The Application of Machine Learning Algorithms in Predicting the Borrower's Default Risk in Online Peer-to-Peer Lending

Huixi He[a], Kazumitsu Nawata[b]

The University of Tokyo

[a]huixihe@g.ecc.u-tokyo.ac.jp, [b]nawata@tmi.t.u-tokyo.ac.jp

Keywords: logistic regression, random forest, neural networks, naive bayes, P2P lending; default risk

Abstract: The machine learning algorithms have excellent performance in classification, regression, clustering, and visualization, which should also be applicated in finance studies such as bank failure prediction, risk management, and so on. This paper focuses on the application of machine learning techniques in predicting the default risk of an individual borrower in a P2P lending platform based on data from one of the most famous platforms. We use four widely accepted classification algorithms to evaluate the loan risk of a certain borrower by predicting whether this borrower will overdue. In the data processing step, we highlight our work in applying *LDA* in text analytics. We found that logistic regression, Random Forest, Neural Networks, and Naive Bayes all performed well with the precision above 86%, and Naive Bayes reaches 97%.

1. Introduction

The machine learning techniques have been extensively accepted recently, especially in the field of finance. Meryem et al. (2009) provide a comprehensive review of studies which used machine learning techniques to measure bank performance. They noticed that neural networks, support vector machines, and logistic regression analysis are commonly used in classification researches [1]. However, the development of network information makes lenders to provide loans to borrowers directly; that is, one transaction can be finished without the traditional financial institutions such as banks. Electronic platforms, which post information about loan applications online and put lenders in contact with borrowers, has emerged recently in China with great potential to increase the effectiveness of financial markets [2] [3].

However, there are few types of research in the assessment of the risk of borrowers as well as the default prediction of online peer-to-peer (P2P) lending. Lin et al. (2013) found that social networking relationship is an essential role in measuring default risk, that is, a powerful social networking relationship may lower the risk of loans [4]. The study of Agarwal et al. (2015) focus on the factors that may influence the necessary time of successfully closed loans, and they got a conclusion that loans with guarantees are easy to be matched, however, such loans may have higher default rate [5]. Guo et al. (2016) worked on the risk of loans in P2P lending and proposed an assessment model to evaluate credit risk based on the risk of borrowers [6].

Unlike existing researches described above, the purpose of this article is to evaluate the default risk of a specified borrower based on machine learning tools such as logistic regression, Random Forest, Neural Networks and Naive Bayes. Besides, because of the information that borrowers provide online has many textual descriptions, Natural Language Processing algorithm is also used in the step of the data process.

The rest of the article is structured as follows. Section 2 describes the data used here and introduces the algorithm used to deal with it. Section 3 discusses the machine learning algorithms that can be applied for classification problems. Section 4 presents the results of experiments using four algorithms. Summarizes about the whole study and further researches are discussed in Section 5.

  DOI: 10.25236/icemit.2019.016

2. Data

2.1 Data Description

Information about borrowers has been collected from one of the largest P2P platforms (renrendai) in China, which has nine years of entrepreneurial history, and the total transaction amount has reached 85.7 billion RMB. This research uses 48,120 loan applications form 1 January 2016 to 31 December 2018 crawled from the site of renrendai that provide some important private information about the borrowers such as social status, credit history, education level, family background and so on (https://www.renrendai.com/loan.html). 252 loan applications have been excluded because of variables loss; that is, 47868 samples left to be analyzed. The summarize of the loan status is presented in Table 1. As shown in Table 1, 0.52% of all the borrowers have been marked as "Bad Debt" and "Failed," which means that 642 million RMB has lost, while 68.39% of all loan applications finished successfully. However, because this P2P lending platform we discussed here was founded mainly by private capital, the number of losses is relatively too much to bear, which need to prevent in advance. Besides, loans marked as "In progress" mean that these loans are still in the repayment period, which means that at this time, they cannot be judged as default or not. Therefore, this part of the data was deleted, and only 32,986 loans entered the next step.

