Feature Selection on Credit Risk Prediction for Peer-to-Peer Lending

Shin-Fu Chen[1(✉)], Goutam Chakraborty[1], and Li-Hua Li[2]

[1] Graduate School of Software and Information Science, Iwate Prefecture University, 152-52 Sugo, Takizawa, Iwate Prefecture, Japan
albirtle93@gmail.com
[2] Department of Information Management, Chaoyang University of Technology, Wufeng, Taichung City 41349, Taiwan

Abstract. Lending plays a key role in economy from early civilization. One of the most important issue in lending business is to measure the risk that the borrower will default or delay in loan payment. This is called credit risk. After Lehman shock in 2008–2009, big banks increased verification for lending operation to reduce risk. As borrowing from established financial institutions is getting harder, social lending also called Peer-to-Peer (P2P) lending, is becoming the popular trend. Because the client information at P2P lending is not sufficient as in traditional financial system, big data and machine learning become the default methods for analyzing credit risk. However, cost of computation and the problem of training the classifier with imbalance data affect the quality of result. This paper proposes a machine learning model with feature selection to measure credit risk of individual borrower on P2P lending. Based on our experimental results, we showed that the credit risk prediction for P2P lending can be improved using Logistic Regression in addition to proper feature selection.

Keywords: P2P lending · Credit risk · Minimum Redundancy Maximum Relevance (mRMR) · Least Absolute Shrinkage and Selection Operator (LASSO) · Logistic Regression

1 Introduction

Peer-to-Peer (P2P) lending platform is emerging as an alternative to banking system. P2P allows individual members to lend and borrow money directly without official financial institution such as banks, playing as intermediate. Since the Lehman shock in 2008–2009, customer trust in financial services declined rapidly. Regulators mandated increased safety measures to approve loans which resulted in banks tightening loan requirements. Financial institutions become more risk averse, causing a loan gap. The needs of risk seeking lenders and high-risk borrowers are not fully served by traditional financial institutions [1]. P2P lending platforms, where a lender has more flexibility to pick and choose a desired risk portfolio, is becoming more and more popular. The difference between traditional lending system and P2P lending platforms is shown in Fig. 1, where different risk taking lenders will find corresponding borrowers.

K. Kojima et al. (Eds.): JSAI-isAI 2018 Workshops, LNAI 11717, pp. 5–18, 2019.
https://doi.org/10.1007/978-3-030-31605-1_1

Traditional banks now work where risk is low. For new small or venture business, it is difficult to obtain loan. P2P lending works over the whole range of risk. In addition, low interest rate or risk based interest rate is an attractive option for P2P lending. Because of open playing ground, it is beneficial to both lender and borrower.

In general, financial institutions analyze customers' credit risk using linear discriminant analysis to build score card system. A reliable model requires a large number of customers' information for statistical analysis [2]. Recently, Machine Learning model has been applied in credit risk area, and shown to have a good accuracy at default prediction [1–3]. Data regarding risk related features, and a balance data, in which the number of timely paid borrowers are nearly the same as the number of default borrowers, would be able to properly train the classifier.

However, in practice, the training data contain only a few borrowers who are default, where most of the borrowers' payback in time. Training with such data, the classifier will be biased to predict all borrowers to be classified as good. This will give high overall accuracy with test data. High accuracy of prediction model doesn't mean the model is good. It predicts correctly of non-defaulting borrowers which predominate the data, wherein fails for cases where the borrower defaulted. This is due to imbalance of available data [4, 5]. Under such circumstances, the model often converges to an overtrained model biased to the predominant class of the available data. How to achieve accurate prediction for bad borrowers is crucial. In addition, compared with traditional banking system, P2P lending do not have sufficient information about customer's financial statistic and historical data. Moreover, the model is needed to be computationally light. How to find important features to reduce the cost of computing becomes another important issue. Fewer features improve the accuracy of classification and generalization, if selected properly.

