Financial Engineering and Risk Management (2020) 3: 14-22
Clausius Scientific Press, Canada
DOI: 10.23977/ferm.2020.030102
ISSN 2523-2576

# *An empirical study on P2P loan default prediction model*

**Yang Yang**

*Department of Insurance and Economics, University of International Business and Economics, Beijing 100029, China*
*yysneaker@163.com*

***Keywords:*** Logit Model; Consumption Loan; Audit Model; Confusion Matrix

***Abstract:*** With the increasing role of personal consumption loan business in economic life, the Internet financial industry has developed rapidly. Due to the lack of supervision and related laws, many irregularities and illegal acts are mixed among them. A large number of enterprises with the purpose of gathering funds are mixed among them. Under the banner of Internet finance, illegal actions to absorb public deposits are carried out. At the same time, due to the lack of supervision and related laws, many irregularities and illegal acts are mixed among them. Industry competition, guaranteed income and other irregularities are prevalent in the industry. Companies in the industry generally use funds for advertising and blindly expand their scale. The most important compliance management and internal wind control mechanism have been in a backward or even missing state for a long time, resulting in a large-scale decline and Thunderstorm in the industry, which has a huge negative impact on industry development and economic development. Under this background, this paper focuses on the internal risk control and studies the application of Logit Model in the auditing model of personal consumption loan. Firstly, starting from the logit model itself, this model is reasoned and introduced. Then, the auditing model of personal consumption loan is established by using logit model. Through the establishment and testing of the model, the model is further optimized and the model is beneficial to personal consumption loan is obtained. Suggestions and suggestions on the design and use of the model.

## 1. Introduction

P2P is the abbreviation of peer-to-peer, which means individual to individual. P2P online loan is a new type of personal to individual loan based on Internet platform. Every loan of P2P online loan involves the interests of three parties. First, the lender transfers the currency to the borrower in the loan activity, collects the principal and interest according to the contract, and then P2P online loan platform, platform As the information media of both parties, it is responsible for reviewing the borrower's information, providing it to the lender, brokering the transactions between both parties, collecting service fees, releasing loan information in P2P lending, obtaining funds, and being responsible for paying interest and repaying principal on time.

Commercial banks have complex procedures for loan review and high thresholds. The traditional group of bank loans has always been state-owned and large and medium-sized enterprises. Traditional small and micro enterprises are difficult to obtain financing through commercial banks, thus forming a small loan market. In this context, private credit has always assumed this role, in the

background of Internet integration with other industries Next, China's P2P network loan industry came into being. In 2007, China's first P2P financial company officially went online, after which the number of P2P online loan platforms showed a rapid growth. As of March 1, 2017, there are 5527 P2P online loan platforms in China .

P2P financial companies are growing up in the unregulated market environment. During the period of rapid development, the industry has covered up many problems. Due to the lack of corresponding supervision, P2P financial companies are mixed with a variety of enterprises for the purpose of absorbing loans and gathering funds. Under the banner of Internet finance, they are laying advertisements online and offline, claiming "capital guarantee" and "platform guarantee" "," national endorsement "," minimum yield ", etc., attract investors to take the bait, and then implement the means of capital mismatch, long loan and short loan, etc., gather funds, run with money at the right time, form a large-scale group lawsuit event, and damage the reputation of the industry. At the same time, in the process of development, the competition in the industry makes P2P financial companies choose to lay advertisements on a large scale, which makes the cost of capital rise, and can not find the right investors, laying the foreshadowing for the breaking of the capital chain.

Since July 2018, P2P financial companies have started a storm wave. When the capital chain is broken, companies have chosen to run, liquidate and close down, and the Internet financial industry is in danger. In this context, P2P online loan companies should strive to improve their own internal, clear their intermediary positioning, protect the interests of investors, strictly review the information of lenders, and improve their service level and quality.

