

# Research on the Default Risk Evaluation of Borrowers from Online Peer-to-Peer Lending Based on Survival Analysis

Yaqiong Pan<sup>1,a,\*</sup>, Qiong LIU<sup>2,b</sup>,

<sup>1,2</sup> School of Management, Wuhan University of Science and Technology, 430065 Wuhan, China

<sup>a</sup>panyaqiong2002@sina.com, <sup>b</sup>2420163427@qq.com

**Keywords:** Online Peer-to-Peer Lending; Default Risk; Survival Analysis; Cox Proportional Hazards

**Abstract.** This paper evaluates borrowers' default risk from online peer-to-peer lending and discuss related influencing factors. Based on the theoretical method of survival analysis, adopting semi-parametric estimation method, and using the loan data of Lending Club, Cox proportional hazards model is developed to research empirically. The findings are shown as follows: (1) Term of loan, credit rating, annual income and line recycling rate are protective factors, and with the increase of these variables, the possibility of loan default is reduced. (2) Rate of loan, debt-to-income ratio and number of bad public records are risk factors, and as these variables increase, the possibility of loan default is increased.

## 1. Introduction

Peer-to-Peer Lending (P2P) is a kind of small unsecured lending between individual investors and individual borrowers through the internet platforms. With the booming development of the Internet, P2P have been growing dramatically recently, but the risks especially the default risk of borrowers, cannot be ignored. Default risk, also called credit risk, refers to the possibility that the lender or intermediary platform may suffer losses due to the borrower's unwillingness or inability to perform the contract conditions for various reasons. Most domestic and foreign scholars adopt the traditional methods such as AHP, fuzzy comprehensive evaluation, Logistic regression, neural network, random forest algorithm and factor analysis to evaluate the default risk of online loans. Survival analysis is an analytical method to study survival phenomena and response time data and their statistical laws. It was initially applied in the fields of biology and medicine, introduced into economics and finance recently. The advantage of this method is that it is a dynamic analysis method, which can provide time point prediction, deal with missing data that cannot be observed or collected and expand the scope of sample study.

Hereby, the purpose of this paper is to evaluate default risk of P2P borrowers by adopting survival analysis and using data from Lending Club. Cox's proportional hazards model will be developed to research empirically.

## 2. Literature reviews

### 2.1 Default risk factors identification of P2P

Most scholars identify borrowers' default risk in P2P from "hard credit information" and "soft credit information". Hard credit information is some factors considered in traditional commercial banks' evaluation of borrowers' credit risk, involving financial features (credit grades, Iyer, 2009; FICO scores, Duarte, 2012; debt-to-income ratio, current line of credit, usage of bank card, Klafft, 2009; numbers of credit check, delinquencies, public records, bankruptcy, Emekter, 2015; income, employment status, job, Freedman, 2008), demographic characteristics (race, age, Pope, 2011; gender, Gonzalez, 2014; employment length, house-ownership, Rajan, 2015) and loan features (amount, purpose, rate, term of loan, Qiu, 2012), etc. Soft credit information is nonstandard, ambiguous and subjective judgment information, including social networks (Lin, 2013; Freedman, 2014), borrower's

narrative (Herzenstein, 2011) and third party certification (credit authentication of platforms, Wang, 2014), etc.

## 2.2 Methodology of Survival Analysis and Its Application

Survival analysis is a method to analyze the correlation between specific risk factors and survival time by using statistical theories and methods to explore the time and rule of specific events such as death, disease occurrence, development and mitigation, failure and status persistence, etc.

A few scholars try to introduce survival analysis into the research of P2P to evaluate the survival time of online lending platforms and the default probability of borrowers. Li Xianling (2015) found that the experience of platform operation and the degree of regional industry competition had a significant positive impact on the survival time of the problem platform. Ge Jiankang (2016) adopted life table method to explore the survival time rules of the platform, and established Cox model to conduct early-warning analysis on the platform with operational difficulties. Lin et al. (2009) and Emekter (2015) used Cox model to study the relationship between loan term and default rate, proving that the longer loan term, the higher possibility of default. Li Siyao (2016) established survival analysis model and found that the rate and length of loan had a positive impact on the default risk of P2P online loans, and the geographical location, income, educational background and credit rating of borrowers had a negative impact. Wu Qingtian (2017) used the Cox model to analyze the factors affecting the prepayment of P2P online loans.

