International Journal of Applied Science and Engineering
2019. 16, 2: 149-157

*Special Issue of The 13th International Conference on Advanced Information Technologies (AIT 2019)*

# Credit Risk Assessment Using Regression Model on P2P Lending

Shin-Fu Chen[a], Goutam Charkaborty[b], Li-Hua Li[a*] and Chi-Tien Lin[c]

[a] *Information Management Department, Chaoyang University of Technology Taichung, Taiwan*
[b] *Software and Information Science Department, Iwate Prefectural University, Iwate, Japan*
[c] *Financial Engineering Department, Providence University, Taichung, Taiwan*

**Abstract:** The financial market crashed after Lehman Shock in 2008 which creates the risk averting environment for banks. Under this kind of environment, it has become difficult for new and small businesses to have access to loans. Since then, P2P (Peer to Peer) lending becomes more and more popular around the world. However, most of the researches who had studied about credit risk on P2P lending consider only the event that the borrower will default instead of the amount of loss. In this work, we consider Net Return Rate (NRR) as the criterion to label the data for prediction training. We train the regression model to assess credit risk. The proposed model predicts the amount of profit from a borrower. In our results, by using our proposed credit risk assessment model, an investor of P2P lending can measure the risk with better accuracy and the proposed model can also predict the amount of profit from a loan.

**Keywords:** P2P lending; random forest; logistic regression; credit risk; credit scoring.

## 1. Introduction

Since Lehman shock, the lending rules from big banks became too strict for new lenders, even with innovative ideas. As a result, lending through social networks or other FinTech platforms keeps growing. Platform of P2P (Peer to Peer) lending as a substitute of traditional lending is flourishing around the world. P2P lending has the advantages of being fast, customizable and low cost [1]. However, like the traditional lending business, P2P lending also involves credit risk problems. Moreover, the traditional banking system can use borrowers' historical financial transaction to build credit scoring and behaviour scoring system to manage customers' credit risk. Without these borrowers' information, it will be difficult to perform the analysis of credit risk and it will be unreliable, too. On P2P lending platform, the detail information about the borrower usually is not complete and sometimes not available. As a result, analysis of the available big data is the only source to be applied for measuring the credit risk on P2P lending platform [1].

For P2P credit risk analysis, most of the researches obtained data from LendingClub (LC), which is one of the biggest P2P lending platforms in the U.S.A. LC provides open data with information supplied by the borrower and her/his repayment status [2, 3, 4]. In the work of Milad Malekipirbazari and Vural Aksakalli [2], they proposed a random forest classifier for predicting borrower's loan status and compared results with other classifiers. They showed that Random Forest (RF) model could achieve the best classification accuracy. However, if the target is just to classify whether the borrower will be default or not, it will lead to mis-judgement of credit risk measurement due to different default situations.

Corresponding author; e-mail: lhli@cyut.edu.tw
doi: 10.6703/IJASE.201909_16(2). 149
Received 28 May 2019
Revised 23 September 2019
Accepted 30 September 2019

For example, the borrower who paid 90% of the loan will be regarded as default in the same way as the borrower who paid only 10% of the loan. In reality, their credit risk levels are different.

In traditional credit risk analysis methods, credit scoring systems are usually designed to estimate the borrower's probability of default (PD) [3]. For P2P lender (investor), a reliable algorithm to measure the balance between PD and profit is very important. In the work of Carlos and Begona [5], they used Internal Rate of Return (IRR) as the output label for the data to train a machine learning model and for credit scoring system to provide advice for investors. In their experiment, they trained the decision trees for credit scoring system which had achieved 5.98% IRR on an average, outperforming the margin of market average 3.92% IRR. However, if we use this type of system, investors can only lend money to 17.17% of borrowers. This low percentage means that the model is too conservative and too strict for loan, which reduced the efficacy of P2P lending platform. On the other hand, LendingClub (LC) does not provide historical payment after 2012. Therefore, the IRR cannot be obtained and calculated after 2012. In addition, one of the most reliable credit-rating companies, namely Fair Isaac Corporation (FICO), does not provide the information of IRR any more.