The development of P2P industry is a history of risk exposure and regulatory improvement. The predecessors mainly focused on the risk early warning and risk management of P2P online lending platform. Starting from the business model of P2P online lending industry, Bai Chuanping divided the domestic P2P online lending industry into online pure intermediary [1], online compound intermediary, and online compound intermediary. There are four types of online and offline composite intermediary type and pure public welfare type. Then, compared with the regulatory differences at home and abroad, in view of the domestic reality, the paper puts forward multi-level regulatory opinions. Chen Jie analyzed the illegal fund-raising and Ponzi fraud and other criminal behaviors of P2P online loan industry from the perspective of law[2], and then made suggestions on the development of P2P online loan industry from the perspective of legislation and law enforcement. Kang Caijuan analyzed the operation of Y company from the micro perspective [7], and put forward countermeasures from the perspective of external government supervision and internal risk control of enterprises. Xu Lida attributed the frequent occurrence of P2P network loan malignant events to the introduction of illegal guarantee [12], and put forward development suggestions from the aspect of P2P network loan platform to guarantee.

On the other hand, the logit model is used to examine loan information, which has been well written by predecessors Et al. Qian zhengming based on the data of personal housing mortgage loan in a certain area of China in 2008[8], using nonparametric random forest method and logit model to study the default risk of personal housing mortgage loan, elaborated the empirical research on the review of housing mortgage loan based on logit model under the background of the increase of non-performing loan rate of commercial banks, and proposed that the review of housing mortgage loan should Focus on the borrower characteristics, loan characteristics, external economic characteristics, and real estate characteristics of these four dimensions, through the comprehensive consideration and weighting of these four dimensions, to review the housing mortgage loans. Hu Yi Based on the panel data of customer credit of commercial banks from January 2009 to April 2012[5], this paper applies the logit model to the research of bank customer default risk, which is mainly aimed at the default risk of large loan customers of commercial banks, mainly aimed at the

research of enterprise loan risk, and puts forward three comprehensive indicators, namely, paying attention to external factors of customers, customer operation level and customer transaction behavior After modeling and analyzing 24 factors of the three indicators, this paper puts forward suggestions on strengthening risk management and internal risk control. Qu Qiushi and Li Li The logit model is applied to the evaluation of personal credit risk of commercial bank customers[9]. The index data of personal credit loan in the database of commercial bank is selected to study the default risk of personal credit loan. The four comprehensive indexes of assets, liabilities, education level and marital status are selected to analyze, and suggestions for personal credit management of commercial bank are put forward. Yu Jia based on the business data of P2P online loan company[13], such as transaction amount, accumulated loan balance, interest rate, number of investors, number of borrowers, average loan cycle, full bid speed, registered capital and registered place, studied the default behavior of online loan company, and obtained the suggestions for external supervision and internal risk control of P2P financial company.

To sum up, for the supervision and development of P2P network loan companies, most of the predecessors focused on the suggestions of the government and law enforcement departments, but for the micro level, how to do internal risk control and audit for P2P network loan companies, there is no relevant literature. At the same time, although there are many works on the application of Logit model, it is not applied to the customer loan audit of P2P network loan companies Therefore, this paper uses logit model, for small personal consumption loan business, selects the consumption loan data of Internet companies, reviews the information of the lender, optimizes the model through the establishment and test of the model, improves the accuracy of the model, and then obtains opinions and suggestions conducive to the development of personal consumption loan business.

## 2. Data source and variable definition

### 2.1 Data source

The data used in this paper comes from the personal consumption loan data published on the website by lending club, a P2P company in the United States. There are 20000 personal consumption loan data in total. The data of each personal consumption loan includes 9 categories: ID, loan date, loan amount, loan interest rate, credit rating, housing status, working years, credit score (FICO) and annual income Among them, ID, loan amount, loan interest rate, credit score (FICO) and annual income are continuous random variables, and credit rating, housing status, working years and loan occurrence date are virtual variables.