Table 1. Loan Status

Status	Number	Percentage
Bad Debt	237	0.5
Failed	10	0.02
Closed	32739	68.39
In progress	14882	31.09

Depending on the data characteristics, variables can be divided into four groups: 1. Variables described by numbers, such as "age," "already pay count" and so on; 2. Variables represented by Boolean, such as "has car" and "has house"; 3. Variables assigned as a single word, such as "gender," "education level," and so on; 4. Variables represented by complex texts. For variables assigned by number, the descriptive statistics of data is shown in Table 2. It can be seen from the Table 2 that borrowers are willing to borrow with a repayment period of 3 years, and the interest rate is 10.2%, which is much higher than the base rate of the central bank (4.35%).

For Boolean variables and variables expressed by a single word, which means that features are not continuous values but categorical values, this study uses One-Hot Encoding to deal with this problem. One-Hot Encoding can binarize the categorical features, which can be embedded as a vector in the Euclidean space [7]. The reason why we use this method here is that in classification problems, the calculation of the distance between features is critical, and the similarity calculation is based on the European space as well. Variables processed by this method are: "borrow type", "has car loan", "credit level", "gender", "education level", "has car", "has house", "has house loan", " work industry", "company size", "company type", "salary" and "work years". "Salary" and "work years" are also processed by One-Hot Encoding because the values of these two variables are numerical intervals rather than absolute numbers.

2.2 Natural Language Processing

Natural Language Processing algorithm is also used here for data processing cause the existence of a variable that is assigned as text. Borrowers who want to get the loan will submit a self-introduction that including work information, social status, purpose of loan, and repayment plans to lenders. A common self-introduction can conclude as: "I am an office worker in Beijing, now engaged in the business services industry with a steady income. The purpose of this loan is house decoration. The information above has been certified by the platform." This self-introduction contains a lot of information that is important for classification, so we put it into consideration.

The algorithm applied in this paper to deal with these texts is *Latent Dirichlet Allocation*, which is first introduced by David M. Blei et.al (2003) [8]. *LDA* is a generative probabilistic model of a corpus,

which consists of three levels: words, documents, and corpora. The essential opinion of *LDA* is that every topic of one document is represented by a multinomial distribution of words, documents are formed from the combination of topics, and the corpus is characterized by documents. In other words, each of the words were generated.

Table 2. Data Description

	N	mean	std	min	25%	50%	75%	max
Age	32986	39.00	8.57	25.0	32.0	37.0	45.0	65.0
Already pay count	32986	1.10	0.358	0.0	1.0	1.0	1.0	10.0
Available credits	32986	48691	53,473	0.0	0.0	41,500	80,500	617,500
Borrow amount	32986	88224	58,116	0.0	47,400	77,850	111,100	701,000
Failed count	32986	0.009	0.097	0.0	0.0	0.0	0.0	2.0
Interest rate	32986	10.12	0.431	8.0	10.2	10.2	10.2	13.0
Loan term(month)	32986	33.41	6.064	3.0	36.0	36.0	36.0	48.0
Overdue count	32986	0.159	1.581	0.0	0.0	0.0	0.0	35.0
Overdue amount	32986	456.38	6,798	0.0	0.0	0.0	0.0	258,027
Successful count	32986	1.16	0.526	1.0	1.0	1.0	1.0	15.0
Credit point	32986	177.6	16.99	0.0	180.0	180.0	180.0	223.0

By selecting a topic in a multinomial probability, and then choose a word from this topic. In terms of topic, it is assumed that the prior distribution is the Dirichlet distribution within a document. For a document d, the distribution of a topic of it can be described as:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta) \quad (1)$$

$$p(\theta|\alpha) = \text{Dirichlet}(\alpha) \quad (2)$$

α, β Are parameters, θ refers to the joint distribution of topics, $\mathbf{z}, \mathbf{w}$ represent the set of *N* topics and *N* words, respectively.