In order to build the credit risk assessment model on P2P lending platform which can provide advice for investors with reasonably and effectively outcome, this research proposes the use of Net Return Rate (NRR) as the output label of the training data. We use regression analysis and machine learning methods to build the credit risk assessment model. In order to understand the performance of regression models, we have conducted experiments using RMSE (Root Mean Squared Error), Mean of NRR, lending ratio, and the cost of computing. Experimental results are compared with various machine learning models. It has shown that using Linear Regression (LR) model and the Random Forest Regression (RFR) model will have a better outcome for P2P lending. As a result, LR model is considered the best choice for investors so that they can successfully identify borrowers and obtain a higher return rate above the market average.

The rest of the paper is as follows. Section 2 is about related works. In section 3, we explain our proposed machine-learning model and the target output. Section 4 describes the experiments and results. Section 5 addresses our conclusions.

## 2. Related works

Credit risk is an important part of financial risk management. Traditional banking systems and financial institutions applied credit scores to measure the credit risk of a loan application.

Lyn C. Thomas [3] did the survey of credit and behavioural scoring. For new applicants, financial institutions can calculate the credit score of applicants, which are based on statistical or operational research methods using basic information provided by applicants. The motivation is to help lenders determine whether to lend or not, to a particular borrower. For the borrowers with a history of repayment, lenders (banks) can use behavioural scores to help them deciding whether to accept an increasing amount of loan or to increase the interest rate on borrowing to reduce credit risk. Recently, machine learning approaches, like logistic regression, decision trees, neural networks, etc. are highly applied for credit risk analysis.

In the work of J. Galindo and P. Tamayo [7], different statistical methods and machine learning methods were used to predicting whether a mortgage loan will default or not. The experimental results pointed out that the CART (Classification and Regression Tree) model can get the lowest error rate in credit risk classification. Their method utilized the mortgage loan information of Mexico. The risk prediction model was trained with 24 variables, including borrowers' repayment history over the past 10 months. After comparing different models, i.e., CART, K-Nearest Neighbour (K-NN), Neural Network and Probit, CART model achieved the lowest error rate at 8.31%.

In contrast to the mortgage loan, P2P platform does not have a repayment history of borrowers, however, it is one of the most significant features in credit risk analysis. In the work of Milad Malekipirbazari and Vural Aksakalli [2], they used the data from the popular social lending platform, i.e., LendingClub (LC), to build the risk assessment model. They proposed and presented a comparison of results from different machine learning models including Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), and K-NN classifiers. Their results show that RF can outperform other classifiers to classify if borrowers' loan statue is default or fully paid. However, in their results, the negative accuracy rate is relatively low and Root Mean Square Error (RMSE) is high with 0.45. It shows that the classification model for P2P lending is biased. In addition, investors need to know not only whether the borrowers will default or not, but also the actual amount of profit by extending the loan. If a borrower defaults after paying a high interest over a long period, then the amount could be more than the principal amount. It is still a good investment.

To consider the profit of lending, Carlos and Begona [5] proposed decision trees. They applied an internal rate of return (IRR) as the output label of the data, to build a Decision Support System (DSS). IRR is calculated by the repayment of the loan in each period, which is closer to the real profit of the lending. In their experiment, they suggested the following three rules: (1) only invest when the annual income greater than $65,000 USD, (2) 1 or 2 inquiries from credit rating agency or bank in the last 6 months, and (3) no investment on small business. On average, this strategy outperformed the market average return by a margin of 3.92%. However, following this strategy, one can only invest in 17.17% of all borrowers in the test set. This implies that their rules are too strict and most of the borrower cannot pass. Since the main idea of P2P lending is to increase the possibility of borrowing and to reduce the borrowing gap of high-risk customers, therefore, a strict model like this is simply not suitable for P2P lending.

## 3. Proposed Method

Recently, many researchers take the credit risk problem of P2P lending as the classification problem. However, in reality, any investor interests more about the profit of lending. To build an efficient credit risk assessment model and to provide practical advice for an investor, this research utilizes the machine learning model which takes the Net Return Rate (NRR) as the output label of data for training. NRR is the outcome that lenders will really interest in. We use the regression model to obtain a real value as the amount of risk.

