ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 91

1st International Symposium on Economic Development and Management Innovation (EDMI 2019)

Probabilistic Regression and Missing Values Compensation for P2P Lending

Huixi He [a] and Kazumitsu Nawata [b]

Graduate School of Engineering, The University of Tokyo, Japan

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

Abstract. With the development of the Internet, peer-to-peer lending has become widely recognized and accepted by the public. P2P lending platforms are overgrowing around the world, which provides many conveniences to medium borrowers who have difficulty in traditional financial institutions. However, due to the qualification of P2P lending platforms and the lack of supervision, many of these platforms have no ability to disclose information to the authorities, which will increase industry risk and instability in the financial system. So, our work focuses on the missing data of P2P lending platforms from the perspective of predicting losing values. We highlight our work on dealing with missing values using Random Forest instead of traditional methods. After that, we applied the Gaussian Process Regression to explore the relationship between variables and got a conclusion that the loan balance of a particular platform can be predicted.

Keywords: Random Forest, Gaussian Process, Data Compensation, P2P lending.

1. Introduction

P2P platforms offer services online entirely, so they can provide a convenient way for borrowers with a lower service fee, which may attract many medium investors. Thus, borrowers who are rejected by traditional financial institutions may turn to P2P platforms for financing, which makes the P2P lending industry developed rapidly [1][2]. The P2P lending platform was first founded in 2005, which was a simple information intermediary website at that time [3]. Since then, this industry has experienced explosive growth because both borrowers and lenders who use such platforms can gain benefits from lower service fee and faster information.

However, P2P lending platforms have been considered as a high-risk industry because the return is much higher than traditional financial institutions. Meanwhile, most of them cannot manage risk effectively.[4] Besides, because there is no specific law in China, which can be used to control this industry yet, the P2P platforms will not disclose their data to official organizations actively. Therefore, all the data we obtained about the P2P platforms come from the third-party websites, which statistics data from the websites of P2P platforms themselves.

The industry data currently used in relative researches also based on those third-party websites, but these studies rarely focus on the problem of handle incomplete data collected by these websites. Because the P2P lending is a new form of industry, the dataset that used for study is relatively fewer, so it is meaningful to manage to complete missing data. Generally, methods for imputing data can be concluded as two significant categories: assigning data according to the average value and the similar value. In this study, we use Random Forest as a new way to approach this problem in this filed. After that, we use Gaussian Process Regression to verify the relationship between variables.

The rest of this study is arranged as follows: Section 2 is a brief introduction of data. Section 3 mainly describes the data process and experiment. Gaussian Process Regression is introduced in Section 4 using complementary data, and Section 5 is conclusion and discussion.

2. Data

In this study, we synthesized data from one of the largest third-party websites in China, wangdaitianyan(https://www.p2peye.com), statistics on this website is widely accepted by relative researches now. Because this website analyze data from the official website of platforms themselves once a month, so this study is also based on monthly statistics. For each sample, seven variables reflect the performance of platforms monthly: the number of lenders and borrowers, the amount of

loan balance, the business volume, rates, the amount of overdue loan, and net capital inflow. Although the data of platforms can be accessed from January 2017 to April 2019, each platform of every month has more or less missing data. So, we picked up one representative month as a sample, then selected 135 samples from all 184 platforms that summarized by this third-party organization. The reason those platforms are chosen is that they founded the risk management departments basically, and the number of missing variables is relatively small. Also, the amount of overdue loan has been deleted from variables since the rate of loss data is nearly to 74%, which is too high for a data completion algorithm.

The condition of data loss of processed samples is shown in Table 1. From Table 1, it can be seen that only 74 samples have interest rates values of 135 examples, which means that 45.1% of all platforms have chosen not to publish this data to the public. Generally, there are two reasons for this. One is that the interest rates are only accessed to potential lenders as a way to protect the private information of borrowers, which makes the third-party website unable to obtain. The other is that the competition in industry causes interest rates of some platforms to change frequently, which makes the cost of acquiring data increase for third-party organizations. Interest rates can reflect many characteristics of a specific platform such as working mechanisms, lending and borrowing behaviors, and risk management [5]. So, the interest rate is a crucial factor that must be considered when studying in this industry.

