Journal of Computers Vol. 31 No. 2, 2020, pp. 83-100
doi:10.3966/199115992020043102009

Research on Default Risk of Peer-To-Peer Online Lending Based on Data Mining Algorithm

Xiao-feng Li[1], Chang Zhang[1,2*], Xu-chen Lin[1], Ting-jie Lv[1], Lin-lin Liu[1,3]

[1] Beijing University of Posts and Telecommunications, Beijing 100876, China
{airmagicz, happy-bee}@163.com

[2] Shi Jiazhuang Posts and Telecommunications Technical College, Shijiazhuang 050021, China
airmagicz@163.com

[3] Weifang University, Weifang 261061, China
693616004@qq.com

Received 5 June 2018; Revised 17 October 2018; Accepted 18 December 2018

Abstract. Online Peer-to-Peer (P2P) lending market has experienced a period of rocketing development far beyond our expectation, which is also facing more challenges, such as the default on a loan. The study aims to explore a data-driven approach to extract knowledge of default risk from borrowers' demographic and the behavior characteristics in the loaning process, which can be used to reduce the default risk of P2P platform. The possibility of credit risk rating automation can also be investigated by estimating the predicting accuracy. A huge dataset from a famous P2P lending platform in China was analyzed, and three default prediction models were employed for the data research of discrete input-output pairs, continuous input-output pairs and continuous input and discrete output pairs. The average percent hit rate (APHR) and the lift analysis were adopted to evaluate the predictive accuracy. A 2-layer artificial neural network (ANN) model has performed brilliantly with continuous input-output data pairs with an average relative-error value of 0.24. The support vector machine (SVM) is highly recognized due to a predictive accuracy of 89.18% with discrete input-output data pairs. Decision tree C5.0 (DT) was utilized to find some important factors affecting risk rate and predict the default risk of the borrower. The behavior data such as delinquency history, the already paid installment and loan remaining was indispensable on obtaining a higher predictive accuracy. Some constructive conclusions on risk management of P2P online lending can be drawn based on data mining results.

Keywords: artificial neural network (ANN), credit risk, decision tree C5.0 (DT), peer-to-peer lending, support vector machine (SVM)

1 Introduction

With the rapidly development of Internet, those business activities on internet-based platforms become a new style. The online peer-to-peer (P2P) lending market is a platform, that individuals make loan without the involvement of traditional financial institutions. Benefiting from the on-line platform, P2P lending can provide more attractive interest rate for lenders and lower transaction cost for borrowers comparing with the traditional bank and financial institution. Consequently, the P2P lending has achieved a prominent development [1]. In August 2017, the monthly turnover of the P2P lending industry exceeded 200 billion yuan in China.

However, transactions in P2P platform sometimes occur anonymously, in other words, information is asymmetrically distributed between borrowers and lenders [2]. For example, some platforms seek to attract customers (borrowers) by improving the user's experience, streamlining the loan process and

[*] Corresponding Author

sparing proof of qualification and collateral provided by borrower. The information asymmetry between borrowers and lenders may lead to some problems, such as moral hazard and adverse selection. Therefore, it would reduce the transaction efficiency and increase the transaction cost [3]. In fact, the online P2P lending market is facing more risks and obstacles like the default on the loan. In China, there is no unified credit scoring system even in conventional commercial bank and other financial institution [4]. Since the absence of personal qualification certificates and financial privacy, the likelihood of the borrower defaulting (ie, moral hazard) is raised to some extent. The borrower's default will lead to a shortage of the platform fund. In order to maintain the normal operation of the platform, the manager of the platform will generally increase the loan rate to compensate for certain bad debts. The increase in the interest rate will enhance the probability of inferior borrowers entering the platform and reduce the probability of high-quality borrowers entering the platform. That is adverse selection. Then the market status becomes ineffective. With the increase in loan rate, high-quality borrowers will voluntarily give up borrowing because they are unable to repay their loans on time. Inferior borrowers do not have any repayment plan and sensitivity to the loan rate, so they just continue to borrow. With the loss of a large number of high-quality customers and the increasing of a large number of inferior borrowers, the bad debts of the platform continue growth. That would make the platform capital flow broke. The lender could not withdraw cash from the platform according to the agreement, which eventually leads to the collapse of the platform. Until June 30, 2018, the number of P2P online lending platforms in China has reached 6,631. But 4,485 of them have problem. In all problem platforms, 81.71% of them have difficulty repaying their lender or loss of communication. To prevent the occurrence of adverse selection, the P2P platform will cooperate with Internet oligarchy companies such as Baidu, Tencent, Ali, telecommunication operators such as China Mobile, China Unicom, China Telecommunications, logistics companies such as SF-express, Green hand, commercial banks such as ICBC, CCB. These companies provide borrowers' risk assessments for P2P platforms based on their own accumulated data of borrowers' behavior. However, the data owned by these data giants has not been shared with each other. In addition, the risk assessment models used by different companies are various. So that the credit level of the same borrower is inconsistent. In order to obtain comprehensive credit information, the P2P platform has to cooperate with several companies at the same time. These actions will greatly increase the cost of P2P platform undoubtedly. Furthermore, in majority developed countries, financial privacy is rigorously protected by government, and it seems unlikely that lenders will be given access to acquire customer personal information in the near future. Therefore, It is quite valuable to estimation and forecasting borrowers' default risk using borrowers' demographic and behavior characteristics in the loan process. P2P platform can collect those data by themselves without a huge investment. On the other hand, the accuracy prediction can alleviate adverse selection and help lenders to allocate their investment more efficiently [5].

The estimation and forecasting methods employed for predicting credit risk can be classified into certain types, such as statistical methods (SM), regression analysis (RA), mathematical programing (MP). Discriminant analysis (DA) is the originally used statistical method to evaluate credit risk. These statistical models are easy to explain. But it needs strict assumptions to get higher accuracy, which makes it difficult to apply to reality. For example, the multivariate normality assumptions for independent variables are frequently violated in financial data sets, which makes these methods theoretically invalid for finite samples. RA is one of the prediction models widely used. For instance, [6] analyses the relationship between demographic characteristics and default risk by binary logistic regression (BLR) based on a new dataset of P2P, and finds that gender, age, marital status, educational level, delinquency history are contributory variables. Mangasarian [7] proposes Linear Programing in pattern separation first. Then Shi et al. [8] extends it to multiple criteria linear programming and applies this approach in credit card portfolio problem. But MP is optimal for small samples and sometimes the convergence is slow.

With the progress of artificial intelligence (AI), a lot of intelligent algorithms are designed to solve the linear and nonlinear problems in financial field [9-11]. AI approach performs splendidly compared with traditional statistical approach [12]. ANN is a data driven model comprising a number of interconnected neurons. Because of the powerful capability of integrating complicated non-linear functions, ANN could explore the underlying relationship only by input-output data pairs [13]. SVM is a novel machine learning technique, which achieving an excellent predictive accuracy and firstly introduced by [14]. As a "black box" model, SVM is difficult to be understood by human. But it's excited to find that the information in the "black box" can be disclosed by sensitivity analysis. It can provide us the importance-analysis of particular input attributes in the model. DT model is another classification method with "if-then" rules. It becomes more accuracy and easily understood by human.