### 2.2 Variable definition

(1) Dependent variable

In this paper, whether the loan is approved or not will be regarded as the binary dependent variable y, in which the loan default represents the binary dependent variable y = 1, and the loan completion represents the binary dependent variable y = 0.and 1 are used as dependent variables for regression. We add loan amount, loan interest rate, credit rating, credit score (FICO) and annual income into the model. At the same time, we add two newly generated variables, loan income ratio and asset liability ratio, into the model, so that the model has seven independent variables.

(2) Loan amount

The loan amount is the total loan amount in the loan data of this article. Since the selected data is the small consumption loan data, the fluctuation range of the loan amount is narrow. The variable is a continuous variable with a value of $100 to $36000. In order to prevent the influence of too large a value on the model, divide the value by the coefficient of 100, so the value range is 1 to 360.

(3) Loan interest rate

The loan interest rate is the interest rate of the loan data in this article, and the value range is 5.42% to 25.42%. The higher the loan interest rate is, the higher the risk of the borrower is, so the loan can only be obtained through the higher capital cost. Multiply the loan interest rate by 100 at the same time, and the value of the loan interest rate becomes 5.42 to 25.42. The loan interest rate is a continuous variable with a numerical value.

(4) Credit rating

The credit rating is a comprehensive rating of the borrower's credit, which is divided into seven levels from a to g. among them, level a represents the best credit and level g represents the worst credit. The level of credit reflects the size of the borrower's default risk. The higher the rating, the lower the default risk. Seven levels a to G are defined as numbers 7 to 1, so that new variables can be added to the model, and the credit rating becomes a continuous variable.

(5) House condition

Housing status is an indicator of personal asset status, which can be divided into three types: rent (lease), own (with or without housing loan) and mortgage (with housing loan). The two types of ownership (with housing loan) and mortgage (with housing loan) are defined as 0, and rent (lease) is defined as 1. The house condition is a logical variable with a value of 0 or 1.

(6) Working years

The working years are the years that the borrower has participated in the work when the loan occurs, which are divided into 12 categories: 1-9 years, less than 1 year, more than 10 years and data loss. The less than 1 year and data loss are defined as 0 year, and the more than 10 years are defined as 10 years. The value range of working years is 0 to 10, and the value of working years is continuous variable and numerical type.

(7) Credit score

Credit score is the credit score matching with personal credit record. Credit score will be increased for each completed credit record, and credit score will be deducted for default or expectation. The value of credit score is 660 to 829. Divide credit score by coefficient 10 at the same time. The value is 66 to 82.9. Credit score is a continuous variable and the value is numerical.

Table 1 Definition of variables

| Variable Name | Meaning | Take value | Definition |
|---|---|---|---|
| Dependent variable y | Loan Default or Not | 0 or 1 | Loan completion is 0, default is 1 |
| Loan Amount | Total Loans | $100-$36,000 | Take the natural logarithm of the loan amount |
| Loan interest rate | Interest rate for a single loan | 5.42%-25.42% | Multiply the loan interest rate by a factor of 100 |
| Credit rating | Describe the level of credit status | A-G seven levels | Divide the seven levels of A-G into the corresponding numbers 7 to 1 |
| Housing condition | Personal housing status | Own, Loan and Lease Categories | Own and loan are defined as 0, and lease is defined as 1 |
| Working years | Working years up to the date of loan occurrence | Twelve categories of 0-10 years and over 10 years | 0-10 years are the numbers 0 to 10 respectively, and more than 10 years are defined as the number 10. |
| Credit score | Score describing the credit status of an individual | Interval 660-829 | 660-829 are defined as numbers 660 to 829, respectively |
| Annual income | Total income for the year | $6,000-$1,786,000 | Take the natural logarithm of |

| | | | |
|---|---|---|---|
| | | | annual income |

(8) Annual income

The annual income is the total income of the borrower in the year when the loan occurs, and it is an important indicator to measure the financial status of the borrower. The fluctuation range of the annual income is $6000 to $1786000. The higher the annual income is, the less the risk of default of the borrower is. Due to the large value of the annual income variable, the annual income variable is taken as the natural logarithm value, and a new variable is generated and added to the model. The annual income variable is a continuous variable, and the value is taken as the numerical type.