The result of data processed by *LDA* is shown in Figure1. For every text, it has 10 topics. We set the topic which has the largest probability as the main topic of this text. This study draws the histogram for all principal topics. As shown in fig.1, more than 12k samples have its main topic's probability significant than 0.9, which means that the *LDA* has excellent interpretability for classification rather than other powerful neural network tools like doc2vec [18]. After data processing, 25 variables are turned to 109 features, and all samples are represented by a matrix, which can use for classification.

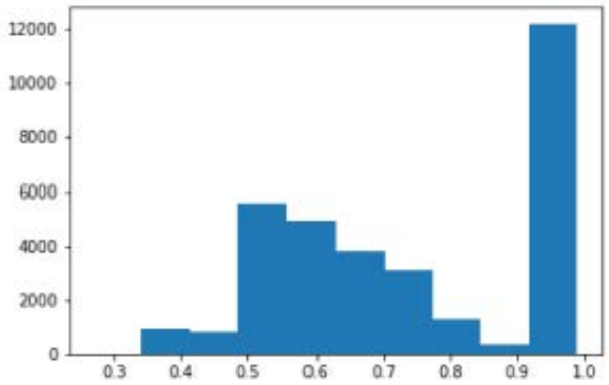

Figure 1. The result of *NLP*

3. Algorithms

This study uses four standard algorithms: *logistic regression, Random Forest, Neural Networks,* and *Naive Bayes,* which are widely used in classification problems with continuous improvement of precision recently.

3.1 Logistic regression

Logistic regression is the preferred binary classification method, which uses for analyzing a dataset that one or more independent variables may influence the outcome [9]. The output is a discrete binary result between 0 and 1, and the purpose is to model the conditional probability $p(\mathrm{Y}=1|\mathrm{X}=x)$ and$p(\mathrm{Y}=0|\mathrm{X}=x)$. It based on the linear regression model, using the sigmoid function to compress the result of the linear model $w^T x$ to [0, 1].

As we all know, linear regression can be described as the following equation:

$$\mathrm{w}\cdot\mathrm{x}=\mathrm{w}_0+\mathrm{w}_1\mathrm{x}_1+\mathrm{w}_2\mathrm{x}_2+\cdots\mathrm{w}_\mathrm{n}\mathrm{x}_\mathrm{n} \tag{3}$$

Formally, the logistic regression model is as follows:

$$\mathrm{p}(\mathrm{Y}=1|\mathrm{X}=\mathrm{x})=\frac{\exp(\mathrm{w}\cdot\mathrm{x}+\mathrm{b})}{1+\exp(\mathrm{w}\cdot\mathrm{x}+\mathrm{b})} \tag{4}$$

$$\mathrm{p}(\mathrm{Y}=0|\mathrm{X}=\mathrm{x})=\frac{1}{1+\exp(\mathrm{w}\cdot\mathrm{x}+\mathrm{b})} \tag{5}$$

And the sigmoid function is:

$$\mathcal{Y}=\frac{1}{1+\exp(-\mathrm{x})} \tag{6}$$

From these above, the output of *Logistic regression* is a discrete binary result between 0 and 1, which indicates the probability that this sample belongs to a certain category [10].

3.2 Random Forest

Random Forest creates the forest with several decision trees. It makes the decision tree more robust and more precise, which has a supervised performance on classification and regression [11], and is widely used to deal with economic problems. Unlike linear models, *Random Forest* can deal with non-linear data as well. The formally models can be described as the following equation:

$$\mathcal{G}(\mathrm{x})=\mathrm{f}_0(\mathrm{x})+\mathrm{f}_1(\mathrm{x})+\cdots+\mathrm{f}_\mathrm{n}(\mathrm{x}) \tag{7}$$

Because of the using of multiple trees, compared to the decision tree, this algorithm reduces the probability of stumbling, which makes the prediction more credible. Besides, by creating multiple estimators, the influence of overfitting is reduced.