Each algorithm has its own feaures and applicable situation. It is meaningful to select an algorithm with perfect matching. And that would improve the efficiency and effectiveness of the system. Some studies have been shown in Table 1. [5, 10, 15-18] focus on the default risk prediction of P2P lending by various methods. [5] explores the P2P loan characteristics and evaluates the credit risk by BLR method. The results show that the likelihood of the loan default increases with the decreasing of credit risk of the borrowers. [15] proposes a credit scoring model using ANN and LR approach. The peer-to-peer loan applications have been classified into default and non-default groups. The NN-based credit scoring model performs effectively in screening default applications. Before the fund is proposed, [10] predicts the credit risk of the P2P loan using a data mining (DM) approach. Five DM models have been constructed and the prediction performance are similar. [16] proposes survival analysis approach, which predicting survival probabilities of loans at different time periods. It is effective in predicting the survival periods of the loans. [17] evolves a LR prediction model by univariate means tests and survival analysis. [18] proposes a default prediction model combined hard information with soft information for P2P lending through random forest (RF) method. The results show that the soft information can improve the recognition rate of loan default. [19] shows that Multi-layer perceptron (MLP) model has an exciting performance in automatic credit scoring systems. The comparison of prediction accuracy of personal credit scoring models has been done in [20-23]. Classification and regression tree (CART) and NN outperform the traditional credit scoring models in terms of predictive accuracy and type II errors [20]. [21] demonstrates that both the mixture-of-experts (MOE) and radial basis function (RBF) models should be considered in credit scoring applications. LR is the most accurate of the traditional methods. [22] builds several non-parametric credit scoring models based on the MLP approach and benchmarks their performance against other models which employ the traditional linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and LR techniques. The results reveal that NN models outperform the other three classic techniques. [23] conducts a personal credit assessment model by ANN and LR models, and the results show that the combine of the two models can lift the classification accuracy. The credit scoring models for firms have been done in [12-13, 24-25]. [12] shows that both SVM and ANN approaches have acquired excellent predictive accuracy. Some works have been done in bank default prediction, and ANN has an excellent performance on predictive accuracy [13]. [24] finds that NN may be more suitable to predict the probability of company bankrupt. However, [25] finds DT has achieved a higher accuracy comparing with ANN and SVM. A comparative analysis of the four DM models has been made and it is found that ANN has achieved the best result [26].

Table 1. Related prediction studies using various techniques

Study	**objects**	Methods	Accuracy	Data	Variables	Sample size
[5]	Predict default risk of loan for P2P lending	BLR		USA Lending Club	credit grade, debt-to-income ratio, FICO score and revolving line utilization	61451 (7 credit grade based on the FICO score
[15]		ANN LR	NN (62.7%; 74.38%) LR (65.34%; 61.03%)	European Bondora	age, country, applied amount, loan duration, use of loan, application type and so on.	4748 (two categories)
[10]		CART CHAIN MLP RBF SVM	CART (71.23%) CHAIN (70.9%) MLP (71.24%) RBF (68.11%) SVM (72.05%)	USA Lending Club	the term of loan, annual income, the amount of loan, debt-to-income ratio, credit grade and revolving line utilization	12517 (three categories)
[16]		SA NN LR		European Bondora	verification type, age, gender, income total, country, rating, marital status, employment status, loan duration and so on.	32714
[17]		SA LR	LR (62%-80%)	USA Lending Club	loan purpose, annual income, current housing situation, credit history and indebtedness	24,449

Table 1. (continue)

Study	**objects**	Methods	Accuracy	Data	Variables	Sample size
[18]		RF	RF (76.02%)	Chinese eloan	hard characteristic, soft characteristic (borrower and loan soft information)	15806 (two categories)
This paper	*P2P lending default risk prediction*	*ANN* *SVM* *DT C5.0*	*ANN (76%)* *(continuous input-output pairs)* *ANN (80.02%)* *SVM (82.31%)* *DT C5.0 (80.81%)* *(continuous input-discrete output pairs)* *ANN (79.02%)* *SVM (89.18%)* *DT C5.0 (81.75%)* *(discrete input-output pairs)*	*Chinese P2P lending platform*	*borrowers' demographic and the behavior characteristics in the loaning process*	*10907*
[20]	Personal credit assessment	DA LR NN CART	DA (65.23%, 62.00%) LR (66.37%, 62.33%) NN (78.85%, 61.52%) CART (72.89%, 65.58%)	Turkish bank	gender, age, marital status, educational level, occupation, job position, income, customer type and credit cards	1260
[21]		DA LR NN KNN KD CART	DA (85.96%) LR (87.25%) NN (87.78%) KNN (85.8%) KD (83.34%) CART (84.38) (Australian data)	German Australian	credit history, account balances, loan purpose, loan amount, employment status, personal information, age, housing, and job	German (1000) Australian (690)
[22]		LDA QDA LR MLP	LDA (93.03%) QDA (91.98%) LR (93.22%) MLP (94.59%)	Peruvian microfinance institution	financial ratios, zone, old, previous loan granted, previously granted credits, previously denied loans, total fees and so on.	5500
[23]		LR BPNNLR	LR (86.25%; 84.87%) BPNNLR (86.33%; 87.96%)	China Unicom	amount of payment arrearage, online duration, accumulative days of communication, number of intranet communicational friends, number of outer communicational friends and so on.	4518 (two categories)
[12]	Firm credit rating prediction	SVM BNN	BNN (80%) SVM (80%)	United States Taiwan	financial ratios, total assets, total liabilities and so on	United States (265 cases, five categories) Taiwan (74 cases, five categories)
[13]		NN DA LR KNN ID3	NN (89.5%) LD (85.8%) LR (85.2%) KNN (75.3%) ID3 (80.8%) (Jackknife Method)	USA	capital, asset, management, equity, liquidity and so on.	118 banks
[24]		NN DA	NN (81.5%) DA (75%)	Moody's Industrial Manuals	working capital / total assets, retained earnings / total assets, earnings before interest, taxes / total assets, market value of equity / total febt x5 sales / total assets and so on	129 firms (two categories)

Table 1. (continue)

Study	**objects**	Methods	Accuracy	Data	Variables	Sample size
[25]		LR RBF SVM CART C5	LR (79.8%) RBF (79.8%) SVM (66.1%) CART (89.8%) C5 (93.7%)	USA Compustat database	total asset, book value per share, inventories, liabilities, receivables, cost of goods sold, total dividends, earnings before interest and taxes, gross profit (loss), net income (loss), operating income after depreciation, total revenue, sales, dividends per share and total market value	1321 records (two categories)

As a summary, many studies of personal credit risk assessment and firms bankrupt risk prediction have been carried on [12-13, 20-25]. However, the research is rather scarce in the field of the online P2P lending by DM approach [27]. The aim of this paper is to test the possibility and potential of default risk prediction of P2P lending by DM approach. The work is to investigate how borrowers' demographic characteristics and the behavior data in the loaning process can be properly utilized in P2P platform to establish data driven models. And the model can extract knowledge from the selected attributes and make a prediction with good accuracy.