## 3. Research ideas, empirical analysis and prediction

### 3.1 Model

The problem that this paper faces is whether the loan is approved or not, that is, a binary selection problem. The models that can be selected for the binary selection model include logit model and probit model. These two models are all discrete selection models, but the assumption of the distribution of error terms after regression is different. The logit assumption of error terms is subject to logical distribution, and probit assumption of error terms is subject to logical distribution Normal distribution, so usually the results of the two are not different. In practice, logit model is more widely used, more acceptable, and easier to interpret the regression results. The model established in this paper is logit model.

$$L_{odds} = Log\left(\frac{P_1}{1-p_1}\right) = \beta_0 + \beta_1 Loanamount_i + \beta_2 Loaninterestrate_i + \beta_3 Creditrating_i + \beta_4 Housingcondition_i + \beta_5 Workingyears_i + \beta_6 Creditrescore_i + \beta_7 Annualincome_i + \varepsilon_i$$

### 3.2 Research ideas

At the same time, we use the random functions of Excel to classify the data sets. One is the training sample, the other is the test sample. The proportion is 7:3. After the above operations are completed, the logit regression is performed on the equation. On the basis of the logit regression results, the confusion matrix is generated, and the probability of class I error (rejecting the original hypothesis) and class II error (accepting the wrong hypothesis) of the model is calculated. On this basis, the original model is improved to improve the accuracy of the model. After that, the test samples are entered into the model, and the results are checked.

On this basis, for the model to make type I errors (reject the original assumption) and type II errors (accept the wrong assumption), both of which are the deviation of the model to the prediction results, which should be avoided in the training. The first type of error refers to the rejection of customers who will not default on the loan, resulting in the loss of transaction volume and the loss of customer resources. The second type of error It refers to the acceptance of customers who will default on the loan, and the loss caused is the loss of investors, that is, the occurrence of loan expectation or event of default. In practice, we take the attitude of being responsible for investors. When we have to choose between two kinds of errors, we choose the first kind of error, namely, forgery and forgery, and the loss of forgery is greater.

### 3.3 Regression analysis

Input the data and model into SPSS software, first is the test set. After the test set forms the

model, the remaining test samples are used to test the model. The regression results of seven variables were obtained.

Table 2 Regression results of the model

| Variable | coefficient | Standard error | Wals | P |
|---|---|---|---|---|
| Loan Amount | 0.0425 | 0.0659 | 0.4155 | 0.51920 |
| Loan interest rate | 0.3078 | 0.0354 | 75.5306 | 0.00000 |
| Credit rating | -0.0025 | 0.0009 | 8.474 | 0.00360 |
| Housing condition | -0.0078 | 0.0017 | 21.8761 | 0.00000 |
| Working years | -0.6461 | 0.0815 | 62.8569 | 0.00000 |
| Credit score | 0.0235 | 0.0107 | 4.8449 | 0.02770 |
| Annual income | -0.0679 | 0.0792 | 0.7338 | 0.39170 |
| constant | -5.9077 | 1.6947 | 12.152 | 0.50000 |

The coefficient of loan amount is positive, which means that the higher the loan amount is, the higher the probability of default is. Because the higher the loan amount is, the higher the borrower's liabilities are, so the greater the investment risk. But the loan amount is not significant under the 95% confidence interval, mainly because the data selected in this paper are small consumer loans, and the difference between each loan data is small, so the loan amount is not significant at the 95% significance level.