3.3 Multi-layer perceptron

Multi-layer perceptron (MLP) [12] has been wildly used among data mining and deep learning domains. Hinton et al. are the first researchers who successfully train the *MLP* model based on back-propagation [13]. Based on several powerful activation functions like the *Rectified Linear Unit* (ReLU) [14] and sigmoid function, *MLP* can approximate any classifier in high dimensional space. If we set inputs as X, the output of each neuron in *MLP* could be described as the following equation:

$$\mathrm{f}(\mathrm{X})=\sum\mathrm{wx}+\mathrm{b}. \tag{8}$$

Where W refers to the weight parameter in the neuron unit, and b is the bias. As the output of each neuron will be activated by the activation function and will be treated as inputs of the next layer.

$$\mathrm{X}'=\max(\mathrm{f}(\mathrm{X}),0). \tag{9}$$

In this research, the last layer will have two outputs which refer to the successful trade or error trade (include bad debt and failed) respectively. For any input, if the first output has a higher value than the second output, it means that this customer may have a higher possibility to finish the transaction punctually and vice versa.

3.4 Naive Bayes

Naive Bayes classifier [15] is one of the most famous classification algorithms for classification tasks. The core equation is based on the Bayes theorem.

$$p(c_i|x) = p(x|c_i)p(c_i)/p(x) \quad (10)$$

Naive Bayes will calculate the probability of the specific category based on the prior information from data. Each category has its own posterior probability, and the largest one shows the possible candidate of its class. $p(c_i)$ Is easy to understand. It shows the ratio of each category. In our research, we have two categories which are the successful trade and error trade, so $p(success)$ refers to the ratio of successful trades in all the transaction data, and $p(error)$ is the remaining part.

As for discrete information, likelihood probability $p(x|c)$ could be calculated as the frequency of the appearance among data. For example, if feature A has appearance frequency n% for class c_1, $p(A|\ c_1)$ = n%.

When it comes to continuous data, we need to use the normal distribution to model the data. For continuous data, it is hard to get such appearance frequency like discrete data. So, we need to calculate the mean and variance of one specific feature dimension, then use the pd.f of Gaussian distribution to get the probability of$p(x|c)$:

$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-(x-\mu)^2}{2\sigma^2}) \quad (11)$$

After understanding the possibility of each feature, just multiply them together, and we can know the posterior probability based on the multiplication law of probability. By comparing the posterior probability of each category, Naive Bayes can realize the classification task by selecting the class of the largest posterior probability.

4. Experiment

4.1 The confusion matrix

In this part, we first discuss the confusion matrix for a two-class problem, which is illustrated in Table 3. This matrix is typically used to evaluate the performance of artificial intelligence algorithms. Traditionally, the minority class is labeled as “negative (0)”, while the majority class is labeled as “positive (1)”. TP and FP represent samples that are correctly classified, FN and TN denote examples that are incorrectly classified [16].

Table 3. The Confusion Matrix for a Two-Class Problem

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

As discussed in section 2, the number of “Closed” samples is much larger than the number of “Bad Debt” and “Failed” samples, so we label “Closed” samples as “positive,” and label “Bad Debt” and “Failed” samples as “negative.” Because for a P2P lending platform, default risk is the crucial factors that need to be correctly estimated, this study focuses on the accuracy of %TN/FN + FP, mark as R (f). The reason why we focus on this rate is that only correctly predicted default loans can play a role in risk management. There are 237 samples of default loans, we use 150 default loans and all positive examples as training samples, leaving only 87 default loans to be tested.

4.2 Parameter setting

Random Forest: We reset the number of trees as 100 for training based on our test that 100 can achieve better results.

Multi-layer perceptron (MLP): In this study, we set a hidden layer size as 10. As for the activation function, we use the ReLU function. After receiving values for outputs, SoftMax function will be used to normalize results. For training such neural network architecture, we use the Adam optimizer [17]. The learning rate was set as 0.001 at the beginning, and it will be judged when optimizing. The batch size was set as 200.