In this paper, 7 models have been built to forecast the borrowers' default risk using demographic characteristics data and the behavior data in the loaning process. These models are employed to predict the default risk of borrowers. The average percent hit ratio (APHR), the lift analysis and average relative errors could evaluate the predictive accuracy. The excellent results mean that these models could solve such problems. It also reveals that the P2P platform data are valuable and need more attention. The accuracy of these models is improved by adding the process data such as delinquency history, the already paid installment and loan remaining. In other words, the processing data would provide more useful information for managers and lenders to make decisions efficiently.

The main innovative contributions of this study are:

Firstly, the determining attributes selection by principal component analysis (PCA) has been experimented, which may help lenders understand the contributory attributes to default risk.

Secondly, a huge dataset from a famous P2P lending platform in China has been analyzed, which broadened the study of credit risk analysis. Until then, most previous researches utilized the data from P2P lending platforms in the U.S.A (such as Prosper and the Lending club).

Thirdly, The ANN model was utilized to predict the default risk by continuous input-output data pairs and the prediction accuracy was evaluated by average relative errors. Different selection strategies for the number of layers and nodes of the black box in ANN are tried here. Furthermore, The ANN, SVM and DT models were applied to predict the credit level with discrete input-output data pairs and continuous input and discrete output data pairs. As a result, the experiment output is highly satisfied, which is worthy to assist managers to make better decisions.

Fourthly, the models of the default risk were established by borrowers' demographic characteristics and the behavioral data in the loaning process. These data can be collected and analyzed by P2P platform itself without a great cost. The prediction result will be conductive to lenders to make an optimal investment strategy, and the managers of P2P platforms could design effective supervision mechanism.

The paper is organized into 6 sections. In the section 2, the data used in the study has been described and summarized. The section 3 has described the approaches to evaluate the default risk of borrowers. In the section 4, the empirical results have been presented. The section 5 has drawn significant conclusions of the whole article. Finally, in the section 6, some discussions and future direction have been made based on the mining results.

2 Data

2.1 Data Preparation

In this section, the data used in the study are described and summarized, including the loan status and the characteristics of loan applicants. The research involves 11,198 loan applications from June 1, 2017 to October 31, 2017. The data are collected from www.yooli.com, which is a famous P2P lending platform in China. Firstly, a data preprocessing and data screening job is conducted by replacing the vacant values with median values and removing useless attributes like ID, contact code and outlier values. For instance, in some loan cases, the age of a borrower is over 100 or borrower's monthly income is more than 100,000 RMB. The true and valid data have been classified into two categories: demographic characteristics attributes (age, gender, marital status, children status, education level, monthly income, loan amount, installment, debt-to-income ratio, monthly repayment, company size, working age) and loaning process attributes (delinquency history, already paid installment, loan remaining). Eventually, 10,907 valid loan cases have been acquired. The total amounts of all these loans are approximately 24.30 million RMB. As shown in Table 2, 22.24% of all the loan applications, in other words, a total amount of 5.40 million RMB will be lost, because the borrowers of these loans have not been paid back in time (either fully or partially). 3.92% of these loans have fully paid, which constituted about 0.20 million RMB. About 18.69 million RMB loans accounting for 76.94% has been in current state. From the perspective of lenders, the most essential concern is the potential default of borrowers in the future.

Table 2. Loan distribution by loan status

Loan status	Number of loan	Per cent	Amount	Per cent
Late in payment	2236	20.50	5,403,550	22.24
Current	8243	75.58	18,693,193	76.94
Full paid	428	3.92	200,315	0.82
Total	10907	100.00	24,297,058	100.00

Lenders will benefit a lot, if some personal characteristics can help to predict the default likelihood of the borrower. Based on the descriptions of loan status, during the study time the platform has provided a total of 10,907 loans, of which 2236 were late payment. The default loans can be converted into a loan default rate of 20.50%. Table 3 provides an overview of the data attributes. According to the sample of 10,907 loan applicants, there are 5,200 borrowers unmarried and only 12.13% of borrowers have default history in the closed periods. The average repayment period is 13.3 months and a borrower needs to pay 285 RMB per month, with a monthly income of 3761 RMB. The average loan amount and debt-to-income ratio is 3360 RMB and 0.088, respectively. On average, the loan remaining is 2209.33 RMB. Male accounts for 62.54%, while female accounts for 37.46%. 48.68% of borrowers are working in small companies with less than 20 employees, 35.40% of borrowers are working in medium companies with 20 to 100 employees, and the remaining 15.92% of borrowers are working in large companies with more than 100 employees. As for borrowers' education level, 2442 borrowers' education experience are below high school, accounting for 22.39%; 5011 borrowers have high school degrees, accounting for 45.94%; 3454 people who achieve Bachelor's degree or above contribute 31.67% of all borrowers. The average already paid installment are 4.4. The average age of borrowers is 27 years old with a mean working age around 4 years.

Table 3. Descriptive statistics

	N	Minimum	Maximum	Mean	SD	Skewness
Marital status	10907	0	1	0.262	0.440	1.085
Children status	10907	0	1	0.228	0.419	1.299
Monthly income	10907	800	10000	3875.49	1819.97	1.552
Installment	10907	6	24	13.31	5.15	0.77
Monthly payment	10907	37.42	1462.67	284.79	137.32	1.588
Debt-to-income ratio	10907	0.007	0.681	0.087	0.056	2.072
Loan amount	10907	800	9976	3360.05	1056.41	0.16
Loan remaining	10907	0	7737.05	2209.33	1041.65	0.040
Gender	10907	0	1	0.375	0.484	0.518
Company size	10907	5	1000	116.73	236.91	2.99
Education level	10907	0	1	0.609	0.274	0.159
Already paid installment	10907	1	24	4.37	2.32	6.23
Delinquency history	10907	0	1	0.021	0.144	6.637
Working age	10907	2	38	4.02	2.46	5.46
Age	10907	20	57	27.19	6.07	1.70

Table 4. Rotated principal components loadings for default risk

Attributer	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Marital status	0.815	0.350		-0.275	0.058	-0.118		0.121		0.122
Children status	0.810	0.343		-0.284	0.051	-0.130		0.137		0.127
Monthly income	0.556	-0.074	0.224	0.495	-0.127	0.330		-0.174	-0.394	-0.126
Installment	0.211	-0.866	0.104	-0.257	0.056		0.052	0.075	0.132	
Monthly payment	-0.223	0.692	0.507	0.193	-0.152	0.108			-0.101	
Debt-to-income ratio	-0.575	0.582	0.317	-0.199		-0.144		0.136	0.199	0.107
Loan amount		-0.094	0.814	-0.105	-0.096	0.245	0.165	0.152		
Loan remaining	0.079	-0.519	0.720	-0.115	-0.149	-0.125		0.079	0.080	
Gender	0.073	-0.144	-0.100	0.698	-0.146	-0.059		0.426	0.265	0.437
Company size	-0.120	-0.071	0.105	0.204	0.656	-0.229	-0.100	0.503	-0.180	-0.380
Education level	-0.233	-0.090	0.276		0.634			-0.312	-0.287	0.527
Already paid installment		0.070	-0.246	-0.182	0.300	0.752	0.385	0.240	0.110	
Delinquency history				0.143		-0.378	0.895	-0.151	-0.057	-0.090
Working age	0.296	0.125	0.228	0.274	0.688	0.058	-0.068	-0.403	0.630	-0.190
Age	0.665	0.213		0.094	0.134	-0.076		0.051		
Variance%	17.781	14.604	12.099	8.231	7.946	6.803	6.657	5.972	5.519	4.821
Cumulative %	17.781	32.385	44.483	52.715	60.661	67.464	74.121	80.093	85.612	90.433
Eigen value	2.667	2.191	1.815	1.235	1.192	1.020	1.001	0.896	0.828	0.723

2.2 Feature Selection

Principal component analysis (PCA) is used to extract significant attributes (principal components: PCs) to analyze the relationship between observed variables. In this section, PCA method has been utilized to find the important attributes. The relatively important attributes could be extracted through importance ranking provided by principal component model. 15 attributes have been selected as the model input features. As shown in Table 4, variation maximum rotations were used to yield a ranked series of attributes on the 10 principal components along with the amount of variance for each component. The first seven principal components account for more than 74% of the total variation. The variables that PCA identified as the primary attributes responsible for the first seven PCs were selected as important attributes for data mining.