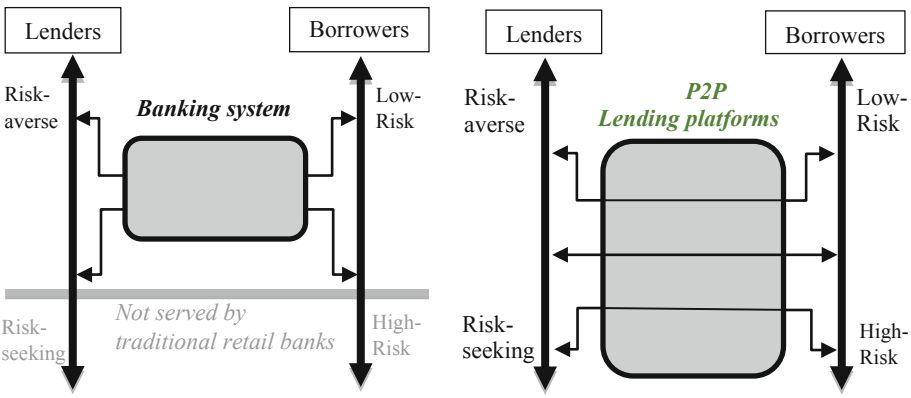

Fig. 1. The difference between traditional lending system and P2P lending platforms

In order to improve identification of credit risk on P2P lending, this research proposed machine learning model to do classification. To improve classification accuracy of both classes, we applied undersampling to deal with the problem of imbalance data. We used Minimum Redundancy Maximum Relevance (mRMR) for

feature selection. This research proposed and presented comparisons of Logistic Regression (LR) and Random Forest (RF) approaches for classification. The experimental results show that LR could achieve similar performance to RF, with less computational cost.

2 Related Works

2.1 P2P Lending

The cause of Lehman shock was that the financial system took too many subprime mortgage debt, which led to the liquidity risk of the financial system [6]. Since then, the financial supervision has been strengthened. The capital adequacy requirement rules for the banking system made the bank's review of the loan stricter, to avoid predatory lending. Due to above reasons, and the development of social networking platforms, P2P Lending is popularized very rapidly [7]. It has the potential to increase economic activities and efficiency of transaction. It can replace financial institutions as a lending medium with appropriate interest rate. For example, LendingClub, one of the most popular P2P lending platform in the U.S, is enjoying a great growth at both number of loans and the total loan issuance, as shown in Fig. 2.

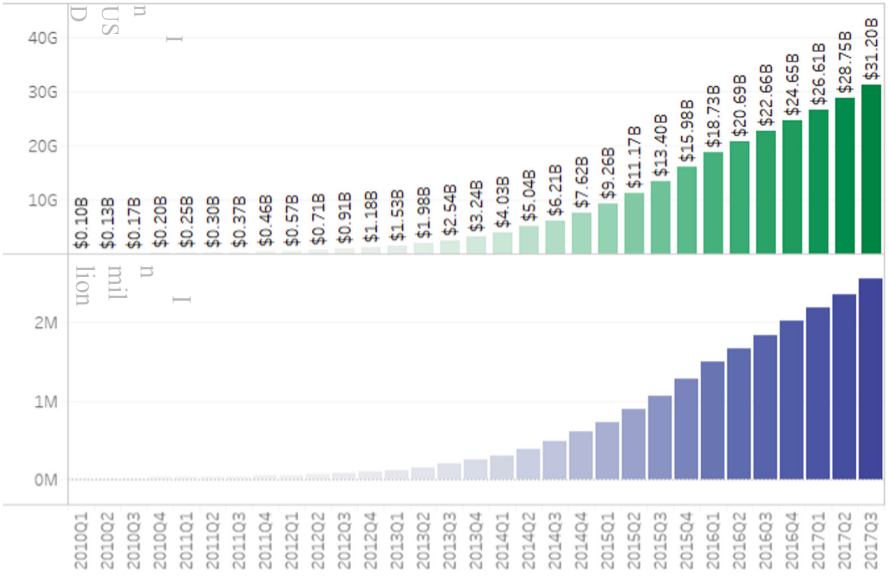

Fig. 2. The growth of LendingClub

There are some major benefits of P2P lending. Firstly, P2P lending platforms have the potential to offer lower interest rates to borrowers. Because of open market place, the interest rate would evolve to its appropriate risk value. Higher returns to investors

are possible because of lower costs compared to banks through extensive use of computerization and the absence of physical store. Secondly, P2P platforms can provide a more convenient service for customers because of a transparent computerized environment for providing loan information and assessing loan risk, which can reduce the search costs of lenders. In addition, P2P lending platform can accelerate the process of lending, because there is no due diligence to borrowers. The online platform can provide faster and more convenient matchmaking mechanism [7]. However, there are a few added risk of P2P lending platform. One of the most important part is credit risk assessment quality, because platforms do not likely have the detailed information such as historical loan information of the borrower, and liabilities information. Therefore, it is not easy to assess the borrower's credibility or the amount of possible losses, making general investors less willing to bear the credit risks. This make some good borrowers with excellent plan, but with poor credit ratings, unable to get much needed fund. This study hopes to establish a better prediction model to help investors make proper risk assessments on lending, thereby increasing investor confidence in P2P platform.

