SCITECH
RESEARCH ORGANISATION

Volume 10, Issue 2
February 14, 2018

Journal of Research in Business, Economics and Management
www.scitecresearch.com

Research on Credit Risk Assessment of P2P Network Platform:Based on the Logistic Regression Model of Evidence Weight

Zhang Yuan
College of Economics and Management, Nanjing University of Aeronautics & Astronautics,China.

Abstract

As an emerging credit model, P2P network credit has been developing rapidly in recent years. At the same time, it also faces many credit risk problems. This paper focuses on the credit risk of borrowers, and constructs a model of WOE and logistic regression to evaluate the risk assessment of China's P2P network platform, Hong ling Venture. The research results show that the main factors that affect the loan success rate of P2P lending platform include loan amount, annual interest rate, bidding transaction amount and proportion of repayment on time and so on. By constructing the model of combination of the logistic regression with weight of evidence, this paper provides an appropriate method to manipulate the borrowing information of loan borrowers and evaluates the borrowing behavior of borrowers simultaneously, so that P2P credit platform can reduce the credit risk caused by borrower.

Keywords: P2P Network Credit; Credit Risk Assessment; Logistic Regression; Weight of Evidence.

1. Introduction

P2P network loan is an important mode of Internet Finance connecting peer to peer. It is a new lending way that individuals can lend their spare funds to the fund shortage through the Internet platform. At present, China's P2P network platform provides effectual way and convenient service for investors and financiers to meet their needs. But due to nonstandard credit platform, imperfect individual credit system, and lack of appropriate laws and regulations to manage and constraint the market, the whole industry's risk is in a sharp rise. Among them, the biggest risk is the borrower's credit risk. Many China's credit platforms are not well prepared for risk control when they are seizing the market and achieving rapid expansion. Many platforms lack perfect credit risk assessment systems to make accurate and effective credit evaluation of borrowers, which results in a huge risk of default and operational crisis. Therefore, it is particularly important for P2P network platform to build a model for quantitative evaluation of borrowers' credit information risk on P2P network credit platform, and provide reliable risk assessment methods for participants of network credit platform.

2. Literature Review

The empirical analysis of the data of the Prosper network platform showed that, the loan default rate is not necessarily related to the risk premium (Sanjeev Kumar 2007)[1]. The more financial indicators that the borrower provides, the easier it is to get a loan. (Freedman and Jin 2008)[2]. The net loan platform that acts as a financial intermediary can improve borrowers' credit. And the intermediary agency of net lending platform can improve the borrowers' credit results (Berger and Fabian Gleisner 2009)[3] . The more the certified borrowings are described, the stronger the borrower's willingness and ability to repay, it is more easily for borrowers to get the loan (Herzenstein et al. 2011) [4]. Both trust in borrowers and trust in intermediaries are significant factors influencing lenders' lending intention, however, trust in borrowers is more critical. In order to develop lenders' trust, borrowers should provide high quality information for their loan request (D Chen et al. 2014)[5]. Hazard rate or the likelihood of the loan default increased with the credit risk of the borrowers ,and higher interest rate charged on the high risk borrowers

are not enough to compensate for higher probability of the loan default (R Emekter et al. 2015)[6]. Predicting whether a borrower will default on a loan by propose a default prediction method for P2P lending behaviors (C Jiang et al. 2017)[7]. The financial and credit status of P2P platforms are key elements in building the trust of investors and impacting their decisions, while the disclosure of information by the borrowers does not significantly affect the number of platform investors (Y Yan et al. 2017)[8].

3. Research Method: The Logistic Regression Model of Evidence Weight

This paper adopts the method of combining qualitative and quantitative. Starting with the borrower credit information firstly, this paper screens appropriate index variables through the value of information and the diagnosis of collinearity, then the raw data will be replaced by WOE, and used in Logistic regression model. Finally, the paper judges the size of credit risk according to the regression results and provide effective credit risk assessment method for credit P2P network platform.

WOE and Logistic regression will be introduced respectively and the concrete algorithm of the regression model of evidence weight is given in the following content.

3.1 Weight of Evidence (WOE)

For the continuous random variable X , the probability density function is $p(x)$, and the entropy is defined as:

$$S(X) = -\int p(x)\ln p(x)dx = -E\left[\ln p(x)\right]$$

There are two continuous random variables X and Y , the relative entropy, namely KL divergence (Kullback – Leibler Divergence) is used to characterize the distance between two continuous random variables. The definition is as follows:

$$S(Y \mid X) = E_Y\left[\ln\frac{p_Y}{p_X}\right] = \int p_Y(x)\ln\frac{p_Y(x)}{p_X(x)}dx$$

Among them, p_X and p_Y stands for the probability density of X and Y respectively, E_Y is the expectation of the random variable Y .