Table 1. Data Loss

Loan Balance	Business Volume	Borrowers	Lenders	Interest Rate	Net Capital Inflow
135	135	135	135	74	135

3. Data Processing

In this chapter, we first focus on the methods to deal with missing data and then describe the experiments about impute missing values of P2P platforms. The processes for handling missing values can be divided into three categories, that is, do nothing about it, delete the samples with missing data, and impute the losing data with predicted value [6]. The first method is an easy one because algorithms such as XGBoost can learn the imputation values based on training loss reduction. The second method obtains a complete data table by moving the samples with losing values. This method is quite useful when the ratio of samples with missing value is relatively low. However, a lot of information hidden in those samples may be discarded, which may influence the analysis of data and lead to erroneous conclusions.

The last one mainly uses various ways to impute the missing values. One typic method deal with numeric data is called average imputation, which uses the mean value of the other samples to replace the missing data. In addition, K-means clustering, Regression, and Expectation Maximization are also used to fill the missing values. In this study, we use Random Forest rather than traditional methods to deal with this problem.

3.1 Random Forest.

Breiman.et (2001) developed classification trees to Random Forest, which integrate a number of decision trees into the forest and combine them together to predict results [7]. Random Forest improved the prediction accuracy without increasing the amount of computation that is needed. More importantly, this algorithm is not sensitive to multicollinearity, which can be used in the field of economic analysis easily. Random Forest uses the Bootstrap Method to resample examples, which construct data by choosing observations from the original dataset and then returning them to it [8]. In other words, this method selects k samples with replacement randomly from the original samples to generate a new training dataset, in which duplicate examples are possible.

Decision tree uses the construct of a tree to deal with classification problems. From the root node, samples assigned to its child nodes based on testing whether a certain feature is belonging to this node. Then, testing and assigning samples recursively until leaf nodes are reached. By doing this, each leaf node represents a specified category, which can achieve the purpose of classification [9].

Based on the decision tree introduced above, Random Forest is an ensemble method that creates the decision trees in bagging. First, the Bootstrap Method is used to generate some training sets. Then, for each training sample, a decision tree is constructed, which extracts parts of the features from all features randomly. After that, *N* decision trees are constructed in the same way. For each of the decision tree, a result is proved. So, there have *N+1* results. The random forest integrates all the results, and the most voted results are the final outputs. For the regression problem, the average of the predicted values of *N+1* decision trees determines the final result.

Because of the idea of integrating, that is, sample both examples and features, Random Forest can avoid overfitting effectively. It makes the decision tree more robust and more precise, which has a supervised performance on classification and regression [10].

A common model of Random Forest can be described as follow:

$$G(x) = f_0(x) + f_1(x) + \cdots + f_n(x) \tag{1}$$

3.2 Estimate Missing Values.

As described above, there are many to handle missing data. We use Random Forest in this study to estimate missing data for the reasons as follows: First of all, because of the large difference between platforms, interest rates are also varying greatly from platform to platform. That is, impute the losing data using average value or middle point of known data may not estimate value accurately. Random Forest, however, can learn the characteristics of yet known samples and predict the unknown samples. It refers to a method of training, classifying, and predicting samples by using multiple decision trees. It can also give the score of each variable while classifying samples, which can evaluate the role each variable played in classification. Secondly, Random Forest can deal with both discrete and continuous data without standardization, which is a better choice for the data we used. What's more, it can process data with high features and does not need to make feature selection as well.

