Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 91 (2016) 168 – 174

Procedia Computer Science

Information Technology and Quantitative Management (ITQM 2016)

Research on Credit Scoring by fusing social media information in Online Peer-to-Peer Lending

Yuejin Zhang[a,] *, Hengyue Jia[a], Yunfei Diao[a], Mo Hai[a], Haifeng Li[a]

[a]School of Information, Central University of Finance and Economics, Beijing 100081, China

Abstract

In recent years, online Peer-to-Peer (P2P) lending market is rapidly expanding in China. In this paper, we use public dataset from PPDai, a leading online P2P platform in China to study the loan default. We construct a credit scoring model by fusing social media information based on decision tree. The experimental result shows that our model has good classification accuracy. From the credit scoring model and classification rules, we get a conclusion that the loan information, social media information, and credit information are most important factors for predicting the default. However, the credit rating is not as important as the platform described.

Peer-review under responsibility of the Organizing Committee of ITQM 2016

Keywords: credit scoring; default risk; decision tree; Peer-to-Peer lending; social media information

1. Introduction

Online peer-to-peer (P2P) lending is a new financing channel which is based on electronic business platform and electronic commerce credit. In P2P lending, borrowers and lenders can use the internet platform to achieve online transactions. There is lower transaction cost, while the loan process is simple and easy to operate. Small and micro enterprises and individual borrowers that are difficult to get loans from the bank do not need loan guarantor and collateral in P2P, so they can get financing more easily. But it means higher credit risk to lenders. Credit risk is the possibility of loss that the bank will suffer after offering loan to the borrowers. It includes not only the actual risk of the borrowers failing to repay the loan on time, but also the potential default risk because of the downgrade of credit or decline of repayment ability of the borrowers. The main reason of credit risk is the asymmetric information between borrowers and lenders [1]. The main method to avoid credit risk in

* Corresponding author. Tel.: 8610-62288896; fax: 8610-62288896.
E-mail address: zhangyj@cufe.edu.cn.

Peer-review under responsibility of the Organizing Committee of ITQM 2016
doi:10.1016/j.procs.2016.07.055

financial institutions is to evaluate the risk by credit scoring [2]. In particular, personal credit scoring is to analyze large amount of data including personal information, credit history and credit behavior to find the characteristics of default and normal borrowers by using data mining technology and statistical analysis method, so as to forecast the credit risk. On the P2P platforms, investors would lend money to a stranger whose personal information may be incomplete, so there is serious information gap.

In this paper we study the online market for P2P in China. In order to reduce the serious problem of information asymmetry between both sides of P2P loans, we make use of social information to describe the behavior characteristics of the borrowers. A person's social behavior and language can reflect the characteristics of their behavior, which can be used as credit data. On the internet, the behavior and language of users can be obtained from social media. Social media refers to website and technology that allow people to write, share, evaluate, discuss, communicate with each other, including social networking sites, micro blog, twitter, micro letters, blogs, forums, and so on. Now, social media have covered almost all aspects of our life, through which we can know the latest developments of a person. Now some credit companies such as Lenddo have already use information of users from Facebook, LinkedIn and Twitter to evaluate credit risk of the consumers. Social data is most useful for people with little or no credit history. A person's social identity, online reputation and professional contacts circle, that should become a factor in the assessment of credit risk.

The main contribution of this paper is to add social behavior factors into the traditional credit scoring model, which are not considered in prior research of credit scoring models. We make an empirical study on PPDai. As a leading platform in China, PPDai has more than 6,000,000 members and 2 billion RMB transaction amount in loans in 2014. We use the public data of borrowers on the platform to create the credit scoring model including the personal information, loan information and social network data. The experimental results show that the credit scoring model based on decision tree can well distinguish between the default and the normal customers, and it has higher forecast accuracy comparing with the model based on logistic regression and neural network.

2. Related Studies

There is little prior research on the credit risk in P2P online lending. The existing literature mainly includes two aspects. The first aspect of research examines the factors in traditional credit scoring models that explain the funding success and default risk. The second aspect of research focuses on the relationship between social media information and default risk.

The main factors in the traditional credit scoring model include the basic information, repayment ability, life stability, credit record, guarantee and some other factors. Herzenstein et al. [3] find that the borrowers' financial strength and their effort when listing and publicizing the loan are more important than demographic attributes for funding success. Qiu et al. [4] show that the personal information, social capital, loan amount, acceptable maximum interest rate and loan period set by borrowers are all significant factors of funding success. Riza et al. [5] use Cox Proportional Hazard regression technique to evaluate credit risk and measure loan performances. They find that credit grade, debt-to-income ratio, FICO score and revolving line utilization play an important role in loan defaults.