## 3. Empirical analysis and results

### 3.1 Variables Explanation

Table 1. Variable description and processing

Name of variable	Variable assignment	Expected direction	Name of variable	Variable assignment	Expected direction
Amount of loan (X1)	Ln (actual amount)	+	Years of working (X7)	<1 year = 0 >10 years = 10	-
Term of loan (X2)	actual loan term	+	House-ownership (X8)	Own=3, Mortgage=2, Rent=1	-
Rate of loan (X3)	borrowing rate*100	+	Debt-to-income ratio (X9)	actual rate	+
Monthly payment (X4)	Ln (actual amount)	+	Line recycling rate (X10)	actual rate*100	+
Credit rating (X5)	A=7, B=6, C=5, D=4, E=3, F=2, G=1	-	Number of arrears exceeding 30 days in the past 2 years (X11)	actual number of times	+
Annual income (X6)	Ln (actual income)	-	Number of bad public records(X12)	actual number of times	+

### 3.2 Sample Selection and Data Sources

This paper takes Lending club, the largest P2P platform in the United States, as the research object. To assess the borrower's default risk, this paper selects the loan transaction data of Lending Club in 2012 for empirical study. The final repayment situation includes Fully Paid /Default/Late (16-30 days)/ Late (31-120 days)/ Charged Off/ In Grace Period, where Fully Paid is regarded as full repayment, and all other states are deemed as overdue repayment, that is default. A total of 53,367 loan records are extracted as the initial sample. After data processing and deletion of missing data, 48180 effective data are obtained. Among them, there are 8000 data of default, and the default rate is about 16.60%.

### 3.3 Model

The Cox proportional hazards model, proposed by D.R. Cox in 1972, is the most widely used function in survival analysis. The model is built as follow:

$$h(t, X) = h_0(t) \exp (\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m) \quad (1)$$

Where:  $X_1, X_2, \dots, X_m$  is the independent variables definite above,  $\beta_1, \dots, \beta_m$  is the partial regression coefficient of each independent variables,  $h(t, X)$  is survival rate,  $h_0(t)$  is the baseline survival function, which needs to be estimated.

### 3.4 Empirical Results

Table 2 illustrates empirical results of Cox proportional hazards model estimated by semi-parametric estimation method using SPSS. B is the regression coefficient of variable, Sig. is the significance level, and Exp (B) represents the relative risk degree. When  $B > 0$  and  $\text{Exp}(B) > 1$ , X is the risk factor, the larger X, the greater risk. On the contrary, When  $B < 0$  and  $\text{Exp}(B) < 1$ , X is the protective factor, the larger X, the lower risk. Where, the Sig. values of variables X2, X3, X5, X6, X9, X10 and X12 are less than 0.05, indicating that these 7 variables have a significant impact on P2P default risk. Therefore, keep the 7 significant variables and eliminate the remaining non-significant variables.

The Sig. values of variables term of loan (X2), rate of loan (X3) and credit rating (X5) are all less than 0.05, which pass the test of significance, indicating that the three variables have significant impact on the survival time of the loan. In particular, the regression coefficient of X2 is -0.895, and its Exp (B) is 0.409, that is  $B < 0$  and  $\text{Exp}(B) < 1$ , showing that X2 is the protective factor. This means that for a particular loan, when other factors keep the same, the longer term of loan, the smaller monthly repayment amount, the less repayment pressure, and the lower corresponding default risk. Similarly, X5 is also the protective factor, which means that the higher credit rating of the borrower, the lower possibility of default. while the regression coefficient of X3 is -0.099, and its Exp (B) is 1.104, namely  $B > 0$  and  $\text{Exp}(B) > 1$ , showing that X3 is the risk factor. This means that the higher interest rate of loan, the higher possibility of default, and the default risk of the borrower with the worst credit rating is 1.104 times of the best credit rating. We also find that the Sig. values of amount of loan (X1) and monthly payment (X4) are all greater than 0.05, which does not pass the significance test, indicating that there is no significant relationship between the survival time of loan and the amount of loan and monthly payment.