As the Random Forest has many advantages described above, we select this method to predict missing value. First of all, we need to test the accuracy of this method. From Table 1, there are 74 perfect samples with values. So, we use 60 samples for training, and 14 samples for testing, the effectiveness of Random Forest can be seen clearly in this way, which can provide guidance when handling with the missing value. The input of each sample is loan balance, business volume, the number of borrowers, the number of lenders and net capital inflow, while the output is the interest rate of this sample.

There are 14 samples which interest rates are not missing, so we can use the difference between precited value and actual value to measure the accuracy of losing data completion. The differences between predicted values and actual values of 14 tested samples are shown in Table 2.

Table 2. Differences Between Predicted Values and True Values

No.	1	2	3	4	5	6	7
Prediction	0.009	-0.011	-0.021	0.019	-0.052	-0.015	0.008
Error	10.9%	10.0%	18.6%	29.5%	36.9%	14.3%	10.9%
No.	8	9	10	11	12	13	14
Prediction	-0.014	0.014	-0.032	-0.006	0.0001	0.007	-0.022
Error	13.1%	19.0%	26.9%	6.56%	0.16%	8.28%	20.7%

As shown in Table 2, the differences between predicted data and actual data are relatively small. The minimum error is at least 0.0001, and the maximum is 0.052. The ratio of error between predict data and real data are also shown in Table 2. The actual interest rates of 14 samples are described in Table 3, the mean value of the interest rate is 0.098, and the maximum interest rate is 0.14. From Table 2, the top three most significant errors are sample5, sample4, and sample10, reached to 36.9%, 29.5%, and 26.9% respectively. However, the real interest rate of the sample5 is the maximum value 0.14, the sample4 is the minimum value of 0.669, and the sample10 is the second-highest value. It

means that the Random Forest works well on standard data, while there are some errors when dealing with outliers.

Table 3. True Values of Interest Rates

Count	Mean	Std	Min	25%	50%	75%	Max
14	0.098	0.02	0.669	0.082	0.099	0.109	0.140

According to the experiment above, except for the three outliers above, the average value of errors is 12.0%, which means that we use this method to handle the missing values of P2P platforms. After processing data in the same way, we obtained 61 predicted interest rates, which filled all the lost data with predicted values. The description of all information is provided in Table 4. As shown in Table 4, the average interest rate is 0.09, and vary from 0.06 to 0.14, which means that the strategy each platform has led to a difference in interest rate. So, the P2P lending industry can provide users with more choices and attract more borrowers than the traditional financial institution. Another noteworthy point is the extreme imbalance between platforms. Loan balance, for instance, the minimum value of this variable is 796, while the maximum value reached to 10,183,954, which indicates that the P2P lending industry in China tends to become an oligopolistic market in the future.

Table 4. Description of Complete Data

	Loan balance(10k)	Business volume(10k)	Borrowers	Lenders	Interest rate	Net capital inflow(10k)
Count	135	135	135	135	135	135
Mean	460,795	50,663	51,321	286,861	0.09	-9,898
Std	1,157,378	112,394	114,092	682,024	0.01	40,787
Min	796	4	11	2	0.06	-439,185
25%	21,924	2,117	1,419	799	0.09	-7,776
50%	65,339	9,064	4,743	11,476	0.09	-621
75%	363,368	48,268	34,244	141,223	0.10	-17.20
Max	10,183,954	744,609	651,336	4,417,561	0.14	12,504

4. Gaussian Process Regression

In the previous chapter, we analyzed several conventional methods that used to deal with missing data and applied the Random Forest to fill the missing values with predicted data as well. In this chapter, we will explore the relationship between the loan balance and other variables. The loan balance is a critical indicator to measure the current operating status of platforms and even can be used to predict future states. In this study, we highlight our work on the application of Gaussian Process for regression analysis instead of traditional methods such as linear regression. The reason for choosing this method is that we want to get the range of possible values of the loan balance.