### 3.1 Net Return Rate (NRR)

In the work of Carlos and Begona [5], their research applies Internal Return Rate (IRR) as the measurement of loan profitability. However, IRR is computed by the initial cash outflow (the loan amount) and the cash inflows (repayment) in each period (like a quarter). Currently, the P2P lending information of LendingClub (LC) provides no historical record of the borrower’s repayment. Only information of the total repayment amount is available. Therefore, we cannot use IRR as a profit label (output of the model). Due to the lack of necessary information, this study suggests that we can use the Net Return Rate (NRR) as the measurement of performance and as the output label of data.

In the LC data, the "*agreement interest rate*" and the "*term*" (duration of the loan) are available for each loan. The expected return of a loan is then calculated as equation (1).

$$\text{expected return} = (1 + agreement\ interest\ rate)^{term} \tag{1}$$

The "*agreement interest rate*" depends on the "*grading*" (like trust) of borrowers. LC specifies the corresponding interest rate and the "term," i.e., the duration of the loan is also specified, either 36 months or 60 months. This research proposed the expected return as NRR by using equation (2).

$$\text{Net Return Rate} = \frac{total\ repayment\ amount}{funded\ amount * expect\ return} \tag{2}$$

The histogram of NRR in 2016 on LC platform is as shown in Figure 1.

### 3.2 Logistic Regression (LR)

Since the credit scoring system is applied to measure the credit risk of a loan, a Logistic Regression (LR) model is considered one of the most suitable methods for credit scoring [3]. Logistic Regression, as a binary classifier, uses the feature variables of each sample as input to calculate the probability of how good the sample is. When using LR model, an output of 1 represents the best grade of investment, i.e., the probability of load repayment is high [8].

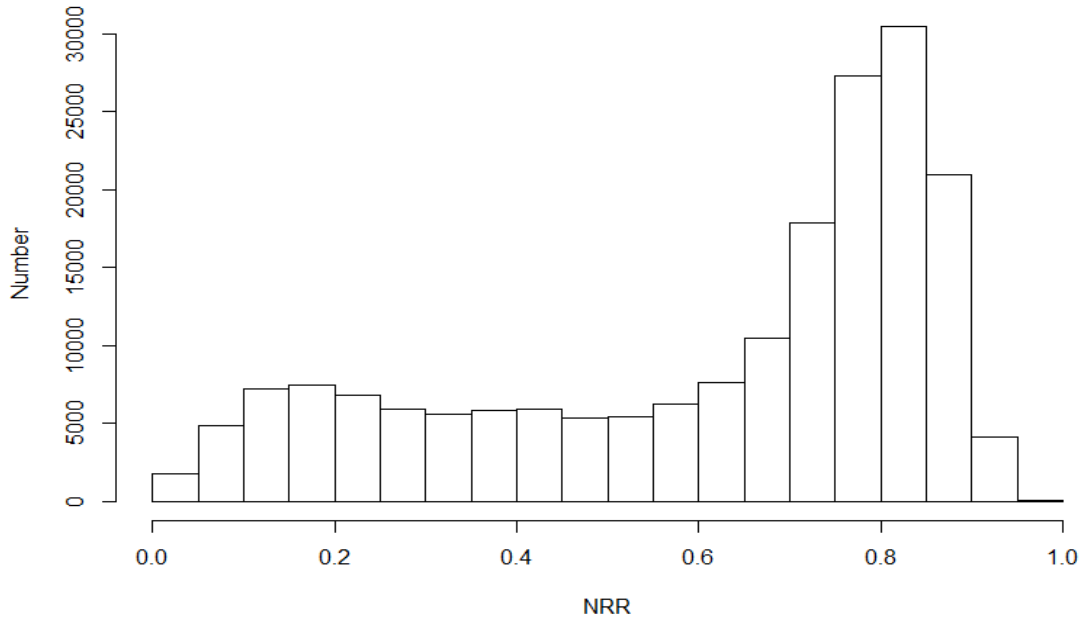


Figure 1. Histogram of NRR in 2016 on LC platform.

Traditionally, Logistic Regression uses the binary label of 0 or 1 to train the model. In this research, we applied a continuous label (NRR) as the data label to train the regression model. The cost function of our proposed model is defined as equation (3).

$$\cos t\_\ function = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i) - y^i)x^i \tag{3}$$

In equation (3), x is the feature vector of a borrower, y is the output which is corresponding to borrower's NRR, and $h_\theta(x)$ is the sigmoid function in our experiment. The sigmoid function is defined as equation (4).

$$h_\theta(x) = \frac{1}{1+e^{-x}} \tag{4}$$

Logistic Regression (LR) can provide not only the expected NRR, but also the significance of each variable to the outcome to help investors understand which information of borrowers is more important [7]. Thus, LR can also be utilized as a tool for feature selection.