Everett [6] analyzes the relationship between the social relationship and the default risk and the interest rate in online P2P lending, and concludes that there is a low default rate between the group members who have actual social relationship on the mutual financing platform. Collier et al. [7] make an empirical analysis of the financing behavior between members of community groups on P2P lending platform, and confirm that by combining individual reputation with the reputation of the group, the group members can supervise each other, which can effectively reduce the adverse selection phenomenon and control moral hazard. Using dataset from Prosper.com, Lin et al. [8] find that the social connections information can effectively solve the problem of asymmetric information in online P2P lending. Their research results show that the borrowers' friendships could increase the probability of successful funding, lower interest rates on funded loans, and these borrowers

are associated with lower ex post default rates. Freedman and Jin [9] also get similar conclusions, and they find that borrowers with social ties are more likely to have their loans funded and receive lower interest rates.

The common methods to establish traditional credit scoring models are divided into statistical methods, operations research methods, artificial intelligence and data mining [10]. Existing researches on online P2P lending mostly use statistical methods. There are rare examines by using other methods. On the other hand, most of these studies use public datasets from Prosper or Lending club, which are leading P2P lending websites of America. There are little works on dataset from other P2P lending platform. In china, online P2P market is developing rapidly. Different from datasets from Prosper or Lending club, the information of public dataset from China is not detailed enough, without FICO score. Therefore by using dataset from PPDai, our study add social behavior factors into the traditional credit scoring model that will provide important signals for investors in Chinese market.

3. Decision Tree

Decision tree is one of the most widely used classification and prediction method of machine learning. Decision tree can deal with both numerical data and non-numerical data (such as gender), so it is very suitable for personal credit rating.

The decision tree is an inverted tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) stores a class label. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. The representation of acquired knowledge in tree form is intuitive and easy to understand by humans. The tree structures are easily converted into the classification rules. The learning and classification steps of decision tree are simple and fast. Generally, decision tree classifiers have good accuracy. Therefore, the decision tree classification algorithm is successfully applied in many areas, such as medical, manufacturing and production, finance, and biology [11].

The main difference between different decision tree algorithms is how to select the attributes and pruning mechanism for creating tree. Based on the concept learning systems Quinlan [12] developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). ID3 adopt a greedy approach, in which decision trees are constructed in a top-down recursive way, and use information gain as an attribute selection measure to ensure that a simple classification tree is obtained.

Set the data partition, *D*, as a training set of class-labeled tuples. Suppose the class label attribute has *m* different values to define *m* different classes $C_i(i = 1, \cdots, m)$. Set $C_{i,D}$ as the set of tuples of class C_i in *D*. $|D|$is the number of tuples in D, and $|C_{i,D}|$ is the number of tuples in $C_{i,D}$. Set node *N* store the tuples of partition *D*. The attribute of the highest information gain is chosen as the splitting attribute of the node *N*.

Firstly, the information entropy (expected information) needed to classify the tuple in *D* is given by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

Suppose we use some attribute *A* to divided tuples in *D*. According to the training dataset, A has v distinct values $\{a_1, a_2, \cdots, a_v\}$. Using attribute *A*, tuples in *D* can be divided into *v* partitions or subsets $\{D_1, D_2, \cdots, D_v\}$. In order to get an accurate classification, the information needs to be measured by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \tag{2}$$

where $\frac{|D_j|}{|D|}$ is the weight of the *j*th partition.

Then, the information gain equals to the difference between the original information demand and the new

demand.

$$Gain(A) = Info(D) - Info_A(D) \tag{3}$$

Both C4.5 and C5.0 algorithms are improved to ID3, in which gain ratio is used to overcome the characteristics that biased toward tests with many outcomes, and they can deal with continuous attributes. The gain ratio is defined as

$$GrainRate(A) = \frac{Grain(A)}{SplitInfo_A(D)} \tag{4}$$

where $SplitInfo_A(D)$ is defined as

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{5}$$

4. Data

We use web crawler to obtain the information of the success loan applications from www.ppdai.com between January 2014 and September 2014. By using the user ID, we also grab the social net information of the loan applicants. According to the repayment period, all of these loans have already been scheduled to expire. According to the traditional credit scoring model and the borrowers' information published on the platform, we choose five aspects of indexes to establish the credit scoring model including basic information, loan information, credit information, social network information, and default information (Table 1). In Table 1, we list the specific variables and the description of every variable.