2.2 Machine Learning Method

In recent years, financial institutions have begun to use machine learning in credit risk analysis. In past, researches used various machine learning models to predict the credit risk of traditional financial institutions, including Logistic Regression, Support Vector Machine (SVM), and Random Forest [2, 3, 8, 9]. However, compared to P2P lending, traditional financial institutions have more financial information about customers. They can also afford longer time for decision making. When customers apply for a loan, financial institutions can analyze the information about the customer's financial history to help them to make more appropriate decision. Relatively, P2P lending does not have financial history about borrowers, so the analysis of credit risk has been more dependent on the big data of peer borrowers.

One of the closet research to this work is by Malekipirbazari and Aksakalli [10]. They used the data set from LendingClub during the time period between 2012 and 2014, and carefully explained every feature. In their experiment, they proposed and presented comparisons of different machine learning models, and showed that the Random Forest have the best performance. To improve the classification accuracy of Random Forest model, they used a cost matrix technique that allows the model to increase costs when misclassifying bad (default) customers to good customers. However, the experiment did not use the feature selection method to retain only important features. One of the feature, namely the external credit score indicator also called FICO scores, could not be obtained in recent years' data set. Using recent data to do credit risk analysis, applying feature selection to reduce feature dimensions, and improving predictions for high-risk lending, are the focus of this research. One of the main emphasis is to remove irrelevant features. Not only they act as noise and reduce classification accuracy, they make the classifier unnecessarily complex and difficult to train with big data.

3 Methodology

Because the original data set is unbalanced, and some of the features are superfluous, we need to deal with these two problems before we train our classifier. To improve the performance of prediction, in this research we applied feature selection to select important feature in credit risk prediction. For feature selection, we tested Minimum Redundancy Maximum Relevance (mRMR) and least absolute shrinkage and selection operator (LASSO). We applied undersampling to deal with data imbalance problem. The steps of the proposed model is shown in Fig. 3.

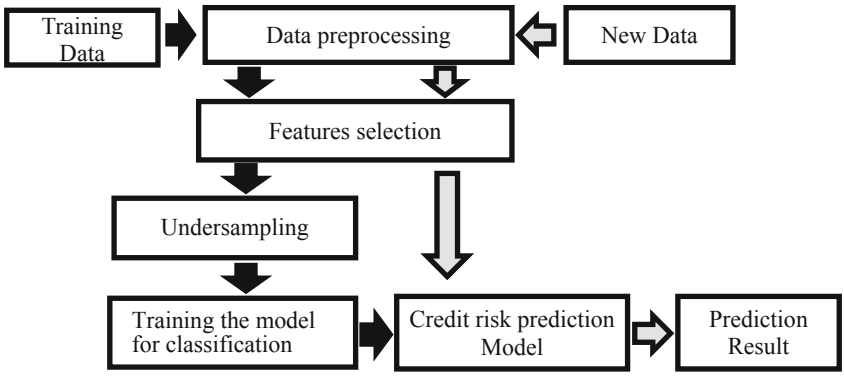

Fig. 3. Structure of the proposed model

3.1 Data Collection and Pre-processing

The original data set is collected from the LendingClub's website [11], one of the most popular P2P platform in the U.S. The period of raw data is from first quarter of 2016 to fourth quarter of 2016, contained 434,407 borrowers with 145 features. For clearing of the raw data, we follow the process: (1) feature irrelevant to the risk assessment, like location or email addresses etc., are manually deleted. (2) Features which appear for only a few numbers are deleted. By the above procedure, the number of feature were reduced to 18. Other than this, some of the borrowers have not finished their loan period yet, which is not suitable for our experiment. Therefore, we retained only the borrowers who has clear loan status, which are default, charge off, or full-pay. There are 187,192 borrowers in this dataset we could use to build our classification model. There are still some borrowers for whom does not has some important features like debt to income ratio (DTI) or revolving utilization. In this experiment, we need such information. Therefore, we omit those borrowers' data. After filtering, the data set contains 117,790 borrowers. The distribution of the loan statue was: 77% of borrowers did full-payment, 23% of borrowers were default or charge off. For this experiment, we used 80% of borrowers' data as training set to build our classification model, and used the rest 20% for model testing.

