P2P lending platforms without further proof.

## REFERENCES

- [1] S.Arya, C.Eckel, C.Wichman. Anatomy of the credit score[J]. *Journal of Economic Behavior & Organization*, 2013(5).
- [2] Xinli Chen. Study on the Personal Credit Score Model of P2P Lending Platform[D]. Xi'an University of Electronic Science and Technology, 2015.
- [3] A.Soner. An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data[J]. *European Journal of Operational Research*, 2012(6).
- [4] D.chen, F.Lai, Z.Lin. A trust model for online peer-to-peer lending: a lender's perspective[J]. *Information Technology Management*, 2014(7).
- [5] R. Emerkter, Y. Tu, B. Jirasakuldech, M. Lu. Evaluating credit risk and loan performance in online Peer-to-peer (P2P) lending[D]. *Applied Economics*, 2014.
- [6] Guozheng Zhang, Zhen Yao. Empirical Study on Microcredit Risk Assessment of Farmers Based on Logistic Regression[J]. *Finance and Accounting Monthly*, 2016(27).
- [7] Minghui Jiang, Pei Xu. Research on the Personal Credit Score Based on Optimizing CBR[J]. *Chinese Soft Science*, 2014(12).
- [8] Qi Dong, Aimin Wang. Exploration on the Credit Risk Management Mode in Small Loan Companies[J]. *Accounting Communication*, 2017(14).