Table 1. Variable sets used in the model

Variable sets	Variable	Type	Value and description
Basic information	age	Discrete	Age of the borrower 1: Under 20 years old 2: 20-25 years old 3: 26-31 years old 4: 32-38 years old 5: Over 39 years old
	sex	Discrete	Sex of the borrower 1: male 0: Female
Loan information	amount	Continuous	Loan amount in [1000, 400000] yuan
	annual interest rate	Continuous	Annual interest rate of the loan in [8%, 24%]
	repayment period	Continuous	Repayment period in [1, 24] months
Credit information	credit rating	Discrete	There are eight credit ratings, including AAA, AA, A, B, C, D, E, F.

	borrowed credit score	Continuous	Borrowed credit score in [10, 134]
	lending credit score	Continuous	Lending credit score in [10, 151993]
	the number of success	Continuous	The number of success applications in [1, 650]
	the number of failed	Continuous	The number of failed applications in [0, 16]
Social network information	membership score	Continuous	Membership score in [0, 2285]
	prestige	Continuous	Prestige score in [-44,563]
	forum currency	Continuous	Forum currency in [0,1997]
	contribution	Continuous	Contribution score in [-44, 563]
	group	Discrete	The user group that the borrower belongs to
Default information	default	Discrete	Default or not 1: Default 0: Normal

From publicly accessible web page of PPdai we can get limited personal basic information including the age and the sex of the borrowers.

Loan information includes the loan amount, annual interest rate and the repayment period. The annual interest rate of the loan application depends on the credit information of the borrower, the loan amount and the repayment period.

Credit information includes the credit rating, the borrowed credit score, the lending credit score, the number of success applications, and the number of failed applications. For each loan, the loan risk model of PPDai will give a risk score, in order to reflect the forecast of overdue. Each score range will be displayed in the form of letters to the borrowers and the lenders. From low to high, the risk is expressed as AAA to F.

Social network information is obtained from the online forum of PPDai, which mainly contains the membership score, the prestige score, the forum currency, contribution score of the users and the user group that they belong to. The users get the membership score by certificating personal information, sharing information and making comments. According to the users' identity in the forum and their activity, they are divided into different user groups.

Default information is about the loan default. According to the repayment period, the loan samples that we obtained have expired for more than 3 months. So they can be divided into normal repayment records and overdue records according to the loan status.

After data preprocessing, we randomly selected 20,000 records to establish the experimental dataset. Our sample includes 10,000 defaults and 10,000 normal customers, so the ratio of the two classes is 1:1 (Table 2). The ten-fold cross validation method is used to obtain the prediction accuracy. In the experiment, the data set is divided into ten parts, in turn, 9 of which are as training set and 1 as the test set. Finally, the mean of the 10 results is used as the prediction accuracy of our model.

Table 2. Data set

	Number of records	Normal records	Default records
Training set (1-10)	18000	9000	9000
Test set (1-10)	2000	1000	1000
Total	20000	10000	10000

5. Empirical Results

Using the training sets above, we construct the credit scoring model based on decision tree algorithm. In the experiment, all of the datasets are balanced data, so we use accuracy to reflect how well the model recognizes samples of the default class and the normal class. The accuracy of a classifier on a given test set is the percentage of samples that are correctly classified by the classifier. When the class distribution is relatively balanced, accuracy is the most efficient. The results of the ten experiments on training set and test set are shown in Table 3. From Table 3 we can see that each average predict accuracy rate of test sets is more than 80%, which is a good result for the actual data. Especially, 82.39% of the potential default borrowers can be identified. The result can be an important signal for the lenders in online P2P lending.

Table 3. Predict accuracy of credit scoring model based on decision tree

Data set	Accuracy of Training set			Accuracy of test set		
	Normal	Default	Total	Normal	Default	Total
1	84.22%	86%	85.14%	78.4%	81.2%	79.8%
2	82.74%	87.81%	85.28%	79.5%	84.5%	82%
3	84.78%	84.96%	84.87%	83%	81.6%	82.3%
4	83.98%	85.47%	84.72%	79.5%	80.2%	79.85%
5	82.93%	86.48%	84.71%	79.7%	80.9%	80.3%
6	83.08%	86.13%	84.61%	79.7%	83.5%	81.6%
7	84.26%	85.60%	84.93%	80.7%	83.4%	82.05%
8	83.47%	86.12%	84.79%	80.2%	82.9%	81.55%
9	83.98%	85.17%	84.57%	80.1%	82%	81.05%
10	83.49%	86.44%	84.97%	79.6%	83.7%	81.65%
Average	83.69%	86.02%	84.86%	80.04%	82.39%	81.22%

In addition to the decision tree, there are other methods commonly used to establish credit scoring models. So we also use logistic regression and neural network to construct the credit scoring model with the same attributes in Table 1. The experimental results and comparisons are shown in Table 4. Note that among the three methods, credit scoring model based on decision tree has higher prediction accuracy.