After we collected data from the website, we preprocess the data to make it suitable for the algorithm. Features are of three data types: binary, real numeric and categorized data. Binary data are used as it is, 0 or 1. Depending on the value and range, some numeric data are normalized to range 0 to 1, and for others the range in scaled down using log of the original numeric value of the feature. We did data encoding to convert categorical type variables into numeric types. For example, the feature "home ownership" has four different categories, "Rent", "Own", "Any", and "Mortgage". In this research we used binarization to generate new features, which are Home ownership (Rent), Home ownership (Own), Home ownership (Any), and Home ownership (Mortgage) to corresponding four feature descriptions. Value of these features are binary, could be 0 or 1.

3.2 Feature Selection

Feature selection is an important problem for machine learning, when there are a lot of redundant features in the dataset, it would be computationally costly to train machine learning model. The other important purpose is that feature selection can help observer to identify which features are important, out of the many possible ones. The general approach is to start with all possible features, and then remove irrelevant ones. In addition, irrelevant features act like noise and deteriorate classification results. Thus, how to reduce the number of features to minimum retaining high classification accuracy, is an important issue.

There are two main approaches, namely dimensionality reduction and feature selection. Dimensionality reduction maps the data onto a lower dimension feature space from original feature space. One of the popular method is principal component analysis (PCA). The problem with dimensionality reduction is that, features of the mapped lower dimension space do not have any physical meaning. Users will not have any idea of which real world features are important. In financial decision, it is important to retain meaningful features.

The other approach is feature selection. Feature selection method is classified into two approaches: Filter method and Wrapper method. In filter method, an individual feature is evaluated using some statistical methods like Chi squared test, information gain or correlation coefficient score. Features are selected according to their scores. In wrapper method a model is used, and a subset of feature is evaluated using the model. The model could be anything, like a regression model, K-nearest neighbor, or a neural network. Searching of optimum subset of features, could be heuristic, stochastic or forward-backward to add and remove features.

Because we want to know which feature from the original set are important, we used feature selection. We used a wrapper method with regression as model, namely absolute shrinkage and selection operator (LASSO). We also tested another statistical method namely Minimum Redundancy Maximum Relevance (mRMR).

The idea of mRMR is that in case of high-dimensional feature space, it is difficult to find which features has the largest dependency on the target class. Selecting features based on maximal relevance criterion is an option. Maximal relevance is to search features which have the approximate maximum dependency to target classes. However, maximal relevance search may select features which are redundant. When two features

highly depend on each other, removing one of them would not change the class discriminative power. Therefore, minimizing redundancy could be used to select one of the mutually exclusive features [12]. The optimization criteria are as shown below:

$$\max D, D = \frac{1}{|S|}\sum\nolimits_{i\in s} I(X_i, C) \tag{1}$$

$$\min R, R = \frac{1}{|S|^2}\sum\nolimits_{i,j\in s} I\left(X_i, X_j\right) \tag{2}$$

$$max\, \Phi\,(D, R),\ \Phi = D - R \tag{3}$$

Where max D is the term for maximum dependency. We need to find the proper feature subset S for which the dependency on the target class C is strongest. $I(\mathrm{X}_i, \mathrm{C})$ is mutual information values between individual feature X_i and class C. min R is the term for the minimum redundancy. We need to find the minimal average mutual information value between each feature in the feature subset S.

3.3 Logistic Regression Model

Our classification problem is to predict whether the borrower will default or not. We regard this as a problem of binomial classification. In this research, we used Logistic Regression (LR) as classification model. LR is one of the most widely used machine learning models for classification purposes, and the computational cost is relatively low. LR can calculate the probability that the sample belongs to 0 or 1.

LR has been used as a credit scoring model for a long time [2, 8, 9]. In LR, Sigmoid function is used for convergent. Sigmoid function is shown follows:

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

Where x is the feature vector of the borrower, and θ is the set of parameter values corresponding to each feature.

The probability of this binomial classification is between 0 or 1, as calculated by function (5):

$$\mathrm{P}(h_\theta(\mathrm{x}) = 0, 1|\theta_0, \cdots, \theta_N) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x1 \cdots \theta_N x_N)}} \tag{5}$$

Since logistic regression gives us a probability that the sample is closer to class 1 or 0, we must set a threshold to classify the result into 1 or 0. If the probability exceeds the threshold, we classify the sample as 1. In our experiment, we set the threshold to 0.5. The imbalance in data leads to strong inclination towards classifying to larger class of data. We use undersampling to balance the dataset. After feature selection and undersampling, we train classification model to train the credit risk prediction model.