When uses the information entropy in the credit scoring system, the paper will give credit score of the debtor F . W and $\overline{W}$ represent the event of default and the normal repayment event respectively. Information Value (IV), as a measurement tool based on relative entropy, distinguishes the difference between the default debtor's score distribution and the normal repayment debtor's score distribution. Assuming density function of the default debtor's score distribution is $P_W(F)$, and the density function of the normal repayment debtor's score distritution is $P_{\overline{W}}(F)$. IV is defined as follows.

$$IV = E_W\left[\ln\frac{P_W(F)}{P_{\overline{W}}(F)}\right] + E_{\overline{W}}\left[\ln\frac{P_{\overline{W}}(F)}{P_W(F)}\right] = \int\left(P_{\overline{W}}(F) - P_W(F)\right)\ln\frac{P_{\overline{W}}(F)}{P_W(F)}df \qquad (3)$$

By calculating IV, the WOE is defined by IV's expression, and the expression is as follows:

$$WOE = \ln\frac{P_{\overline{W}}(F)}{P_W(F)} \qquad (4)$$

So the increase of WOE means the reduction of default risk. Because information value is a measure of the difference between the debtor and the normal debtor, the value of information should be as large as possible for the credit rating system with higher screening ability.

3.2 Logistic Regression

Collect sample data $H = \{x_i, y_i\}$, x_i represents the customer's index variable and y_i represents two classified variable. $y_i = 1$ means the ith debtor is a good customer to repay normally. $y_i = 0$ indicates that the ith ebtor is a bad customer who breaks a contract. This paper evaluate the P2P network platform by follow formula.

$$P(y=1|X)=\frac{e^{\alpha_0+\alpha^T X}}{1+e^{\alpha_0+\alpha^T X}}$$

In the formula, *X* is a vector that the dimension is equal to the number of the customer's index variables. α is the estimated parameter, and the dimension is the same as X. α is also the weight of each index variable.

The formula above can be drawn as follows:

$$\ln\frac{p(y=1|x)}{1-p(y=1|x)}=\alpha_0+\alpha^T X$$

The left side of the equation above is the logarithm of the ratio of the debtor's probability of repayment to the debtor's probability of default under an index variable. Among them, The ratio of the debtor's normal repayment probability and the debtor's probability of break a contract, which is the ratio of good customers to bad customers, will be a very critical index in the model.

3.2 Algorithm

(1) After calculating the WOE value and the IV value, customers will be grouped according to IV's value. Given $X=(x_1,x_2,\cdots,x_n)$ as an index variable of n debtors; Y_i represents the actual situation of the ith debtor. The debtor of a normal repayment $Y_i=1$ is called a good customer, and a default debtor $Y_i=0$ is called a bad customer. In order to find out the reasonable segmentation of index variables, we set the index variable *X* in ascending order, written as $X^{'}=(x_{(1)},x_{(2)},\cdots,x_{(n)})$, divide *X* into *k* intervals according to the equal probability method, and find out a $k-1$ quantiles ($2<k<10$). Secondly, G_i and B_i represent the number of good customers and bad customers inith interval respectively. *G* and *B* represent the total number of good customers and bad customers respectively, and calculate the values of WOE_i of each interval and the whole value of IV. Finally, By comparing the IV values of all subsections, we select segmentation of the maximum IV value as the reasonable segmentation of the index. And among them , $WOE_i=\ln\frac{G_i/G}{B_i/B}, i=1,2,\cdots,k$, $IV=\sum_{i=1}^{k}\left[\frac{G_i}{G}-\frac{B_i}{B}\right]WOE_i$.

(2) Eliminate the index variables through the IV value and the collinear diagnosis.

(3) We take the processed index variables of WOE into new design matrix, instead of the original value, to be taken into logistic regression model, and then use the corresponding results model (parametric test, goodness of fit test and hypothesis testing) to get final model. After all, we can calculate ROC curve (Receiver Operating Characteristic curve)with the success probability of loan repayment, training samples and the test samples.

In order to prove the prediction accuracy of the WOE and logistic regression model, we conduct an empirical analysis through the data from P2P platform Hongling Venture website.

4. Empirical Analysis

4.1 Data Description

We use the web information collector to grab the loan information from 2012 to 2016, which is announced by Hongling venture website. Among them, 1152 loan data are used as training samples and 348 loan data are used as test samples, and training samples contain 960 normal customers and 192 default customers.