The coefficient of loan interest rate is positive, which means that the higher the loan interest rate is, the higher the probability of default is. At the same time, the loan interest rate is significant under the 95% significant confidence interval, which shows that the loan interest rate is an important index to reflect the borrower's default probability, which can well reflect the borrower's default risk.

The coefficient of credit rating is negative, which indicates that there is a negative correlation between credit rating and loan default probability. The higher the credit rating is, the lower the default risk is. The credit rating also passed the 95% significance test. Credit rating is a certain indicator given by lending club after comprehensive evaluation of the borrower's default risk. From the regression results, the indicator has high credibility and can accurately reflect the borrower's default risk.

The coefficient of credit score is negative, which shows that credit score has a negative correlation with the probability of loan default. As an indicator of credit status associated with individuals, credit score is significant at 95% confidence interval. It shows that credit score can reflect personal credit status well, and it is a correct and necessary choice to incorporate credit score into personal loan default prediction model.

The coefficient of annual income is negative, which means that the higher the annual income is, the lower the borrower's default probability is, and the negative correlation between them is also consistent with our common sense. Annual income is significant at 95% confidence level, indicating that personal income is an important financial indicator, which can well reflect the financial situation and default risk.

The coefficient of working years is negative, indicating that with the increase of working years, the probability of default increases, and the two are positive correlation. At the same time, the working life is significant at 95% confidence interval. Generally speaking, the working years are positively correlated with personal income, but from the regression results, there may be other implicit variables such as personal expenditure with the increase of personal income that are not

taken into account, so the working years are positively correlated with the probability of default.

The coefficient of housing condition is negative. In the definition, the self owned and loan are defined as 0, and the lease is defined as 1, which means that the borrower's default probability of the leased housing is lower. At the same time, the housing condition is not significant under the 95% confidence interval, so this variable should not be added to the model.

### 3.4 Prediction results

Table 3 Prediction results

| | observe condition | | | | |
|---|---|---|---|---|---|
| Prediction situation | | Defaulter number | Number of non defaulters | Total | Accuracy rate |
| | Defaulter number | 5619 | 232 | 5851 | 99.65% |
| | Number of non defaulters | 20 | 131 | 151 | 36.09% |
| | Total | 5639 | 363 | 6002 | 95.80% |

Based on the regression of the above test set, the default prediction model of personal consumption loan is obtained. The lowest value of the default probability of personal consumption loan is 0.5, that is, the probability of exceeding 0.5 is predicted as the default customer, and the probability of less than 0.5 is predicted as the non default customer. According to this principle, the training set is predicted, and the prediction result is the prediction group, and the real result is The observation group can get the real accuracy of the model by forming the confusion matrix of the results of the two parts. The model accuracy includes two parts: one is the prediction accuracy of the defaulting customer, that is, the real default and the predicted default, the other is the prediction accuracy of the non defaulting customer, that is, the predicted non default and the real non default. Both constitute the prediction accuracy of the model.

It can be seen from table 3-3 that the model has a high accuracy in predicting non defaulting customers, but it is not satisfactory in predicting defaulting customers, which also meets the above-mentioned model's type I errors (rejecting the original hypothesis) and type II errors (accepting the wrong hypothesis). In these two types of errors, we prefer to make type I errors, mainly comparing the two types of errors In other words, the cost of making type II mistakes is higher. From the test results, we can see that the improvement direction of this model is to improve the accuracy of forecasting default customers.

The test set is entered into the model formed by the training set, in which the good person represents the loan is returned on time and the bad person represents the loan is overdue. From the above table, it can be seen that the logit model has a comprehensive accuracy rate of 95% in identifying the good person group and the bad person group, so it is feasible to apply the logit model to the loan audit.

## 4. Conclusions and recommendations

### 4.1 From the perspective of government

The government should strengthen the supervision of the P2P network loan platform, clarify the nature of the financial information intermediary of the P2P network loan platform, clearly record the platform, and prevent the P2P network loan companies from using their own business advantages to gather funds and illegally use funds. In order to protect the property security of the

people, it is necessary to prevent P2P network loan companies from losing connection and running away, set up a reasonable early warning mechanism, monitor the operation of the platform in real time. It is strictly prohibited to conduct financial fraud such as illegal fund-raising on P2P online lending platform, and corresponding legal sanctions should be taken for illegal and criminal acts.