The significant assumption of Gaussian Process Regression is that given the values of some independent variables X, the dependent variables Y are assumed to obey the joint normal distribution [11]. In other words, for (x_1, y_1), (x_2, y_2), (x_3, y_3) … y_1, y_2, y_3…are assumed to obey the distribution described as follow:

$$Y \sim \mathcal{N}(\mu, \sigma^2) \tag{2}$$

So, Gaussian Process Regression focuses on making inferences about the relationship between X and Y without calculating the coefficients. That is, $p(y|x)$ is estimated based on Bayesian Linear Regression [12].

The units between variables are quite different, which will lead to lower prediction accuracy. So, we first used regularization to deal with data, which can reduce overfitting as well [13]. As there are 135 samples, we use 110 samples for training and 25 samples for testing. The input of each sample

is, business volume, the number of borrowers, the number of lenders, interest rate and net capital inflow, while the output is the loan balance of this sample.

The results are shown in Table 5, and the Gaussian Process Regression works well on this dataset. The average prediction error is 8%, except for one outlier that reached to 66%. Each of the prediction results of Gaussian Process Regression is a distribution that obeys a normal distribution $\mathcal{N}(prediction\ mean, prediction\ std)$, then combine these distributions to form a Gaussian Process.

Table 5. Results of Gaussian Process Regression

Results	Mean	Std	Min	25%	50%	75%	Max
Prediction Error	0.08	0.13	0.00	0.03	0.04	0.06	0.66
Prediction Mean	-0.38	0.06	-0.46	-0.41	-0.38	-0.37	-0.14
Prediction Std	0.10	0.22	0.04	0.04	0.04	0.05	1.15

5. Conclusion and Future Work

We first introduced the problem of missing interest rates in P2P lending industry and analyzed reasons for this as well. Then, we proposed a method to handle missing values based on the machine learning algorithm. Compared with other traditional methods, Random Forest performs well because of the large gap between samples. After filling in the missing data with predict values, we applied the Gaussian Process Regression to analyze the relationship between variables, and this method works well overall expect for one outlier. One of the drawbacks of this study is the selection of dataset. We only selected one month as a sample, which may not prevent randomness. Further research, based on time series data, should be considered in the future.

References

[1]. 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.

[2]. Greiner M E, Wang H. The role of social capital in people-to-people lending marketplaces[J]. ICIS 2009 proceedings, 2009: 29.

[3]. Inman, D.J. 1998. “Smart Structures Solutions to Vibration Problems,” in International Conference on Noise and Vibration Engineering, C. W. Jefford, K. L. Reinhart, and L. S. Shield, eds. Amsterdam: Elsevier, pp. 79-83.

[4]. Emekter R, Tu Y, Jirasakuldech B, et al. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending[J]. Applied Economics, 2015, 47(1): 54-70.

[5]. Zhao H, Ge Y, Liu Q, et al. P2P lending survey: platforms, recent advances and prospects[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2017, 8(6): 72.

[6]. De Silva H, Perera A S. Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data[C]//2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, 2016: 141-146.

[7]. Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

[8]. Danielsson J, de Haan L, Peng L, et al. Using a bootstrap method to choose the sample fraction in tail index estimation[J]. Journal of Multivariate analysis, 2001, 76(2): 226-248.

[9]. Safavian S R, Landgrebe D. A survey of decision tree classifier methodology[J]. IEEE transactions on systems, man, and cybernetics, 1991, 21(3): 660-674.

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

[11]. Rasmussen C E. Gaussian processes in machine learning[C]//Summer School on Machine Learning. Springer, Berlin, Heidelberg, 2003: 63-71.

[12]. Quiñonero-Candela J, Rasmussen C E. A unifying view of sparse approximate Gaussian process regression[J]. Journal of Machine Learning Research, 2005, 6(Dec): 1939-1959.

[13]. Scholkopf B, Smola A J. Learning with kernels: support vector machines, regularization, optimization, and beyond[M]. MIT press, 2001.