4.2 Empirical Analysis

4.2.1 Calculation of IV Value and WOE Value and Grouping the Indexes

This paper has collected 19 variables, including loan amount, annual interest rates, loan term, loan types, repayment way, number of bidding, membership level, borrowing credit, lending credit, repayment state ,auditing and certification of all data. As there are so many indicators are involved in the paper, we take the amount of money as an example. 1152 data were collected, including 960 good customers and 192 bad customers. Good represents the number of customers. Bad indicates the number of bad customers. $P_1=G_i/G$ indicates the proportion of good customers in the ith interval to the total good customers. $P_0=B_i/B$ indicates the proportion of bad customers in the ith interval to the total bad customers. The specific grouping is as follows:

Table 1: Grouping According to the Amount of Borrowed Money						
Value Range	**Good**	**Bad**	**P_1**	**P_0**	**WOE**	**IV**
[0,10000]	155	148	0.1615	0.7708	1.5632	0.9526
（10000,60000]	233	40	0.2427	0.2083	0.1527	0.0052
（60000,300000]	304	2	0.3167	0.0104	3.4144	1.0457
（300000，800000]	268	2	0.2792	0.0104	3.2884	0.8838
Total	960	192	1	1	8.4188	2.8873

By comparing IV of different grouping method, we find that IV get the maximum value when the loan amount is split into 4 groups. So the reasonable grouping of the loan amount is 4.

4.2.2 Eliminate the Index Variables Through the IV Value and the Collinear Diagnosis

IV means information value. With reference to the experience of FICO and other institutions, the threshold is set as 0.03. When the IV value is more than 0.03 and less than 0.1 8, the discrimination of index variable is limited, and the index can be excluded. When the IV value is more than 0.18, it is considered that the index variable has higher degree of distinction, and this index variable can be used. After data processing and maneuverability calculation, 9 index variables are selected and the total IV values are shown in Table 2 as follows.

Table 2: Primary Selection Index Variables and Total IV Value	
Index Variables	**Total IV Value**
Credit of Borrowing	3.3995
Borrowing Amount	2.8873
Annual Interest Rate	1.7375
Modes of Repayment	3.0253
Number of Late Payment	1.1233
Number of Bids	0.9861
Number of Unpaid	4.7391
Numbers of Repayments on time	0.3849
Offer of Income Certification	0.3732

A common linear diagnosis is carried out by the index variable which is eliminated by the IV value. After multiple linear regression analysis, the results show that, the tolerance of all factors are greater than 0.1, and the variance inflation factor (VIF) is far less than 10. But there are multiple collinearity between these indexes. Through stepwise regression analysis, 6 variables are approved by the test, including annual interest rate, borrowing Amount, number of unpaid, number of repayment on time, number of bids and offer of income certification.

4.2.3 Comprehensive Test of Model Coefficient

Grouping six indicators according to the method of getting the highest IV value separately, and take the WOE value into the logistic model instead of the original value. And then we do regression fitting by SPSS software, the results are showed in table 4.

From table 4 we can see that, the overall parameter is not 0 significantly, P values are lower than the 0.05 level of significance, so the parameter estimation is good. The modified R-square is 0.875, the data is high, so that it can accept the model fitting, 6 variables are all passed the Wald test.

Table 3: Comprehensive Test of Model Coefficient				
		Chi-S	***df***	**Sig.**
Step 1	Step	640.393	1	.000
	Module	640.393	1	.000
	Model	640.393	1	.000
Step 2	Step	62.084	1	.000
	Module	702.477	2	.000
	Model	702.477	2	.000
Step 3	Step	38.235	1	.000
	Module	740.712	3	.000
	Model	740.712	3	.000
Step 4	Step	28.213	1	.000
	Module	768.925	4	.000
	Model	768.925	4	.000
Step 5	Step	8.622	1	.003
	Module	777.547	5	.000
	Model	777.547	5	.000
Step 6	Step	5.379	1	.020
	Module	782.926	6	.000
	Model	782.926	6	.000

The regression results are shown in Table 4. From table 4, we can see that when the significance level is 0.05, the significance of all variables is less than 0.05, which shows that 6 variables have a significant impact on the success of borrowing.