4.3 Results

This paper uses four methods to predict. Each method was repeated for 10 times. And the sample sequence was disrupted for each time to ensure that a different training dataset and testing dataset were used for every experiment. The results of the tests are illustrated in Table 4. As shown in Table 4, the prediction accuracy of each algorithm is quite good, in which Naive Bayes achieves an accuracy to 97%. It means that this study can correctly predict whether a borrower will default the loan or not based on the borrower's information submitted to the P2P lending platform.

Table 4. Results

	Logistic regression	Random Forest	Multi-layer perceptron	Naive Bayes
R(f)	0.864±0.026	0.951±0.024	0.8674±0.145	0.9674±0.007

5. Conclusion

In this study, we work on the default risk of a particular borrower of P2P lending based on data from one platform. To predict the default risk, logistic regression, Random Forest, Neural Networks, and Naive Bayes are used in this study.

There are no research works on the forecast of the default risk of a certain borrower, as far as we know. One of the innovative points is that we developed a risk warning system based on predicting the risk of every borrower by using logistic regression, Random Forest, Neural Networks, and Naive Bayes. It means that if a borrow is judged as negative by our method, then the probability of default is above 86%. Another innovation is the use of Natural Language Processing to transform textual information, which has not been used in previous P2P lending studies. Our experiments show that this step is vital for evaluating the risk of the loans.

The weakness of this study is that although we used four classification methods, we did not analyze the differences between them. We don't know why Naive Bayes is doing so well on this dataset, which requires further researches.

References

[1] Fethi M D, Pasiouras F. Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey [J]. European journal of operational research, 2010, 204 (2): 189 - 198.

[2] Bachmann A, Becker A, Buerckner D, et al. Online peer-to-peer lending-a literature review [J]. Journal of Internet Banking and Commerce, 2011, 16 (2): 1.

[3] Milne A, Parboteeah P. The business models and economics of peer-to-peer lending [J]. 2016.

[4] Lin M, Prabhala N R, and Viswanathan S. 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.

[5] Agarwal S, Li Y, Liu C, et al. Personal guarantee and peer-to-peer lending: Evidence from china [J]. Downloaded on November, 2015, 11: 2015.

[6] Guo Y, Zhou W, Luo C, et al. Instance-based credit risk assessment for investment decisions in P2P lending [J]. European Journal of Operational Research, 2016, 249 (2): 417 - 426.

[7] Beck J E, Woolf B P. High-level student modeling with machine learning[C]//International Conference on Intelligent Tutoring Systems. Springer, Berlin, Heidelberg, 2000: 584 - 593.

[8] Blei D M, Ng a Y, Jordan M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3 (Jan): 993 - 1022.

[9] Hosmer D W, Lemeshow S, Sturdivant R X. Introduction to the logistic regression model [J]. Applied logistic regression, 2000, 2: 1-30.

[10] Hosmer D W, Lemesbow S. Goodness of fit tests for the multiple logistic regression model [J]. Communications in statistics-Theory and Methods, 1980, 9 (10): 1043 - 1069.

[11] Liaw A, Wiener M. Classification and regression by randomForest [J]. R news, 2002, 2(3): 18-22.

[12] Gardner M W, Dorling S R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences [J]. Atmospheric environment, 1998, 32 (14-15): 2627 - 2636.

[13] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors [J]. Cognitive modeling, 1988, 5 (3): 1.

[14] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines [C]//Proceedings of the 27th international conference on machine learning (ICML-10). 2010: 807-814.

[15] Rish I. An empirical study of the naive Bayes classifier[C]//IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001, 3 (22): 41 - 46.

[16] Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]//International conference on intelligent computing. Springer, Berlin, Heidelberg, 2005: 878 - 887.

[17] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv:1412.6980, 2014.

[18] Lau J H, Baldwin T. An empirical evaluation of doc2vec with practical insights into document embedding generation [J]. arXiv preprint arXiv:1607.05368, 2016.