Table 5 demonstrates the selected attributes. According to the final dataset, the selected attributes of each loan case include marital status, children status, monthly income, installment, monthly payment, debt-to-income ratio, loan amount, loan remaining, gender, company size, educational level, already paid installment, delinquency history, age and working age.

Table 5. Final set of selected attributes

No.	Attributes	Type	Description
1	Marital status	nominal	Marital status includes unmarried, married
2	Children status	nominal	Children status includes no kid and having kid.
3	Monthly income	numerical	The monthly income provided by the borrower.
4	Installment	numerical	Installment includes 6, 12, 24.
5	Monthly payment	numerical	The monthly payment provided by the lender.
6	Debt-to-income ratio	numerical	A ratio represents the proportion of the borrower's total monthly repayment on total monthly income.
7	Loan amount	numerical	The amount of loan applied by the borrower.
8	Loan remaining	numerical	The amount on the loan.
9	Gender	nominal	Gender includes female and male.
10	Company size	numerical	The number of staff provided by the borrower.
11	Education level	nominal	The capability of learning included below high school, high school, undergraduate and postgraduate.
12	Already paid installment	numerical	The number of paid up loan.
13	Delinquency history	numerical	Whether there was overdue previously.
14	Age	numerical	Age provided by the borrower.
15	Working age	numerical	Working age provided by the borrower.

3 Methodologies of Data Mining

3.1 Data Mining Models

In this study, ANN, SVM and DT C5.0 have been used and compared to each other by accuracy measurements. IBM SPSS Modeler 18.0 software has been employed to construct the credit risk models.

ANN model. As an approach widely adopted by artificial intelligence, ANN model has been studied for more than two decades. Many researches have focused on the understanding of the brain functionality, with the purpose of acquiring human-like performance in many domains of knowledge engineering. The processing elements of an ANN model are called neurons and the interconnected neurons constitute layers. A typical network comprises the input layer, hidden layer and the output layer [28].

The adjustable connect weight, W_i, will multiply the input neuron X_i, which is summed at each neuron. There is a threshold value b, which is added to each neuron. The combined input is then passed through a transfer function, f, including linear and nonlinear, to produce the output of the neuron, Y. The output of one neuron provides the input to the next neuron. Equation (1) shows us the process clearly [28].

$$Y = f(\sum_{i=1}^{n} w_i x_i + b) = f(W * X + b)\,. \tag{1}$$

In equation (1), $X = [x_1, x_2, \cdots, x_n]^T, W = [w_1, w_2, \cdots, w_n]$.

The widely used linear transfer functions includes POSLIN and PURELIN functions, while the typical nonlinear transfer functions comprises LOGSIG and TANSIG functions. Classically, the most familiar and effective function in the case is the Logarithmic Sigmoid function defined as equation (2)

$$f = \frac{1}{1+e^{-av}}\,. \tag{2}$$

Where v is the input sum for a neuron and a is a constant ranging from 0.01 to 1.00 (Javad and Narges 2010).

In this study, demographic characteristic and loaning process attributes data are used as the input and rating outcome as the output.

SVM model. The SVM is an intelligent algorithm which could transfer the input space into a high m-dimensional space. The kernel function makes the input data more separable compared with the original input space. Then, the SVM finds the best linear separating hyperplane to separate the data. It is reasonable to assume that a lower error means a large parallel distance between these hyperplanes [28].

In recent years, there has been explosive development in different aspects of SVM, including theoretical understanding, algorithmic strategies for implementation and real life application [29-30]. Due to the complexity, SVM model often achieves accurate prediction, but can't be understood easily by human. There are two interesting ways to open the "black box", rule extraction and sensitively analysis [31-32].

DT model. DT has been becoming increasingly popular, which is a powerful classification algorithm with the capability of extracting rules from massive attributes data.

Popular DT algorithms in the field of data mining include classification and regression tree (CART), chi-squared automatic interaction (CHAID), iterative dichotomiser 3 (ID3), and C4.5/C5.0. Attributes of data applied in the CART are continuous, while the CHAID and ID3 algorithm can only handle discrete data. Both C4.5 and C5.0 are supervised learning classification algorithm, which can tackle continuous data. In this study, the DT C5.0 algorithm is applied to make a prediction of default risk.

3.2 Evaluation

To evaluate the prediction performance, two evaluation methods have been employed: APHR analysis and the lift analysis.

APHR analysis. APHR analysis is popular as a performance metric for classification methods [33]. It is clear that a bigger value of APHR means a better prediction accuracy. APHR can be summarized in Equation (3).

$$APHR = \frac{a}{b} . \tag{3}$$

Where *a* is the number of samples correctly classified and *b* is the total number of samples.

Lift analysis. The lift analysis is widely employed to estimate the essence of establishing classification models. A useful lift cumulative curve is obtained by plotting the population samples versus the cumulative percentage of real responses captured. Obviously, 1.0 corresponds to ideal method, while 0.5 means a random baseline.

4 Empirical Results

4.1 ANN Models for Continuous Output

Table 6 shows the different parameters in ANN model. The mean square error (MSE) is considered as a standard measurement in a training set.

Table 6. Final set of selected attributes

ANN parameters
Network type: Feed-forward back-propagation network
Number of hidden layers: 1, 2, 3
Transfer function: Log-Sigmoid

Equation 4 shows the calculation method of MSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[p-a]^2 . \tag{4}$$

Where *p, a* and *n* represent the predicted output, the actual output and the number of input-output data pairs, respectively [28].

Table 7 shows the relative average errors for ANN model with different parameters. In general, A 2 layered networks gives the best prediction accuracy, opposed to the 3 or 4 layer networks in the present investigation. Equation 5 shows the calculation method of the average relative error.