### 4.2 From the perspective of P2P network loan platform

The model established in this paper has a limited number of variables. In practice, enterprises can get more data, so as to select variables in a larger range, so as to make more accurate prediction. During the loan review process, the credit status of the applicant can be limited by setting the application threshold. For example, the borrower is required to have a certain number of credit records, determine the amount of the borrower according to the borrower's income, and reject the loan beyond the amount. In the audit process, setting thresholds can reduce the loan default rate, exclude the borrowers with low credit and low qualification, focus on multiple factors can reduce the loan overdue rate and default rate, achieve more accurate positioning, and set the interest rate corresponding to the borrower's loan risk.

With the maturity of Internet technology, more and more risk control operations can be realized by using Internet technology, and finally more reliable credit approval evaluation opinions can be obtained through quantitative evaluation.

P2P financial companies should focus on the construction of internal risk control system, position the company as an unsecured information intermediary, connect the lender and lender through the Internet. The core technology of the company is to collect the information of the lender's loan application, quantify the customer's default risk through rating method through its own screening model, with different ratings corresponding to different interest rates, and grade it The latter loan is provided to the investor. By rating the loan, the lower the risk of the loan, the lower the corresponding interest rate, which not only requires the company to select and filter in the massive data, but also requires the government to establish a more clear social credit system, combine personal information records and personal credit, and provide more for individuals Timely and accurate credit rating, the company can establish a more accurate model and achieve a more accurate loan rating only on the basis of getting correct and sufficient customer information, so that each loan corresponds to the appropriate interest rate.

## References


*[1] Bai Chuanping. Risk monitoring of P2P network platform. Zhejiang: Zhejiang University, 2017*

*[2] Chen Jie. Analysis and Countermeasures of fund raising risk of P2P online lending platform in China. Zhejiang: Zhejiang University, 2018*

*[3] Chen Ruiji. Risk prevention of Internet finance. China's collective economy, 2019 (08): 84-85*

*[4] Han Zhengchen. Generation and development of Internet finance. Business information, 2018, (12): 47, 49*

*[5] Hu Yi, Wang Jue, Yang Xiaoguang. Early warning research on default risk of bank customer loan based on panel logit model. System engineering theory and practice, 2015, 35 (7)*

*[6] Hu zhenkai. Study on credit risk assessment of borrowers in P2P network lending. Heilongjiang: Harbin University of technology, 2016*

*[7] Kang Caijuan. Research on risk management of P2P network lending of Y company. Hunan: Hunan University, 2018*

*[8] Qian Zhengming, Li Haibo, Yu Yanping. Study on default risk of personal housing mortgage loan. Economic research, 2010, 45 (S1): 143-152*

*[9] Qu Qiushi, Li Li. Personal credit risk assessment of commercial banks based on logit model. Commercial economy, 2010, (12): 72-73, 75*

*[10] Wang Lei. Analysis of personal non-performing loans based on probit and logit model. Shanghai University of Finance and economics, 2010*

*[11] Wang Jingyue. P2P network loan default prediction based on user behavior data. Shanghai: Shanghai Normal*

*University, 2017*
*[12] Xu Lida. Research on "de guarantee" of P2P online loan platform. Beijing: Graduate School of Chinese Academy of Social Sciences, 2018*
*[13] Yu Jia. Research on the default behavior of P2P online loan platform based on Panel Data logit model. University of foreign economic and trade, 2017*
*[14] Zhang Jing. An Empirical Study on the causes of personal loan risk of commercial banks based on logit model. Journal of Ningbo Vocational and technical college, 2014, (4): 83-86*