The Application of Random Forest in Individual Credit Risk Management

Yuqing Gu

No.99 Shangda Rd. Shanghai, China

Yuqing_gu@shu.edu.cn

Keywords: Credit Risk Management; Random Forest; Model Comparison; P2P Lending

Abstract. Credit risk has always been one of the primary risks in the field of peer-to-peer (P2P) lending. As an attempt to address the problem, in this paper, we use the random forest method to conduct an empirical study on individual credit forecast using personal credit data from RenRenDai, a popular online P2P lending platform in China. Additionally, the five-fold cross-validation method is introduced to compare the effectiveness across different models including random forest, logistic regression and support vector machine (SVM). Results show that factors such as income, Marriage, birthplace, credit card limit, loan and credit card overdue, have significant effects on predicting personal credit risk; And the random forest model established is superior to logistic regression and support vector machine (SVM) in terms of Accuracy and Specificity.

Introduction

Peer-to-peer (P2P) online lending is the fruit of China's recent financial innovation and "Internet plus" innovation as it exploits the potential of online banking and the internet technology to establish credit relationship between borrowers and lenders. At present, P2P online lending has become one of the fastest growing industries in the internet finance sector due to its low threshold and insecure loan. However, despite its rapid development, the industry is also facing huge challenges. Since the second quarter of 2018, Online P2P platforms have experienced waves of bankrupts that shocked the whole country. Credit risk has always been the main risk faced by the industry [1]. Commercial banks and other traditional financial institutions have established relatively sound credit risk management processes and operational standards based on the Basel Capital Accord, keeping the non-performing asset ratio to a certain level. In contrast, emerging Internet P2P lending platforms, facing a large number of potential credit needs, have not yet established a complete and effective credit risk management system, which prevents the further development of the industry. As the basis of credit risk management, individual credit data analysis is self-evident. Through modeling and analyzing personal credit data, P2P Online platforms are able to predict borrowers' default probability and classify potential borrowers, which lowers business risks and promotes the healthy development of China's Internet finance.

Literature Review

Researches in credit risk assessment abound, with various scholars conducting empirical studies using both traditional and emerging methods. Traditionally, models employed in the finance industry can be classified into default models and market-to-market models. The former models, such as Credit Portfolio View developed by McKinsey, emphasis on default loss while the later ones, such as Credit Matrix by JP Morgan, exam the day-to-day change in the market value. Recent years witness studies digging into the impact of new factors on individual credit risk, for instance, the value of social networks[2], culture differences[3] and text describe[4]. With the availability of internet credit data and the recent progress in artificial intelligence, machine learning has become one of the leading methods to explore the area. clustering[5], decision tree, support vector machine (SVM), logistic regression[6], random forests[7], neural network[8] and their hybrids[9, 10] has been applied to credit risk management. Malekipirbazari and Aksakalli (2015) compared popular machine learning methods including k-means clustering, Logistic regression, support vector machine (SVM) and

  DOI: 10.25236/icmcs.2019.036

random forests on the data from Lending Club suggesting RF-based method comes superior to other models in terms of accuracy and volatility and outperforms the FICO credit scores as well as LC grades in identification of good borrowers. Random forest has been employed in many fields and considered one of the most promising methods.[11]

Random Forest

Random Forest (RF) is an ensemble learning method proposed by Breiman in 2001. It deploys decision tree as base learners and introduces averaging methods with the help of bagging (Bootstrap AGGregatING). As a result, random forests correct for decision trees' habit of overfitting to their training set, with the expense of slightly higher bias and loss of interpretation. In this paper, RF-based methods are used to model individual credit risk. First, bootstrap sampling is used to extract k subsets namely $\{D_1,D_2,...,D_k\}$ from the whole data set D; Secondly, decision tree is run on each subset so we can obtain k classification model sequences and their corresponding result sequences namely $\{h_1(x), h_2(x),...,h_k(x)\}$, where x is the matrix of selected personal credit risk features. Finally, voting is adopted to average the prediction results of k base learners. The combination rules are as shown in Eq. 1:

$$H(x) = arg \max_{Y} \sum_{i=1}^{k} I(h_i(x) = y). \quad (1)$$

Where $H(x)$ is the final result, with $I(\cdot)$ being indicative function and y being the output given by the decision tree i.