Table 7. List of neural networks

No.	Network details	MSE (training)	MSE (test)	Average relative error (test)	Prediction accuracy
1	15-[5]-1	0.034	0.034	0.25	0.75
2	15-[6]-1	0.033	0.035	0.25	0.75
3	15-[7]-1	0.032	0.034	0.24	0.76
4	15-[8]-1	0.034	0.034	0.25	0.75
5	15-[9]-1	0.033	0.035	0.25	0.75
6	15-[10]-1	0.033	0.035	0.25	0.75
7	15-[15]-[7]-1	0.033	0.035	0.26	0.74
8	15-[10]-[7]-1	0.033	0.035	0.25	0.75
9	15-[7]-[7]-1	0.033	0.036	0.26	0.74
10	15-[30]-[15]-[7]-1	0.034	0.035	0.26	0.74
11	15-[20]-[10]-[7]-1	0.035	0.036	0.27	0.73
12	15-[15]-[7]-[7]-1	0.036	0.038	0.28	0.72
13	15-[10]-[7]-[7]-1	0.036	0.037	0.27	0.73

$$Average\ relative\ error = \left[\sum_{i=1}^{n} \frac{1}{n} \left\{ \frac{abs(p-a) \times 100}{a} \right\} \right]. \quad \textbf{(5)}$$

Where *p, a* and *n* represent the predicted output, the actual output and the number of input-output data pairs, respectively [28].

The errors suggest the network with a15-[7]-1 architecture is the optimum one with an average relative error value of 0.24.

4.2 ANN, DT and SVM Models for Discrete Output

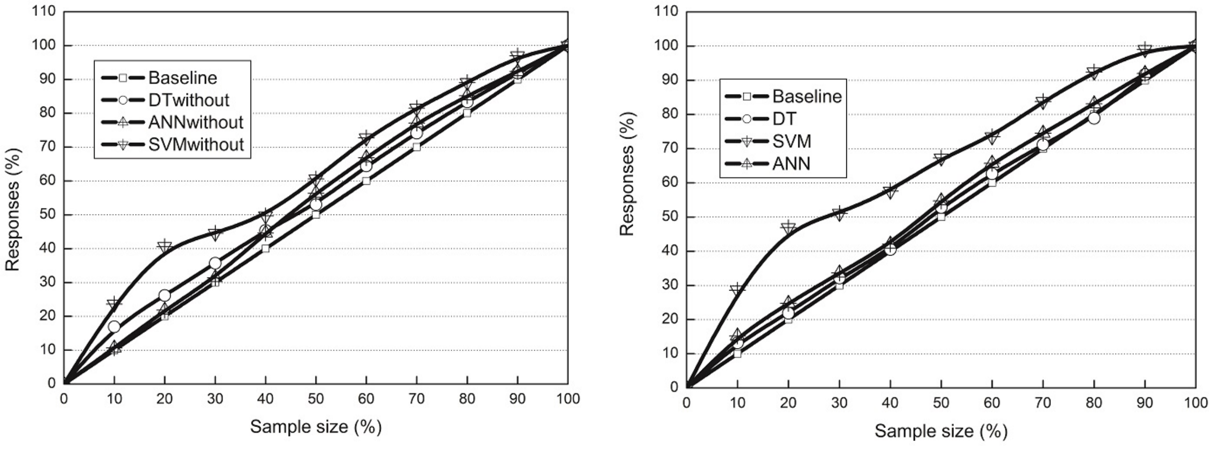
In this section, the selected attributes have been classified into two categories: process attributes (delinquency history, the already paid installment and loan remaining) and non-process attributes (installment, debt-to-income ratio, children status, education level, monthly income, monthly repayment, marital status, company size, loan amount, age, gender and working age). The predictive accuracy of these three models (DT, SVM and ANN) have been compared. At the same time, the importance of these attributes have been investigated.

ANN, DT and SVM models with continuous input value. Table 8 summarizes the prediction accuracy of the three models (DT, SVM and ANN) without and with process attributes by APHR analysis. The SVM has achieved the best performance with a predictive accuracy of 79.87% without process attributes and 82.31% with process attributes. It is explicit that the process attributes lift the prediction accuracy.

Table 8. The prediction accuracies (APHR) of the three models without and with process data under continuous input and discrete output data pairs

	DT (%)	SVM (%)	ANN (%)
APHR (without PD)	78.68	79.87	77.43
APHR (with PD)	80.81	82.31	80.02

Fig. 1 shows the lift cumulative curves for the three models (DT, SVM and ANN) without process attributes (a) and with process attributes (b). Obviously, the area of SVM is the largest, meanwhile, the area of DT and ANN are almost equal.



(a) and with process attributes (b) under continuous input and discrete output data pairs

Fig. 1. Lift cumulative curves for the three models (DT, SVM and ANN) without process attributes

Table 9 provides the examples of DT, SVM and ANN lift cumulative response values without process attributes (a) and with process attributes (b).

Table 9. The examples of DT, SVM and ANN lift cumulative response values without process attributes (a) and with process attributes (b) under continuous input and discrete output data pairs

Sample size (%)	DT (%)	SVM (%)	ANN (%)	DT (%)	SVM (%)	ANN (%)
	Without process attributes			With process attributes		
10	16.91	23.64	10.65	13.20	28.62	15.21
20	26.16	40.61	21.83	21.85	46.90	24.73
30	35.69	44.55	31.40	32.17	51.03	33.64
40	45.38	49.70	44.61	40.52	57.59	42.03
50	53.03	60.61	56.33	52.81	67.24	54.69
60	64.45	72.73	66.85	62.67	73.45	65.75
70	74.13	81.52	77.09	71.17	83.79	74.50
80	83.38	89.10	85.18	78.91	92.41	83.10
90	91.91	96.97	92.32	91.65	98.97	92.01

Fig. 2 shows the relative importance of each variable for SVM and ANN models. The process attributes such as delinquency history, the already paid installment make an important contribution to the prediction accuracy of default risk. The non-process attributes such as marital status, company size, working age, also play an important role in the accurate prediction of loan default.

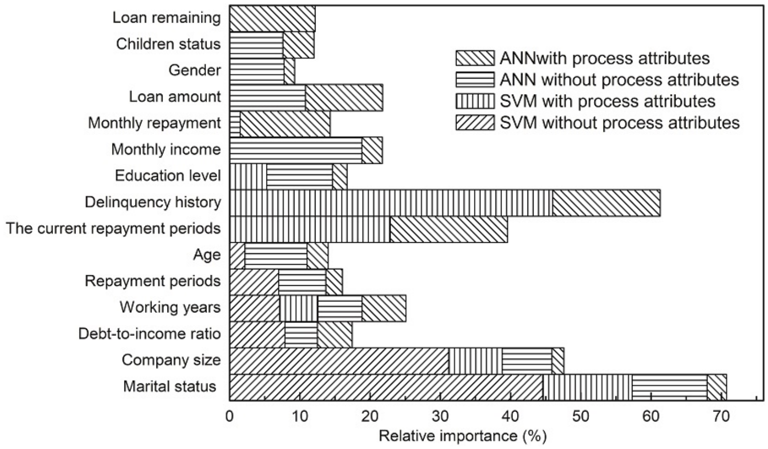


Fig. 2. The relative importance of each attribute for models under continuous input and discrete output data pairs

ANN, DT and SVM models with discrete input value. In order to simplify the model, increase the stability of the model and reduce the computational complexity, the alternative indicators are preprocessed, and some numerical alternative indicators are grouped into categories. As shown in table 10, monthly income was divided into lower (below 3000), medium (from 3000 to 5000) and high (over 5000) and monthly payment was divided into small (below 200), medium (200 to 400) and large (over 400). Loan amount was divided into small (from 800 to 4000), medium (4000-7000) and large (over 7000). Loan remaining was divided into mini (below 1500), small (from 1500 to 2500), medium (from 2500 to 4000) and large (over 4000).