### 3.3 Random Forest (RF)

As an ensemble of decision trees, the method of Random Forest (RF) can significantly improve the generalization accuracy for classifier and regression [9, 10].

Random forest regression is generated by integrating a large number of regression trees. To avoid the overfitting problem, each regression tree is trained by a different subsample of training data and subsample variables. The training data is randomly extracted under the assumption of independent and identically distributed data. For regression, the prediction of random forest is the unweighted average over the collection of results of each regression tree [11].

To evaluate the method of Random Forest Regression, the error of prediction based on the test set is estimated by using the loss of squared error for each out-of-bag case. The squared error loss function is defined as equation (5).

$$Squared\ Error\ Loss = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

Here, *yi* is the accurate NRR value and $\hat{y}_i$ is the predicted NRR value generated by using RF model. As the number of tree increases, the square error loss will reduce. When the squared error loss is less than a pre-assigned threshold, then we say the training has converged. Once the training process of RF has completed, the model of RF Regression is ready to use.

## 4. Experiment and results

The dataset we used in the experiment is obtained from Lending Club (LC) website [12] which is one of the biggest P2P lending platforms in the U.S.A. We use the data in year 2016 and the raw data contains 434,407 samples with 143 variables. In the process of data cleaning and data pre-processing, variables with missing data or incomplete information will be deleted. As a result, there are 33 variables are retained and will be used as input features.

We first calculate the average Net Return Rate (NRR) base on the raw data and we derive the NRR is 61.17%. To find the best model for credit risk prediction and for repayment of P2P lending, we have implemented four different models in our experiments. These four models are: (1) Logistic Regression (LR) model, (2) Support Vector Machine Regression (SVR) model, (3) CART decision tree model and (4) Random Forest Regression (RFR) model. To train these models, we separate the dataset using 7:3 ratio, i.e., 70% of data is used as training data and the other 30% of data is used as testing data.

### 4.1 Experiment--using 33 variables as input features

To compare the effectiveness and the prediction performance of these four models (see Table 1), i.e., LR, SVR, CART, and RFR. We take all 33variables as input for training. After these models are all trained, we then use the trained models to predict the test data set. The test set contains 58,834 samples with an average of 61.11% NRR. To evaluate the performance of these models, we calculate RMSE of predicted NRR for each model. Three other measurements are also calculated, i.e., mean of NRR, lending ratio, and cost of computing.

To measure the profit for the investors, we set a threshold for each model to decide whether the lender should invest or not. If the predicted repayment probability is less than 0.5, we set the model to reject the loan requested from the borrower. For all the loans that are classified as acceptable, the average NRR will be calculated to measure model performance. The lending ratio is obtained by calculating the ratio of the number of accepted borrowers versus the total number of all borrowers. This ratio can help us understand how strict our model is set.

All of these four models are built using R language and also the training and testing process. Among these four models, RFR model was built with 200 regression trees and 6 feature subsets. Total of 33 variables and the square root of variables are calculated. The results of various models are as shown in Table 1.

From Table 1 We observe that CART model produces the highest mean value of NRR if compare to other models. However, the lending ration for CART has only 24.2% which is much lower than LR model (76%) and RFR model (75%). Also, the RMSE for CART model is too high which also implies the model has higher error of prediction the NRR result. In turns, this may increase the risk of using this model for P2P lending.

The SVR model with RDF kernel in Table 1 shows excellent lending ratio, i.e., 85%, however, the cost of computing is too high which means this model cannot be practical for on-line service. If we compare SVR to other models, SVR model also has a higher RMSE value. From the experimental results, as shown in Table 1, we observe that the LR model and RFR model have similar performances. However, the cost of computation of RFR is higher than LR model. Therefore, if we take all of the measurements into consideration, this research suggests that Logistic Regression (LR) model should be the best choice to predict the credit risk for P2P lending.