4 Experiment and Results

The period of the data, collected from LendingClub website, is from 2016 Q1 to Q4. After data preprocessing, we have 117,790 borrowers with 18 features available for the whole dataset. Those features can be categorized into loan information, applicant information and some other information, as shown in Table 1.

In the other information part, Purpose, Home ownership and Verification status are text type features. We did binary coding to map those features to numerals, e.g., "Home ownership" – yes or no is changed to 1 or 0. After the processes, the total number of available features are 33.

Table 1. Features for the whole dataset.

Issues	Loan information	Applicant information	Other information
Features	Loan amount	Annual income	Purpose
	Term	Employment length	Public recall
	Installment	Debt to income ratio	Home ownership
	Interest rate	Total account	Verification status
		Open account	Delinquency in 2 years
		Inquire in last 6 months	Earliest credit line
		Revolving balance	
		Revolving utilization	

For feature selection we tested both mRMR and LASSO. Figure 4 is the change of feature coefficients as the value of changes, in LASSO regression. We applied cross-validation to find the best lambda, which is shown as blue dashed lines. The best lambda value after cross validation check is 0.0005467 (nature log equal to −7.51). After the selection, LASSO regression suggested to keep 27 features.

However, with LASSO regression it can only reduce 6 features from 33 to 27. As most of the feature were retained, the reduction in computational cost is not very significant. Due to that, we applied mRMR as the other feature selection method, and selected top 10 features, as shown in Table 2 below.

The ten important features selected by mRMR have 6 different type information of borrowers, they are: (1) Interest rate is the loan interest rate recommended by LC after evaluating the borrower, (2) Home ownership is the relationship between the borrower and the owner of the house in which the borrower currently resides. (3) Debt to income ratio is calculated by dividing borrower's total recurring debt by borrower's monthly income. (4) Delinquencies is the number of delinquencies for borrower in the last two years. (5) Income to payment ratio is calculate by dividing borrower's monthly income by borrower's monthly payment. (6) Inquiries in 6 months is the number of credit inquiries for borrower in the last six months.

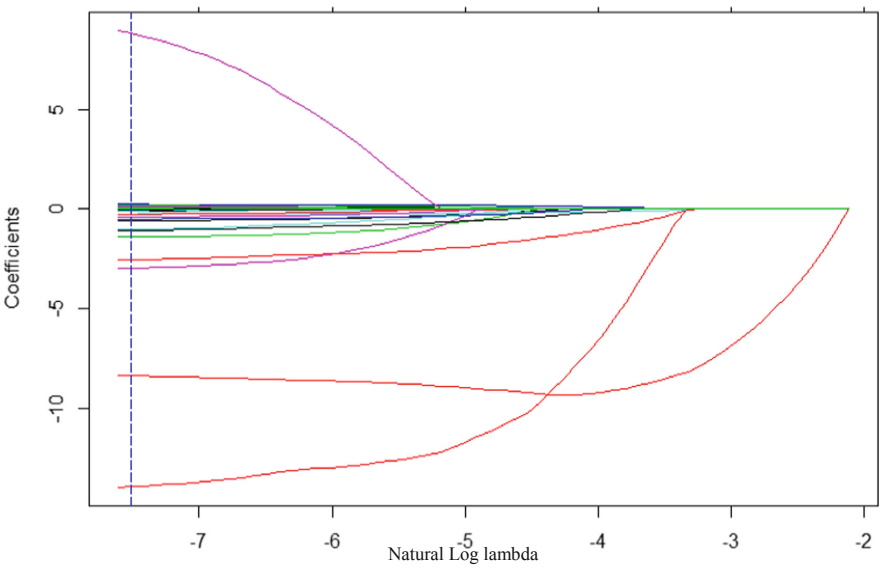

Fig. 4. The convergence changes in LASSO regression

Table 2. Selected features by mRMR

Feature	Original data type	Data manipulation
Interest rate	Numeric	Normalization-range 0 to 1
Home ownership (ANY)	Binary	None
Home ownership (RENT)	Binary	None
Loan purpose (Medical)	Binary	None
Debt to income ratio (DTI)	Numeric	Log or scale
Delinquencies	Numeric	Normalization-range 0 to 1
Loan purpose (Small business)	Binary	None
Income to payment ratio (ITP)	Numeric	Log or scale
Inquiries in last 6 months	Numeric	Normalization-range 0 to 1
Loan purpose (renewable energy)	Binary	None

Loan interest rate is the decision after the evaluation of borrower, an evaluation by LC. On the other hand, both DTI and ITP in important financial information of an applicant in the lending business, because higher is the DTI, higher is the risk that the borrower will default. This is the opposite of ITP. To observe the relationship between those important features mRMR selected, we illustrate them in Figs. 4 and 5. For the quality of data visualization, we take 0.1% of the data by random sampling from raw data, which contains data points from about 150 borrowers.