Table 10. Description of selected attributes

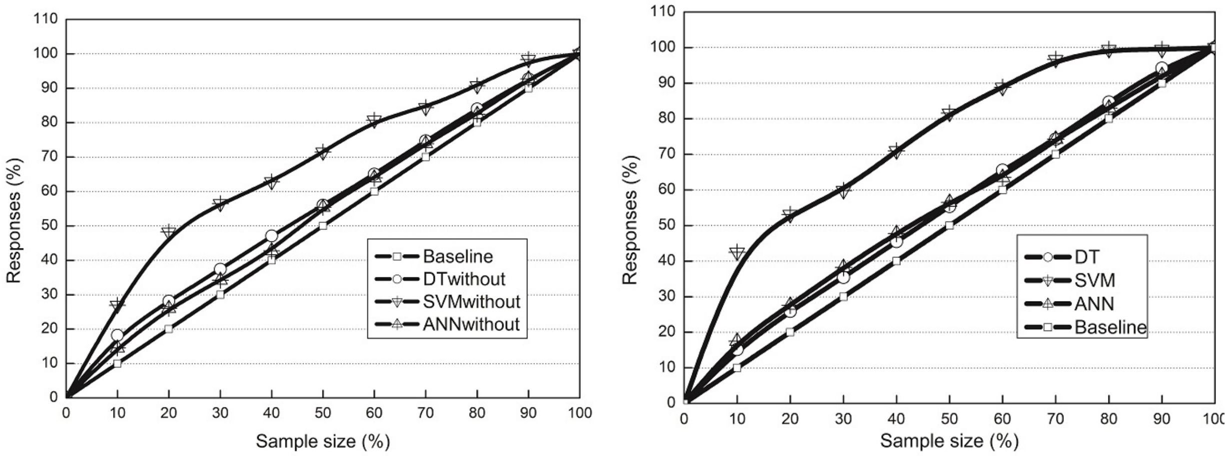
No.	Attributes	Description
1	Marital status	Marital status includes unmarried, married, assigning value 0, 1, respectively.
2	Children status	Children status includes no kid and having kid, assigning value 0, 1, respectively.
3	Monthly income	Monthly income was divided into lower, medium and high, assigning value 0, 1, 2, respectively.
4	Installment	Installment includes 6, 12, 24, assigning value 0, 1, 2, respectively.
5	Monthly payment	Monthly payment was divided into small, medium and large, assigning value 0, 1, 2, respectively.
6	Debt-to-income ratio	Debt-to-income ratio was divided into below.5%, from 5% to 10%, from 10% to 30%, over 30%, assigning value 0, 1, 2, 3, respectively.
7	Loan amount	Loan amount was divided into small, medium and large, assigning value 0, 1, 2, respectively.
8	Loan remaining	Loan remaining was divided into mini, small, medium and large, assigning value 0, 1, 2, 3, respectively.
9	Gender	Gender includes female and male, assigning value 0, 1, respectively.
10	Company size	Company size was divided into small company, medium company and large company, assigning value 0, 1, 2, respectively.
11	Education level	Education level was divided into below high school, high school and undergraduate or postgraduate, assigning value 0, 1, 2, respectively.
12	Already paid installment	Already paid installment was divided into below 3, from 3 to 6 and beyond 6, assigning value 0, 1, 2, respectively.
13	Delinquency history	Delinquency history (whether there was overdue previously) includes no default record and having default record during the repayment process, assigning value 0, 1, respectively.
14	Age	Age was divided into below 30 years old and over 30 years old, assigning value 0, 1, respectively.
15	Working age	Working age was divided into below 5 years and over 5 years, assigning value 0, 1, respectively.

Table 11 summarizes the prediction accuracy of these three models (DT, SVM and ANN) without and with process attributes. The SVM has achieved the best performance with a prediction accuracy of 82.08% without process attributes and 89.18% with process attributes. The process attributes strongly lift the prediction accuracy.

Table 11. The prediction accuracies (APHR) of the three models with process data under discrete input and output data pairs

	DT (%)	SVM (%)	ANN (%)
APHR (without PD)	79.22	82.08	77.78
APHR (with PD)	81.75	89.18	79.02

Fig. 3 plots the lift cumulative curves for the three models (DT, SVM and ANN) without process attributes (a) and with process attributes (b). Obviously, the area of SVM is the largest, meanwhile, the area of DT and ANN are almost equal.



(a) and with process attributes

(b) under discrete input and output data pairs

Fig. 3. Lift cumulative curves for the three models (DT, SVM and ANN) without process attributes

Table 12 provides the examples of DT, SVM and ANN lift cumulative response values without process attributes (a) and with process attributes (b).

Table 12. The examples of DT, SVM and ANN lift cumulative response values without process attributes (a) and with process attributes (b) under discrete input and output data pairs

Sample size (%)	DT (%)	SVM (%)	ANN (%)	DT (%)	SVM (%)	ANN (%)
	(a) Without process attributes			(b) With process attributes		
10	18.24	26.91	14.49	15.16	42.46	17.45
20	28.11	48.17	25.99	25.89	53.07	27.61
30	37.37	56.48	34.23	35.43	59.78	38.20
40	47.09	62.79	42.76	45.49	70.95	47.78
50	56.05	71.43	55.40	55.37	81.56	56.51
60	65.02	80.73	63.92	65.58	88.83	63.52
70	74.74	84.39	73.86	74.28	96.65	74.25
80	84.01	90.70	82.37	84.67	99.44	82.98
90	92.68	100	92.76	94.21	99.44	92.13

Fig. 4 shows the relative importance of each variable for models. From Fig. 4, it could be concluded that the process attributes such as delinquency history and the already paid installment play an important role in the prediction accuracy of loan defaults. The non-process attributes such as gender, debt-to-income ratio, age are critical factors on the prediction of credit risk.

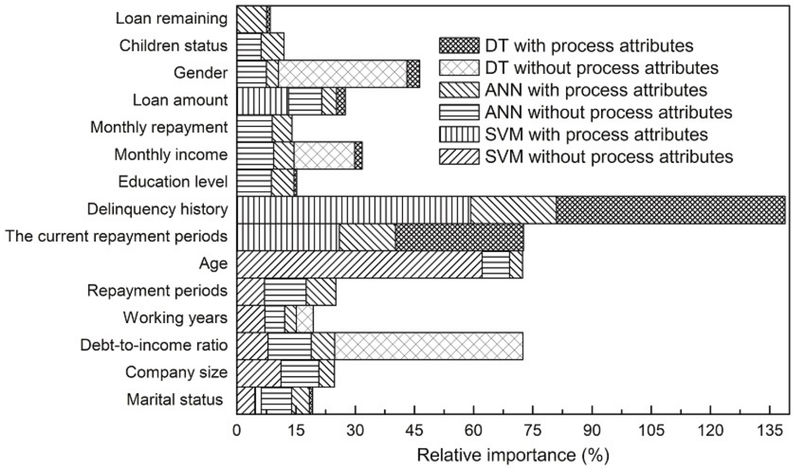


Fig. 4. The relative importance of each variable for models under discrete input and output data pairs

DT C5.0 model analysis results:

(1) If a borrower has delinquency history, then his probability of loan default will increase dramatically. Managers or lenders of the platform should not lend money to the one who has delinquency history.