Table 4: Model Regression Results						
	B	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**
Borrowing Amount	.952	.213	19.936	1	.000	2.591
Annual Interest Rate	.827	.151	30.187	1	.000	2.286
Number of Bids	.545	.237	5.277	1	.022	1.724

Number of Late Payment	-.756	.086	77.201	1	.000	2.130
Offer of Income Certification	.744	.263	8.010	1	.005	.475
Numbers of Repayments on time	3.327	.496	45.061	1	.000	27.855
Constant	1.262	.196	41.517	1	.000	3.532

The sample model is as follows: $Ln\left(\frac{P_1}{P_2}\right) = Ln\left(\frac{P_1}{1-P_1}\right) = Z$ = 1.262+0.952 * (WOE value of the loan amount) +0.827 * (WOE value of annual interest rate) +0.545 * (WOE value of the bid number) -0.756 * (WOE value of late payment number) +0.744 * (WOE value of the income certificate), +3.327 * (WOE value of repayment on time number).

Then the probability of judging a customer as a good customer is showed as follows:

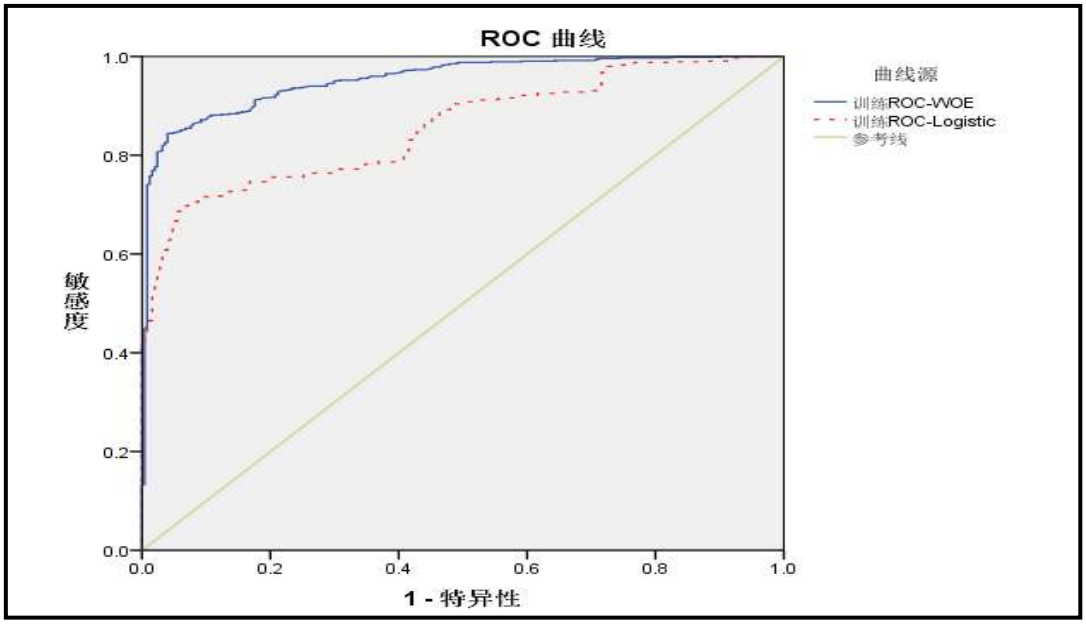
$$P_1 = \frac{EXP(Z)}{1+EXP(Z)} \quad (7)$$

4.2.4 Test the training samples

To get the coefficients of variables through WOE and logistic regression model, we can calculate the probability of whether a customer is good customer, but we don't know how much the probability is to be a good customer, so it should be segmented, and then make an objective evaluation on fitting model. The method of determining the optimal segmentation point is that , determined by the maximum deviation degree of ROC curve. Under this probability, the difference between the proportion of the good customer and the bad customer reaches the maximum, which is the optimal cutting point.

By drawing ROC curve, the area under the ROC curve of the training sample is 0.948, the fitting effect is very good, its significance is obviously. Through the test, it can distinguish the good and bad customers significantly.

Fig.1 ROC Curve of Training Samples



We calculate the calculating the probability by 0.872* (1-0.052) =0.82557, which is usually used to calculate the optimal segmentation point, the result is shown in table 5. Therefore, when the return probability is greater than 0.802433, the customer will be predicted as a good customer, and when the return probability is less than 0.802433, it is predicted to be a bad customer.