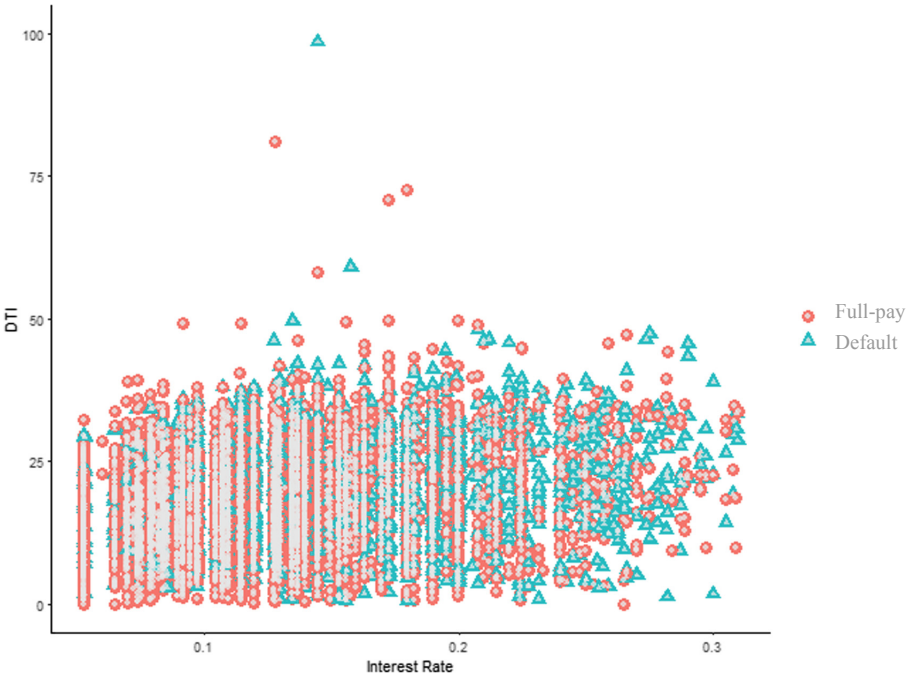

Fig. 5. Debt to income (DTI) ratio and interest rate in different loan status.

In Fig. 5, orange circles represent who fully pay back the loan. The green triangles represent default borrowers. It is clear that most of the borrowers who have both low DTI and low interest rate are full-pay borrowers. As the interest rate and DTI both getting higher, more borrowers default of their loans. This is reasonable, because higher interest represents a borrower with bad credit rating, and higher DTI also means that the borrower's debt may exceed the borrower's affordable range leading to non-payment.

In Fig. 4, it is more difficult to find the relation between ITP and loan status. Once the distribution of defaulting borrowers is closely observed, we find that, there is a high concentration of ITP lower than 0.25. This can be a reference when an investor need to be treated important goal, choose a borrower to lend money. In the financial market, both DTI and ITP can represent as credit rating. However, neither can significantly express whether the borrower will default or not. It is apparent that more information is needed for better classification of loan status at P2P lending (Fig. 6).

For classifier, we compared Logistic Regression (LR) and Random Forest (RF). In this research, we used Negative Accuracy (NA) to measure the accuracy of prediction of bad borrower, Positive Accuracy (PA) to measure the accuracy of prediction good