(2) If a borrower does not have delinquency history and the already paid installment >2, then the probability of default falls to 17.3%.

(3) If a borrower does not have delinquency history and the already paid installment≤2 and the gender is female, then the default risk is low. However, if the gender is male, the probability of default is high.

(4) If a borrower does not have delinquency history and the already paid installment≤2 and the loan remaining is medium (from 2500 to 4000) or high (over 4000), then the default risk is high.

(5) If a borrower does not have delinquency history and the already paid installment≤2 and the loan remaining is mini (below 1500) and the debt-to-income ratio is medium (from 10% to 30%) or high (over 30%), then the default risk is low. However, if a borrower without delinquency history and the already paid installment≤2 and the loan remaining is small (from 1500 to 2500) and the debt-to-income ratio is small (from 5% to 10%), then the default risk is high.

(6) If a borrower does not have delinquency history and the already paid installment≤2 and the loan remaining is mini (below 1500) and the debt-to-income ratio is medium (from 10% to 30%) and education level is high school or undergraduate or postgraduate, then the default risk is low. However, if the education level is below high school, then the default risk is high.

(7) If a borrower does not have delinquency history and the already paid installment≤2 and the loan remaining is mini (below 1500) and the debt-to-income ratio is medium (from 10% to 30%) and working age is over 5 years, then the default risk is low.

(8) If a borrower does not have delinquency history and the already paid installment≤2 and the loan remaining is mini (below 1500) and the debt-to-income ratio is medium (from 10% to 30%) or high (over 30%) and marital status is married, then the default risk is low. However, if the marital status is unmarried or divorced, then the default risk is high.

(9) If a borrower does not have delinquency history and the already paid installment≤2 and the loan remaining is small (from 1500 to 2500) and the debt-to-income ratio is medium (from 10% to 30%) and children status is having kid, then the default risk is low. However, if children status is having no kid, then the default risk is high.

5 Conclusions

In the present study, SVM, DT C5.0 and ANN models have been built to predict the credit risk of P2P loan. All of these models have achieved satisfactory results. The high prediction accuracy obtains the conclusion that DM approach is effective and efficient to make default risk prediction for P2P platform by borrowers' demographic characteristics and loaning process characteristics data. Especially, the dataset comes from this P2P platform. The data driven models have extracted knowledge from the selected attributes and made a prediction with good accuracy.

(1) The adjusted ANN is an effective tool for predicting the borrower's default risk with continuous input-output attributes data. The optimum ANN architecture includes 15 neurons in the input layer, 7 neurons in the hidden layers, one neuron in the output layer respectively, and the average relative error value is 0.24. The capability of handling continuous input-output means the excellent performance in predicting accuracy. For example, the DT and SVM models only provide discrete target prediction. If two borrowers both are rated 0 by DT or SVM models, a lender can not decide which one he will choose. Because he can not distinguish preciously the difference between two borrowers with the same credit level. However, the lender can apply ANN model, which provides the continuous output value to screen a better borrower.

(2) The SVM is highly recognized due to an exciting prediction performance. The SVM model has achieved the best performance by discrete input-output data pairs, with a predictive accuracy of 82.08%, opposed to DT with 79.22% and ANN with 77.78% without process data; a predictive accuracy of 89.18%, opposed to DT with 81.75% and ANN with 79.02% with process data. Moreover, the SVM model has outperformed other models by continuous input and discrete output data, with a predictive accuracy of 79.87%, opposed to DT with 78.68% and ANN with 77.43% without process data; a predictive accuracy of 82.31%, opposed to DT with 80.81% and ANN with 80.02% with process data.

(3) DT model is utilized to analyze the relationship between the default risk and selected attributes. Some useful and innovative knowledge have been provide for administrators, and the platform should pay more attention to the borrower who is having delinquency history, junior, male, unmarried, without kid, with low debt-to-income ratio and in low education level. For example, The more working age, the better the borrower's financial status will be. Meanwhile, people who have worked for a long time are more likely to realize the importance of credit. So that, the default rate of a loan decreases with the increasing of the borrower's working age. It is found that the woman displays a common trait of less risk-seeking behavior than man, which is consistent with the results under [6, 34]. In other words, male is more risky. A borrower who married and having kid has a better credit level. A borrower with low debt-to-income ratio may not care about the economic cost of default, because the cost of overdue payment is relatively low. Then, the default probability of the borrower is higher. The improvement in the education level means the higher improved monthly income, that may decrease the default rate. As a conclusion,

Each algorithm has its own characteristics and applicable situation. If the target attributes data are all continuous, internet financial institutions can do accurate predictions by using ANN method. Some different data can be used to conduct the qualitative analysis. If the target attribute values of the collected data are all discrete, SVM, DT and ANN methods can be used to do qualitative prediction to guide the decision-making. The SVM method has the highest accuracy here. However, if internet financial institutions would like to get some detailed and distinctive prediction results, DT methods is powerful, which can discover some predictable knowledge.

(4) The borrowers' behavior attributes stem from the platform can improve the prediction accuracy. The process attributes such as delinquency history, the already paid installment and loan remaining play a significant role in evaluating credit risk. For instance, there is a lifting of about 7.10% by adding loan process information into the SVM model.

6 Discussion and Future Direction

In this work, DM approaches have made default risk prediction of P2P platform effective and efficient by borrowers' demographic characteristics and loaning process characteristics data. As we all know, the accuracy prediction will alleviate adverse selection and decrease transaction cost. So, it's imperative to establish a data warehouse based on decision analysis, which can collect, preprocess, share and analyze the dynamic and massive data of borrowers efficiently. The demographic characteristics data can be utilized to evaluate credit risk before fund is proposed, and then, the behavior data in the loaning process such as delinquency history, the already paid installment and loan remaining can be employed to supervise the borrowers.

In the research, the demographic characteristics data include age, gender, marital status, children status, education level, monthly income, loan amount, installment, debt-to-income ratio, monthly repayment, company size and working age. There are some false information in the dataset, which may influence the prediction accuracy of default risk. So the manager or the lender should design proper mechanism, which could motivate borrowers to provide the true and valid personal information such as their background, their financial standing and the purpose of the loan. For example, the platform can provide a higher discount on loan interest rate for those borrowers who upload the certification materials. More attention should be paid on the borrower who provide false or vacant information. In general, the borrower who provides false information or vacant information may be more risky. Because he or she has lower default cost than the borrower who provides real information.

The process attributes play an important role in credit rating, so it is critical important to collect and track process behavior data of borrowers. The effective mechanism should be designed to decrease the default efficiency and increase the transaction cost. For example, the DT show us that when the already paid installment >2, the probability of default falls to 17.3%. This is mainly due to the decreasing in the repayment amount as the repayment period increasing. The borrower's transformation cost increases. As the number of installment increases, the time cost and capital cost of the borrower increase. Stop repayment means that all of these cost become sunk cost. So, the wise choice for borrowers is to repay. In addition, more attention should be paid to track the borrower's payment behavior, and effective mechanism should be designed to motivate the borrower to repay on time in an early phase. For instance, the platform can provide a floating interest rate for the borrower. In the beginning-three months, the repayment interest is free or low. A high overdue charge and a high interest will be paid by the borrower once the overdue behavior happens. A borrower repayment on due can increase the credit amount or enjoy other preferential policy.