The out-of-bag data (OOB) generated by the Bagging method can be used to estimate the generalization error of random forest and the importance of features. To measure the importance of a particular feature, we first permute this feature among the training data and calculate the out-of-bag error. The importance score for the feature is computed by averaging the differences in out-of-bag error before and after the permutation over all trees. The larger the change, the more crucial the feature is. The Gini index we used in this paper is calculated as Eq. 2 and Eq. 3

$$Gini(D^v) = 1 - \sum_{i=1}^{|y|} p_i^2. \quad (2)$$

$$Gini(D,i) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) . \quad (3)$$

Where $|y|$ is the number of outcomes of feature i; $p_i$is the proportion of samples falling into category i; $|D|$ is the total number of samples; $|D^v|$ is the number of samples that fall into category v.

Empirical Study

Data. Data used in this paper is collected through RenRenDai, one of the leading online P2P platforms in China using web spider programs. We collected a total number of 30000 records and address the miss value problem by multiple imputation (MI) method. However, the distribution of the positive and negative sample is extremely uneven as most of the borrowers made the repayment on time. The number of negative samples is much larger than the positive samples. Since most models use 0.5 as the threshold, there seem to be difficulties for some of the models to make the prediction. Therefore, this paper employs the downsampling method which takes all 1875 positive samples in the data set and randomly selects 1875 negative samples from the data set. The final sample size is 3,750.

Feature selection. Whether the borrower will default or not is decided by his ability and willingness to repay. The repayment ability makes sure that the borrower objectively has sufficient to pay the principal and interest. The borrower's repayment ability is largely affected by the income in that the borrowers with higher income are more likely to make the repayment. The willingness to repay is whether the borrower subjectively wants to make the repayments. Borrowers' repayment willingness is largely influenced by their general behavioral habits. Borrowers with good credit history tend not to default, while borrowers with awful credit history tend to continue to default.

Based on the analysis above, we choose a total number of 13 features related to personal credits risk, including basic information, the ability and willingness to repay, as shown in Table 1.

Table 1 Feature Explanation

Category	Feature	Meaning	Remarks
Basic information	has_fund	Has Housing provident funds or not	1 = has housing provident funds
	is_local	Local or not	1 = local
	married01	Married or not	1 = married
	divoced01	Divorced or not	1 = divorced
	bachelor	Own bachelor degree or not	1 = own bachelor degree
Ability	salary	Level of salary	
Willingness	fraud	Conducted fraud or not	1 = conducted fraud
	current_od	Loan overdue this month of not	1 = loan overdue this month
	sumoc	No. of loan overdue	
	sumom	No. of loan overdue month	
	credit	Credit card limit	
	current_od_cd	Credit card overdue this month or not	1 = credit card overdue this month
	curren	Own foreign currency credit card or not	1 = own foreign currency credit card

Result. By calculating the change of the Gini index of 13 variables in Table 1, it is found that income level, credit card limit, loan overdue related information, bachelor degree, housing provident funds, foreign currency credit card and birthplace seems to be some most important features, while credit card overdue related information is of less important. Therefore, in this paper we consider to select the top 10, 12, 13 features respectively to establish 3 random forest models.

Table 2 Empirical Results

Set	Indicators	10 Features	12 Features	13 Features
		Model I	**Model II**	**Model III**
Training set	Accuracy	0.92	0.95	0.94
	Specificity	0.95	0.93	0.91
	Sensitivity	0.89	0.97	0.97
Test set	Accuracy	0.93	0.94	0.93
	Specificity	0.96	0.90	0.88
	Sensitivity	0.91	0.97	0.97

As shown in Table 2, with the increase of variables in the model, the accuracy of prediction maintains roughly the same. What needs to be mentioned is that while sensitivity is on the rise, specificity tends to decline with the additional variables. A declining specificity means the probability of predicting a default borrower as a non-default borrower increases, thus making the lending platform loss money. Therefore, we consider Model I to be the best model.