Table 5: The Coordinates of the Training ROC Curve			
Hypothesis is Positive	**Sensitivity**	**1- Specificity**	**Sensitivity * Specificity**
0.000000	1.000	1.000	0
.005214	1.000	.989	0.012
……	……	……	……
.802433	.872	.052	0.82665
……	……	……	……
.999878	.001	0.000	0.0011121
1.000000	0.000	0.000	0

4.2.5 Test of the Model

From table 6 we can see that, the accuracy of the overall prediction model of training samples reached at 86.45%, of which 192 are bad customers, only 8 bad customers are misjudged as good customer, the accuracy rate is 95.83%. 162 customers were misclassified as bad customers for 960 good customers, and the accuracy is 84.58%.

Table 6: Prediction Accuracy of Training Samples			
Customer types	**Prediction of Good Customers**	**Prediction of Bad Customers**	**Accuracy**
Good Customers	812	162	84.58%
Bad Customers	8	184	95.83%
Total Accuracy			86.45%

4.2.6 Test of Model's Prediction Accuracy

To further verify the validity of the model, 348 test samples are taken into the model. After the data processing process above, we get the test samples' area under the ROC curve, which is 0.972, the regression results is good, and the significance also passes the correlation test.

The overall prediction accuracy of test sample is 94.25%, which is shown in Table 7 as follows.

Table 7: Prediction Accuracy of Test Sample			
Customer Types	**Prediction of Good Customers**	**Prediction of Bad Customers**	**Accuracy**
Good Customers	230	16	93.49%
Bad Customers	4	98	96.07%
Total Accuracy			94.25%

The result of the test sample is consistent with the prediction accuracy of training samples. The prediction accuracy is high and the model is stable. The WOE and logistic regression model is valuable for popularized application and prospect.

5. Conclusions

After the model test of the previous section, it shows that the model has a good prediction effect, and some conclusions are obtained as follows:

i) In all the variables with greater impact, the number of punctual repayment pens has the greatest impact on the success rate of the loan. Among the selected 6 index variables, the amount of borrowing, the annual interest rate, the number of tender pens, and the number of punctual repayment pens are proportional to the success rate of the loan, that is, the greater the above index variables, the higher the success rate of the loan; only the number of overdue repayment pens is negatively related. The platform should be screened when examining the borrower's information.

ii) Before testing the model, a certain number of training samples should be tested, and then the test samples should be substituted for the test of the band when the correct rate is guaranteed. The number of credit information for good and bad customers is limited because of the official website of the Hongling Venture, but the test results of the sample need more data support. On the premise of full model sample, credit platform should continuously test and repair the model so as to achieve better results.
(3) Some of the variables in the model, which are eliminated because of their objective impossibility, may exclude some factors that have a greater impact on the success rate of the loan, thus, the result of model construction deviates from the reality. In order to further improve the model, credit platform should strengthen the audit and improvement of credit information and update the model dynamically so that it has practical value.

iii) In this paper, the risk assessment of credit platform based on the borrower's credit information is combined with the combination of the weight of evidence and the logistic regression. That is, the WOE is introduced into the Logistic model and the WOE is used to code the independent variables. It not only improves the prediction effect of the model, but also improves the comprehensibility of the model. In spite of this, the prediction model needs further improvement and optimization from the point of view of the rigor of data processing and the adequacy of sample selection.

References

[1] KUMAR S. Bank of One: Empirical Analysis of Peer-to-Peer Financial Marketplaces; proceedings of the Reaching New Heights Americas Conference on Information Systems, Amcis 2007, Keystone, Colorado, Usa, August, F, 2007 [C].

[2] JIN G Z. Do Social Networks Solve Information Problems for Peer-to-Peer Lending? Evidence from Prosper.com [J]. Seth Freedman, 2008, 8-43.

[3] BERGER S C, GLEISNER F. Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending [J]. Social Science Electronic Publishing, 2009, 2(1): 39-65.

[4] HERZENSTEIN M, SONENSHEIN S, DHOLAKIA U M. Tell Me a Good Story and I May Lend You Money: The Role of Narratives in Peer-to-Peer Lending Decisions [J]. Journal of Marketing Research, 2011, 48(SPL): S138.

[5] CHEN D, LAI F, LIN Z. A trust model for online peer-to-peer lending: a lender's perspective [J]. Information Technology & Management, 2014, 15(4): 239-54.

[6] EMEKTER R, TU Y, JIRASAKULDECH B, et al. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending [J]. Applied Economics, 2015, 47(1): 54-70.

[7] JIANG C, WANG Z, WANG R, et al. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending [J]. Annals of Operations Research, 2017, (2–3): 1-19.

[8] YAN Y, LV Z, HU B. Building investor trust in the P2P lending platform with a focus on Chinese P2P lending platforms [J]. Electronic Commerce Research, 2017, (2): 1-22.