Annual income (X6), debt-to-income ratio (X9), line recycling rate (X10) and number of bad public records(X12) all pass the significance test, which indicate that the four variables related to borrower information have a significant impact on survival time of loan. Where,  $B < 0$  and  $\text{Exp}(B) < 1$  of X6 and X10 indicate that these two variables are protective factors, that is, the higher annual income and line recycling rate, the lower corresponding default risk. Since  $B > 0$  and  $\text{Exp}(B) > 1$ , both X9 and X12 are risk factors, indicating that the higher debt-to-income ratio and number of bad records, the higher borrower default risk. But years of working (X7), house-ownership (X8) and number of arrears exceeding 30 days in the past 2 years (X11) does not pass the significance test, indicating that these three variables have no significant correlation with the survival time of the loan.

Table 2. Empirical results of Cox proportional hazards model

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
X1	-0.610	0.387	2.481	1	0.115	0.543	0.254	1.161
X2	-0.895	0.143	39.095	1	0.000	0.409	0.309	0.541
X3	0.099	0.013	57.597	1	0.000	1.104	1.076	1.133
X4	0.704	0.384	3.362	1	0.067	2.022	0.953	4.292
X5			52.297	6	0.000			
X5(1)	-0.141	0.215	0.431	1	0.511	0.868	0.569	1.324
X5(2)	-0.243	0.183	1.761	1	0.185	0.784	0.548	1.123
X5(3)	-0.057	0.158	0.131	1	0.717	0.944	0.693	1.288
X5(4)	0.057	0.129	0.195	1	0.659	1.059	0.822	1.363

X5(5)	0.163	0.096	2.877	1	0.090	1.177	0.975	1.422
X5(6)	0.103	0.068	2.332	1	0.127	1.109	0.971	1.266
X6	-0.454	0.029	250.397	1	0.000	0.635	0.600	0.672
X7	0.004	0.003	1.311	1	0.252	1.004	0.997	1.010
X8			2.184	2	0.336			
X8(1)	0.012	0.044	0.077	1	0.781	1.012	0.929	1.102
X8(2)	-0.025	0.045	0.316	1	0.574	0.975	0.894	1.064
X9	0.006	0.002	12.199	1	0.000	1.006	1.002	1.009
X10	-0.002	0.001	18.056	1	0.000	0.998	0.996	0.999
X11	-0.019	0.018	1.176	1	0.278	0.981	0.948	1.016
X12	0.141	0.056	6.382	1	0.012	1.151	1.032	1.284

#### 4. Conclusion

Based on the theoretical method of survival analysis, 12 variables are selected to establish Cox proportional hazards model, the data of Lending Club are used and semi-parametric estimation method is adopted to conduct an empirical study on the default risk of P2P borrowers. The findings are shown as follows:

(1) Term of loan, credit rating, annual income and line recycling rate belong to protective factor, which have a negative impact on the default risk of P2P. It shows that the possibility of loan default can be reduced with the extension of loan term, the improvement of credit rating, the increase of annual income and line recycling rate.

(2) Rate of loan, debt-to-income ratio and number of bad public records are risk factors, which have a positive impact on the default risk of P2P. It indicates that with the increase of borrowing rate, debt-to-income ratio and the number of bad public records, the possibility of loan default will be increased.

(3) There is no significant relationship between five variables including amount of loan, monthly payment, years of working, house-ownership and number of arrears exceeding 30 days in the past 2 years and the probability of loan default.

#### Acknowledgment

This research was financially supported by Hubei province department of education humanities and social science research project: (Grant NO. 19D012), special soft science research project from Hubei province technical innovation (Grant NO. 2018ADC102) and center for industrial policy and management research (Grant NO. 17CYY09).

#### References

- [1] Lin M.F., Prabhala N R and Viswanathan S, Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending, *Management Science*, vol. 59, pp. 17-35, 2013.
- [2] Emekter R, Tu Y and Jirasakuldech B, Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer (P2P) Lending, *Applied Economics*, vol. 47, pp. 54-70, 2015.
- [3] Li Xianling, Empirical analysis on the causes of failure of P2P Lending Platforms, *Journal of Financial Development Research*, vol. 3, pp. 51-55, 2015.
- [4] Li Siyao, Wang Jitian and Liu Lichao, Study on Influencing Factors of the Default Risk of P2P Lending Network Based on Survival Analysis, *Reform of the Economic System*, vol. 6, pp. 156-160, 2016.
- [5] Wu qingtian and Liu Yan, Study on Influencing Factors of the Prepayment Risk of P2P Online Loan Based on Survival Analysis, *Journal of Commercial Research*, vol. 1, pp. 150-156, 2017.