Table 1. Experimental results of four different models.

| Measurement of Performance | LR | SVR(RDF) | CART | RFR |
|---|---|---|---|---|
| **RMSE** | 0.197 | 0.212 | 0.541 | 0.197 |
| **Mean of NRR** | 68.3% | 66.1% | 80.2% | 68.8% |
| **Lending ratio** | 76% | 85% | 24.2% | 75% |
| **Cost of computing (Sec.)** | 1.31 | 14,909.5 | 6.32 | 6262.9 |

### 4.2 Experiment--using significant features

In order to find out which important feature (variable) may affect the NRR, we examine the coefficients of logistic regression model as shown in Table 2, and the difference of Mean Squared Error of Random Forest Regression model in table 3 after training.

Table 2. Significant features using logistic regression model.

| | ***Estimate*** | ***Std. Error*** | ***Pr(>\|z\|)*** |
|---|---|---|---|
| ***(Intercept)*** | 3.108 | 0.032 | < 2E-16 |
| ***Term*** | -1.776 | 0.039 | < 2E-16 |
| ***Interest Rate*** | -9.857 | 0.146 | < 2E-16 |
| ***Employee length*** | 0.085 | 0.016 | 1.43E-07 |
| ***Home Ownership**--Mortgage* | 0.177 | 0.013 | < 2E-16 |
| ***Home Ownership**--Own* | 0.08 | 0.02 | 6.15E-05 |
| ***Verification Status** Not_Verified* | 0.0513 | 0.014 | 3.20E-04 |
| ***Debt to Income*** | -5.355 | 0.671 | 1.48E-15 |
| ***Delinquency in 2 years*** | -0.465 | 0.135 | 5.37E-04 |
| ***Inquest in Last 6 Months*** | -0.207 | 0.034 | 1.17E-09 |
| ***Revolving Utilization*** | -0.247 | 0.042 | 5.11E-09 |

In Table 2, we list out all the significant variables based on the Logistic Regression model. These significant variables all have *P value* less than 0.05. The smallest *P value* means the most significant feature which includes: (1) *Term* (duration of the loan), (2) *Interest Rate*, and (3) *Home Ownership--Mortgage*.

In Table 3, *%IncMSE* is defined as the difference of result for each tree when the feature on the leaf column is included or not. If the value of *%IncMSE is high, it means the pruned leaf is very important which causes the big difference of final value. Therefore,* a high value of *%IncMSE* means the feature is significant. In Table 3, we observe that the *int_rate* (interest rate) has the highest *%IncMSE*, which means this feature affects MSE the most. We choose features with high *%IncMSE* values, i.e., $\%IncMSE \geqq 20$ as the significant features as shown in TABLE 3.

Table 3. Significant features using random forest regression model.

| | ***%IncMSE*** |
|---|---|
| ***int_rate*** | 213.35 |
| ***term*** | 85.25 |
| ***RTI*** | 57.97 |
| ***annual_inc*** | 54.45 |
| ***dti*** | 50.44 |
| ***revol_util*** | 46.81 |
| ***open_acc*** | 34.35 |
| ***installment*** | 27.29 |
| ***funded_amnt_inv*** | 27.28 |

By comparing Table 2 and Table3, we can conclude that *int_rate* (interest rate) and *term* (duration) are the two most important features for our experimental models. To demonstrate the relationships between these two variables and the relationship to NRR, we visualize the relationship as shown in Figure 2.