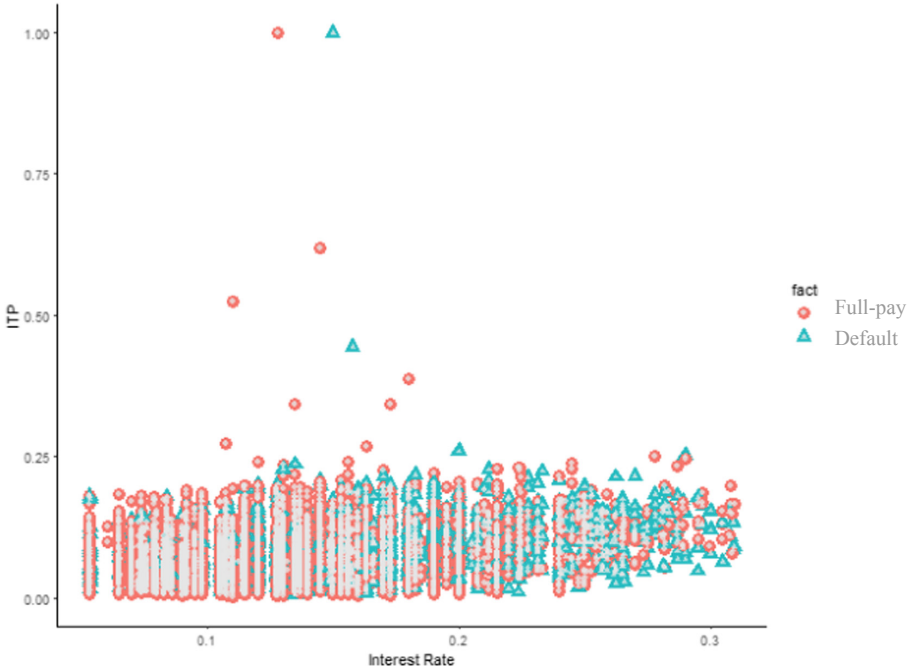

Fig. 6. Income to payment (ITP) ratio and interest rate in different loan status.

borrower, Geometric mean (GM) which is calculated by square root of PA multiplied by NA, and Total Accuracy (TA) which is the overall accuracy of the test data set. TA, PA, NA and GM are calculated by True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), using the following formulas.

$$\text{TA} = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$\text{PA} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{NA} = \frac{TN}{FN + TN} \tag{8}$$

$$\text{GM} = \sqrt{PA * NA} \tag{9}$$

We compared results using all features and with features selected by mRMR. Experiments were done using unbalanced data and after balancing using undersampling. The results using two different classifiers, LR and RF, are shown in Table 3. Our target is to get the best GM value

Table 3. Results of comparison.

Classification model	Logistic regression				Random forest			
Experiment	NA	PA	GM	TA	NA	PA	GM	TA
No feature selection No data balance	0.160	0.957	0.391	0.751	0.153	0.956	0.382	0.748
LASSO No data balance	0.763	0.547	0.646	0.748	0.764	0.553	**0.649**	0.749
mRMR No data balance	0.154	0.957	0.384	0.749	0.104	0.973	0.313	0.748
No feature selection Balance data	0.663	0.635	0.647	0.650	0.681	0.600	0.639	0.649
mRMR Balance data	0.637	0.650	**0.649**	0.644	0.673	0.604	0.641	0.646
LASSO Balance data	0.648	0.637	0.642	0.642	0.641	0.647	0.644	0.644

It is interesting to note that without data balancing, LASSO can achieve better GM (0.649 compared to 0.3), though the model is very low at positive accuracy. This means the model is too strict, and a lot good borrowers had been classified as bad borrower. But the performance is better. So we are interested in to know which features had been filtered out as unnecessary features. They are listed below (Table 4).

Table 4. Features LASSO throw away.

No.	Features
1.	Funded amount in investment
2.	Installment
3.	Home ownership “ANY”
4.	Home ownership “OWN”
5.	Verification status “Source Verified”
6.	Purpose “credit card”

For comparison the result from different experiments, we used histogram to show the GM at from each experiment, shown in Fig. 4. LR mean Logistic Regression was used as classifier, RF means Random Forest was used classifier. The best GM outcome, 0.649, apparent twice (1) imbalance data feature selection by LASSO and classifier RF, (2) balance data with feature selection by mRMR. In additional, we can observe that GM is significantly improved after balancing the data using undersampling with or without feature selection and using mRMR for feature selection. With LASSO for feature selection, both LR and RF can reach very high GM. However, LASSO can eliminate only 6 features out of 33. With mRMR, we can reduce features from 33 to 10, leading to much more computing cost reduction for training of the classifier. On the other hand, mRMR selected only 10 featured out of 33. Even with only 10 features,

high GM and TA were achieved. This means that mRMR can effectively reduce the number of features. Finally, we did not observe that Random Forest have obviously better performance than Logistic Regression, where LR is trained much faster compared to RF (Fig. 7).