Table 4. Comparisons with other methods

Classifier	Accuracy of Training set			Accuracy of test set		
	Normal	Default	Total	Normal	Default	Total
Decision tree	83.69%	86.02%	84.86%	80.04%	82.39%	81.22%
Logistic regression	73.04%	82.27%	77.65%	72.80%	81.40%	77.10%
Neural Network	80.11%	76.23%	78.17%	79.30%	75.50%	77.40%

6. Conclusion

In this paper, we used public dataset from PPDai to study the loan default. Being different from datasets from America, our dataset is not detailed enough, without FICO score. In order to solve the asymmetric information between borrowers and lenders, we added social behavior information into the variable sets to build the credit scoring model. The experimental result showed that although the dimension of the dataset is still not high, our model has good classification accuracy. Additionally, comparing with other classification methods such as logistic regression and neural network, decision tree is more powerful in predicting the potential default borrowers.

On the other hand, from the credit scoring model and classification rules, we found that borrowed credit score, the number of success, prestige, the number of failed, repayment period, forum currency are most important attributes for predicting the default. In these factors, prestige and forum currency are social network information, repayment period is loan information, and others are credit information. However, the credit rating is less important than the six attributes above, which is the main decision-making basis provided by the platform for the lenders. Actually, some borrowers with credit rating of AA default in the end. Therefore, we suggest that the lenders of PPDai do not only reference to the credit rating. In addition, we hope that our work can make help to improve the risk control model of PPDai and other online P2P platforms.

Acknowledgements

This research has been partially supported by 121 of CUFE Talent project Young doctor Development Fund (No. QBJ1427), National Natural Science Foundation of China (No.61309029, No.61100112, No. 71401188), and Beijing Higher Education Young Elite Teacher Project (YETP0987, YETP0988).

References

[1] J.E. Stiglitz, A. Weiss, Credit Rationing in Market with Imperfect Information[J], The American Economic Review, 1981, 71(3), 393-410

[2] Allen N.Berger, W.Scott Frame, Nathan H.Miller, Credit Scoring and the Availability, Price, and Risk of Small Business Credit[J]. The Journal of Money, Credit and Banking, 2005, 37(2).

[3] M. Herzenstein, R. Andrews, U. Dholakia, et al. (2008) The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities, Working Paper. Available at SSRN www.prosper.com (accessed 30 September 2014).

[4] J. Qiu, Z. Lin, and B. Luo, Effects of borrower defined conditions in the online peer-to-peer lending market. E-life: web-enabled convergence of commerce, work, and social life, Lecture Notes in Business Information Processing, 2012, 108, 167–179.

[5] Riza Emekter, Yanbin Tu, Benjamas Jirasakuldech & Min Lu, Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending, Applied Economics, 2015, 47:1, 54-70,

[6] C. Everett, Group membership, relationship banking and loan default risk: the case of online social lending, SSRN Electronic Journal 03/2010. http://www.researchgate.net /publication /228200235

[7] B. Collier, Hampshire R., Sending Mixed Signals: Multilevel Reputation Effects in Peer-to-Peer Lending Markets [J]. Decision Support System, 2010, 49: 52-70.

[8] M. Lin, N R Prabhala, S. Viswanathan, Judging Borrowers by the Company They Keep: Friendship Networks and Information Asymmetry in Online Peer-to-Peer Lending[J]. Management Science, 2013, 59(1):17-35.

[9] S. Freedman, and G. Jin, The information value of online social networks: lessons from peer-to-peer lending, NBER Working Paper No. 19820. Available at http://www.nber.org/papers/w19820 (accessed 30 September 2014).

[10] L.C. Thomas, A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, 2000, 16, 149–172.

[11] J.W. Han. Micheling Kamber. Jian, Data Mining Concepts and Techniques (Third Edition). Beijing: China Machine Press. 2012. 213-220.

[12] J. R. Quinlan. Induction of decision trees. Machine Learning, 1986, 1:81–106.