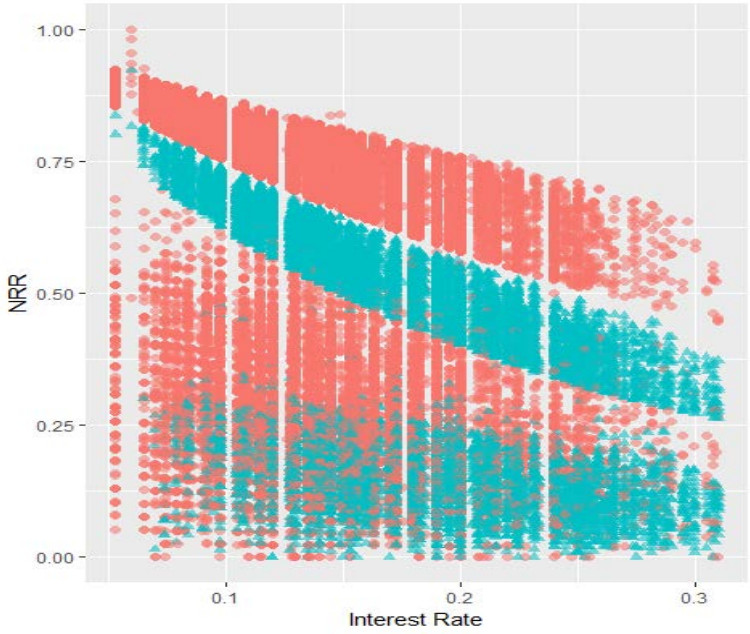


Figure 2. Relationships of *term*, *interest rate*, and *NRR*.

In Figure 2, the orange dots represent the loan of 36-month duration, and green triangles represent the loan of 60-month duration. The *X*-axis in figure 2 is the agreed interest rate and *Y*-axis is the NRR value. It is noticed that, in Figure 2, when the agreed interest rate increases, the proportion of NRR will decrease. It is reasonable because higher the interest rate will make the repayment more difficult which may create more defaults. On the other hand, loan borrowers with a shorter period of time such as 36-month have higher NRR than those with a longer borrowing period of time. It also makes sense that the longer the loan period, the higher the risk that the borrower will default.

## 5. Conclusions

How to find the balance between profit and risk has always been one of the most difficult issues in the financial world. For P2P lending, higher interest rates are often accompanied by higher credit risk.

This research has experimented on four different machine learning models, i.e., LR model, SVR model, CART model, and RFR model. This research has suggested that we should consider the extent of the borrower's overall repayment and the expected return. In this way, we can help investors to measure the profit and risks, therefore, a better lending decision can be made.

Based on our experimental results, it has shown that both Random Forest Regression model and Logistic Regression model can achieve the lowest RMSE if compare with SVR and CART models. In our experiments, we had set the 0.5, as the threshold, of NRR to filter out borrowers. Using this threshold we can achieve the mean of NRR to 68.3% with LR model and 68.8% with RFR model. If we compare with the average NRR in the market of P2P lending, which is 61.11%, then our proposed LR and RFR model gain 7% more than the average market value.

This research suggests that LR model is both effective and practical because LR model has the highest lending ratio (76%) with the lowest cost of computing (1.31) and the lowest RMSE value. By using LR model we can choose the borrower based on the model suggestion and still we can receive better NRR (68.3) than other models. If compare with the NRR value (17.177%) of [5] our LR model shows great increase (68.3%). We like to conclude that the LR model can be an effective credit risk prediction model for P2P lending.

## References

[1] Working Group of CGFS and FSB. 2017. FinTech Credit: Market Structure, Business Models and Financial Stability Implications, Report. *Bank for International Settlements and Financial Stability Board*, https://www.bis.org/publ/cgfs_fsb1.htm, ISBN 978-92-9259-051-2 online.

[2] Malekipirbazari, Milad & Aksakalli, V. 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications*, 40, 10 : 4621-4631.

[3] Thomas, L. C. 2000. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International journal of forecasting*, 16, 2 : 149-172.

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

[5] Carlos Serrano-Cinca & BegoñaGutiérrez-Nieto, 2016, "The use of profit scoring as an a lternative to credit scoring systems in peer-to-peer (P2P) lending," Decision Support Sys tems, vol. 89, pp. 113-122.

[6] Serrano-Cinca, C. and Gutiérrez-Nieto, B. 2016. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, 89 : 113-122

[7] Galindo, J. and Tamayo, P. 2000. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics*, 15, 1-2 : 107-143.

[8] Wiginton, J. 1980. A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitative Analysis*, 15, 3 : 757-770

[9] Breiman, L. 2001. Random forests. *Machine learning*, 45, 1 : 5-32.

[10] Barandiaran, I. 1998. The random subspace method for constructing decision forests. I*EEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 8.

[11] Segal, M. R. 2004. Machine learning benchmarks and random forest regression. *Technical Report, Center for Bioinformatics & Molecular Biostatistics*, University of California, https://escholarship.org/uc/item/35x3v9t4.

[12] 2018. Open Data of Lending Club, Available at: http://www.lendingclub.com/info/down load-data.action, extracted at Sep.