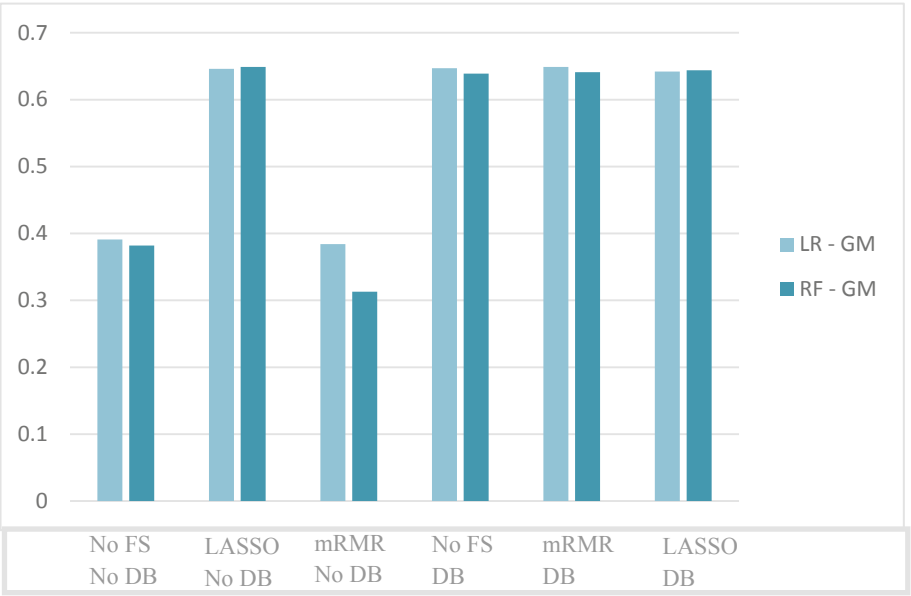

Fig. 7. Comparison of each experiment

5 Conclusion and Feature Work

In this work, we used dataset from LendingClub. The period is from 2016Q1 to 2016Q4. We applied both LASSO regression and mRMR for feature selection and compared machine learning models Random Forest and Logistic Regression, with all features and selected features.

The results show that NA and GM are poor if the data is not balanced. The training data has 33 features. LASSO selected 27 features out of 33, and achieved higher GM. We therefor conclude that those 6 feature were irrelevant for classification. mRMR on the other hand selected only 10 features out of 33, and achieved almost similar result with balanced data. When, lowering number of feature for faster training, mRMR is much effective.

For quick physical interpretation lower number of features is important. Feature selection can help investors find more meaningful indicators, to help them make right investment decisions. Also, we can observe that Logistic Regression, with much lower computational cost, can achieve similar or even better performance than Random Forest. Another important result is that, after balancing the data using undersampling, the accuracy of default borrower prediction improved significantly. However, the

positive accuracy and total accuracy declined, which mean that there will be many good borrowers mistakenly judged as bad borrowers.

How to improve negative accuracy without reducing PA or total accuracy, is one of our feature works. In additional, the motivation of credit risk is not just predicting whether a customer will default or not. The total return on the loss amount is also an important parameter. We will extend our work on estimation of loss and gain on an investment.

References

1. McWaters, J., et al.: The Future of Financial Services: How Disruptive Innovations are Reshaping the Way Financial Services are Structured. Provisioned and Consumed. World Economic Forum (2015)
2. Thomas, L.C.: A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. Int. J. Forecast. **16**(2), 149–172 (2000)
3. Sandberg, M.: Credit Risk Evaluation using Machine Learning (2017)
4. Birla, S., Kohli, K., Dutta, A.: Machine learning on imbalanced data in credit risk. In: 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1–6. IEEE, Vancouver (2016). https://doi.org/10.1109/iemcon.2016.7746326
5. Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. **39**(3), 3446–3453 (2012)
6. Ashcraft, A.B., Schuermann, T.: Understanding the securitization of subprime mortgage credit. Found. Trends Finan. **2**(3), 191–309 (2008)
7. Board, Financial Stability, FinTech Credit: Market Structure, Business Models and Financial Stability Implications. Financial Stability Board, Basel (2017)
8. John, C.W.: A note on the comparison of logit and discriminant models of consumer credit behavior. J. Finan. Quant. Anal. **15**(3), 757–770 (1980)
9. Dong, G., Lai, K.K., Yen, J.: Credit scorecard based on logistic regression with random coefficients. Procedia Comput. Sci. **1**(1), 2463–2468 (2010)
10. Malekipirbazari, M., Aksakalli, V.: Risk assessment in social lending via random forests. Expert Syst. Appl. **42**(10), 4621–4631 (2015)
11. LendingClub, June 2018. https://www.lendingclub.com/info/download-data.action
12. Peng, H., Fuhui, L., Chris, D.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)