In the end, there are still some limitations of the work. Although the authentic dataset of P2P platform is used to do analysis and some conclusions are drawn, the total data quantity for research is still insufficient and the data quality is not perfect also, which may lead to the analysis results inaccurate. At present, the massive, unstructured and dynamic data is really difficult to be mined. Firstly, massive data is hard to be calculated by existing data mining software or computer compiling system because of the limitation of algorithm performance. Secondly, complex and changing data is difficult to be modeled quickly and effectively, even leading to the failure of data algorithm model. So the performance improvement of the data mining and data warehouse system is an important direction of data mining application in future, which makes data processing more effectively and efficiently. In addition, some optimizational and well-targeted data mining algorithms and more reasonable distributed processing methods should be created to make data analysis more effectively and accurately. The data mining approaches and technology will be increasingly developed in future.

Acknowledgments

This work is supported by Major Program of the National Social Science Foundation of China under Grant No.15ZDB154, and the science and technology youth fund project of colleges and universities of Hebei province under Grant No. is QN2018318. The contents and conclusions of this work are based on deep study of sub-topic of the research project.

References

[1] S-C. Berger, F. Gleisner, Emergence of financial intermediaries in electronic markets: the case of online P2P lending, Business Research 2(1)(2009) 39-65.

[2] W. Dobbie, P.-M. Skiba, Information asymmetries in consumer credit markets: evidence from payday lending, American Economic Journal Applied Economics 5(4)(2013) 256-282.

[3] G.-A. Akerlof, The market for “lemons”: quality uncertainty and the market mechanism, Quarterly Journal of Economics 8(4)(1970) 488-500.

[4] S.-C, Carlos, G.-N. Begoña, The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending, Decision Support Systems 89(2016). DOI:10.1016/j.dss.2016.06.014.

[5] R. Emekter, Y. Tu, B. Jirasakuldech, M. Lu, Evaluating credit risk and loan performance in online peer-to-peer (P2P) lending, Applied Economics 47(1)(2015) 54-70.

[6] X.-C. Lin, X.-L. Li, Z. Zheng, Evaluating borrower’s default risk in peer-to-peer lending: evidence from a lending platform in China, Applied Economics 49(2017) 1-8.

[7] O.-L. Mangasarian, Linear and nonlinear separation of patterns by linear programming, Operations Research 3(13)(1965) 343-514.

[8] Y. Shi, Y. Peng, W.-X. Xu, X.-W. Tang, Data mining via multiple criteria linear programming: applications in credit card portfolio management, International Journal of Information Technology & Decision Making 1(1)(2002) 131-151.

[9] H.-J. Chiang, C.-C. Tseng, C.-C. Torng, A retrospective analysis of prognostic indicators in dental implant therapy using the C5.0 decision tree algorithm, Journal of Dental Sciences 8(2013) 248-255.

[10] Y. Jin, Y.-D. Zhu, A data-driven approach to predict default risk of loan for online peer-to-peer (P2P) lending, in: Proc. Fifth International Conference on Communication Systems and Network Technologies, 2015.

[11] B. Widrow, D.-E. Rumelhart, M.-A. Lehr, Neural networks: applications in industry, business and science, Communications of the Acm 37(3)(1994) 93-105.

[12] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, S.-S. Wu, Credit rating analysis with support vector machines and neural networks: a market comparative study, Decision Support Systems 37(2004) 543-558.

[13] K.-Y. Tam, M.-Y. Kiang, Managerial applications of neural networks: the case of bank failure predictions, Management Science 38(7)(1992) 926-947.

[14] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag New York, New York, 1995.

[15] A. Byanjankar, H.-K. Markku, J. Mezei, Predicting credit risk in peer-to-peer lending: a neural network approach, in: Proc. 2015 IEEE Symposium Computational Intelligence, 2015.

[16] A. Byanjankar, Predicting credit risk in peer-to-peer lending with survival analysis, in: Proc. 2018 IEEE Computational Intelligence, 2018.

[17] C. Serranocinca, B. Gutiérreznieto, L. Lópezpalacios, Determinants of default in P2P lending, Plos One 10(10)(2015) e0139427.

[18] C.-Q. Jiang, R.-Y. Wang, D. Yong, The default prediction combined with soft information in online peer-to-peer lending, Chinese Journal of Management Science 2017(11)(2017) 22-32.

[19] Z.-Y. Zhao, S.-X. Xu, B.-H. Kang, M.M.J. Kabir, Y.-L. Liu, R. Wasinger, Investigation and improvement of multi-layer perceptron neural networks for credit scoring, Expert Systems with Applications 42(7)(2015) 3508-3516.

[20] H. Ince, B. Aktan, A comparison of data mining techniques for credit scoring in banking: a managerial perspective, Journal of Business Economics & Management 10(3)(2009) 233-240.

[21] D. West, Neural network credit scoring models. Computers & Operations Research 27(11)(2000) 1131-1152.

[22] A. Blanco, P.-M. Rafael, J. Lara, Credit scoring models for the microfinance industry using neural networks: evidence from Peru, Expert Systems with Applications 40(1)(2013) 356-364.

[23] Y.-J. Huo, H.-Z. Chen, J.-C. Chen, Research on personal credit assessment based on neural network-logistic regression combination model, Open Journal of Business and Management 5(2017) 244-252. DOI:10.4236/ojbm.2017.52022.

[24] D.-L. Olson, D. Delen, Y.-Y. Meng, Comparative analysis of data mining methods for bankruptcy prediction, Decision Support Systems 52(2012) 464-473.

[25] M.-D. Odom, R. Sharda, A neural network model for bankruptcy prediction, in: Proc. 1990 IJCNN International Joint Conference on Neural Networks, 1990.

[26] S. Moro, P. Cortez, P. Rita, A data-driven approach to predict the success of bank telemarketing, Decision Support Systems 62(2014) 22-31.

[27] H.-P. Hu, B.-H. Huang, H.-B. Yao, Z.-D. Lu, R.-X. Li, Identifying local trust value with neural network in P2P environment, Journal of Chinese Computer Systems 27(2006) 1503-1505.

[28] G. Javad, T. Narges, Application of artificial neural networks to the prediction of tunnel boring machine penetration rate, Mining Science and Technology (China) 20(2010) 727-733.

[29] S. Begum, D. Chakraborty, R. Sarkar, Cancer classification from gene expression based microarray data using SVM ensemble, in: Proc. CATCON- 2015 IEEE 2nd International Conference, 2015.

[30] L.-F. Ren, W.-J. Wang, An SVM-based collaborative filtering approach for Top-N web services recommendation, Future Generation Computer Systems 78(2018) 531-543.

[31] P. Cortez, M.-J. Embrechts, Using sensitivity analysis and visualization techniques to open black box data mining models, Information Sciences 225(2013) 1-17.

[32] B. Yeo, D. Grant, Predicting service industry performance using decision tree analysis, International Journal of Information Management 38(2018) 288-300.

[33] R. Sharda, D. Delen, Predicting box-office success of motion pictures with neural networks, Expert Systems with Applications 30(2006) 243-254.

[34] M. Powell, D. Ansic, Gender differences in risk behaviour in financial decision-making: An experimental analysis, Journal of Economic Psychology 18(1997) 605-628.