Model Comparison. The results of the 5-fold cross-validation method in Table 3 shows that compared with the logistic regression and support vector machine models, the random forest model has a higher overall accuracy, with the prediction accuracy and F1 value reaching 92%, which is higher than 72% of logistic regression and 74% of support vector machines. Besides, the recall and precision rate of the RF model is also relatively high, making it superior to the other two models.

Table 3 Model Comparison

Model	Accuracy	Sensitivity	Specificity	F1
RF	**0.92**	**0.95**	**0.89**	**0.92**
Logistic	0.69	0.8	0.58	0.72
SVM	0.71	0.83	0.58	0.74

For non-default samples, RF has a prediction accuracy of 95% higher than that of the logistic regression and support vector machine method. In the actual lending business of financial institutions, the forecast of default samples is critical. The loss of predicting a normal sample as a default sample is much lower than the other way around. In terms of the prediction accuracy of default samples, the accuracy of random forest prediction is 89%, which is much higher than 70% of logistic regression and 56% of support vector machines. In summary, the random forest method is superior to logistic regression and support vector machine algorithms in terms of overall prediction accuracy and sensitivity, especially its predictive effect (predictive ability for default samples) is much higher than other models thus having high application value.

Conclusion

Risk credit assessment and management is of great significance for standardizing the online P2P lending market and the orderly development of the financial system. With the development of network credit investigation and the progress in artificial intelligence, machine learning methods have become one of the leading techniques to explore the area. In this paper, a personal credit assessment system is established based on borrowers' basic information, repayment ability, and repayment willingness and then applies to random forests method for further analysis. Empirical study shows that income, Marriage, birthplace, credit card limit, loan and credit card overdue, have significant effects on predicting personal credit risk. In terms of prediction ability, random forests method is superior to logistic regression and support vector machine (SVM) in Accuracy and Specificity. Its excellent performance makes it an ideal model in personal credit risk field.

Surely, although random forest method has relatively high performance, there is still room for improvement, especially in the prediction accuracy of default samples (Specificity). To further improve the performance of the model, a more flexible individual credit risk assessment system could be built to capture the credit risk in a more comprehensively and accurately way. Besides, a more industry-specific missing value filling method can be established to reduce the noise in the training and test set which could also be the future direction of work.

References

[1] X. Lei, Discussion of the Risks and Risk Control of P2P in China, Modern Economy 07 (2016) 399-403.

[2] M. Lin, N. Prabhala and S. Viswanathan, Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending, Management Science 59 (2013) 17-35.

[3] G. Burtch, A. Ghose and S. Wattal, Cultural differences and geography as determinants of online pro-social lending, MIS Quarterly 8 (2014) 773-794.

[4] G. Dorfleitner, C. Priberny and S. Schuster, et al, Description-text related soft information in peer-to-peer lending - evidence from two leading European platforms, Journal of Banking and Finance. 64 (2016) 169-187.

[5] H. Li, Y. Zhang and N. Zhang, et al, Detecting the abnormal lenders from P2P lending data, Procedia Comput. Sci. 91 (2016) 357-361.
[6] R. Emekter, Y, Tu, B and Jirasakuldech, et al, Evaluating Credit Risk and Loan Performance in Online Peer-to-Peer Lending, Applied Economics 47 (2015) 54-70.
[7] X. Ye, L. Dong, and D. Ma, Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score, Electronic Commerce Research and Applications. 32 (2018) 23-36.
[8] H. Bekhet, S. Eletter. Credit risk assessment model for Jordanian commercial banks: neural scoring approach, Rev. Develop. Finance 4 (2014) 20-28.
[9] X. Ma, J. Sha and D. Wang, et al. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning, Electronic Commerce Research and Applications 31 (2018) 24-39.
[10] G. Ke, Q. Meng and T. Finley, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 31st Conference on Neural Information Processing Systems (Long Beach, CA, USA).
[11] Malekipirbazari, M. and V. Aksakalli, Risk assessment in social lending via random forests, Expert Systems With Applications 42 (2015) 4621-4631.