Analysis of P2P Online Lending Default Based on Random Forest

Haojiang Cai
School of Science, Xi’an Shiyou University, Xi’an 710065, China
imagic725@163.com

Abstract. This article, based on Renren Loan website, will utilize SMOTE to process the debit and credit data in a balanced way. In addition, the importance for the variable-Random Forest and cross-validation conception will be applied for feature selection before the parameter optimization by grid searching, so that the model of basic Random Forest will be derived. Finally, featured modules of LDA will be added to further explore the reference value of loan description. The research findings show that the model has a satisfactory performance on test set in terms of predicting the results.

1. Introduction
With the craze of Internet finance, P2P online lending has developed rapidly in China as a new credit model for providing peer-to-peer credit loans[1]. P2P network lending is an Internet financial model in which borrowers and lenders operate directly through the network platform without any financial intermediaries. The online lending platform uses social platforms to expand users, cloud computing to improve computing power and data mining technology to improve the risk control capacity. However, due to the burgeoning P2P online lending platform in China, the credit information system is still not perfect, and the situation of platform cheating and borrower defaults still occurs[2]. By August 2018, the number of normal operation of P2P online lending platform is 1595, the number of problem platforms is 4811, total number of platforms is 6406, which is a record high.

At the same time, the P2P online lending platform is generating massive transaction data every day. In the era of big data, how to use this data to obtain useful information about users, thus improving the risk control capability is particularly important. The research started late and is still in the induction phase compared to foreign countries[3].

In order to better establish a model suitable for China’s situation, two year’s loan data have been captured from the domestic “Renren Loan” platform by using the web crawler program. Next, data analysis will be carried out and the model will be established, and finally, variable selection as well as model optimization and evaluation are to be conducted.

2. Research design

2.1. Model introduction
At present, domestic and foreign scholars analyze the machine learning tools of P2P default loans, which are generally traditional logistic regression models or a simple neural network model[4-5]. However, the Logistic model requires that the independent variables are independent of each other and sensitive to multicollinearity. The actual situation of the data is difficult to meet the conditions of the Logistic model[6-7]. In addition, the neural network is equivalent to the black box, which can not
effectively explain the model in the sense of economics. The random forest model can solve the above two problems well.

2.2. Generalized error theory of random forests

At present, the researches on classification hope that the classifier can correctly classify the unknown category data. The algorithm optimization goal is to reduce the generalization error and improve the classification accuracy[8].

Theorem 1 As the number of trees increases, for all parameter sequences \( \theta_1, \theta_2, \cdots \), generalization error \( PE \) will always converges to

\[
\lim_{n \to \infty} \max_{\theta \neq \theta'} P_{\theta}[h(X, \theta) = j] \leq \epsilon.
\]

When the number of classification trees in the forest is large enough, the generalization error of the random forest algorithm will converge to a limit value[9-10].

Theorem 2 The upper bound of the generalization error can be given as follows[9]:

\[
PE^* \leq \frac{\beta(1-s^2)}{s^2}.
\]

The theorem gives an upper bound of generalization error and also shows that increasing the classification intensity of a single decision tree and reduce the correlation between any two decision trees can improve the generalization performance of combined classifiers.

2.3. Advantages of a random forest model

Random forest model has the following advantages:

1. The random forest algorithm can deal with complex high-dimensional data and give the variable importance measures (VIM), reducing the complexity of the model and making the model more efficient;

2. In the face of imbalanced data, random forests are still suitable for SMOTE[11];

3. When sampling by using the Bagging[12], about one-third of the data is used for verification. The error estimated by the validated sample has the same accuracy as that of the sample set separately from the validation set.

2.4. Data introduction and preprocessing

This paper builds a model by using the data of Renren Loan. The dependent variable is the loan status, which is divided into Paid and Default, and Paid in the original data is 89.02%, Default is 10.98%, the data is not balanced.

There are a total of 14169 data and 35 characteristics. There are many factors affecting the P2P loan default. According to the literature, it can be divided into 3 categories: hard information of borrowing: borrowing interest rate, borrowing amount, etc.; soft information of the borrower, such as: age, gender, marital status, etc.; soft information of borrowing: type of loan, description of loan, method of repayment, etc. [13]

2.5. Model evaluation method

For classification problems, the label “1” is a positive class, representing a default record, the label “0” is a negative class, representing the paid record. The confusion matrix is used to present the classification result, as shown in Tab.1.

| Actual category | Forecast category |  |
|-----------------|-------------------|---|
| +               | True positive (TP)| False negative (FN) |
| -               | False positive (FP)| True negative (TN) |
The confusion matrix contains the basic information of the classification results of the classifier. In order to compare the performance of the classifier more specifically, a more detailed classification evaluation index is given:

1. **Accuracy** indicates the ratio of the number of correctly classified samples to the total number of samples:
   \[
   \text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP}.
   \]

2. **Precision** indicates how many of the samples with positive predictions are true positive samples:
   \[
   \text{Precision} = \frac{TP}{TP + FP}.
   \]

3. **Recall** indicates how many positive examples in the sample are predicted correctly:
   \[
   \text{Recall} = \frac{TP}{TP + FN}.
   \]

4. **F₁-Score** is the mean of the reconciliation between the accuracy rate and the recall rate:
   \[
   F₁ - \text{Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.
   \]

5. **ROC curve and AUC**: Receiver Operating Characteristic curve is a graphical analysis tool\cite{14}. ROC Space defines the false positive rate (FPR) as x axis, true positive rate (TPR) is defined as y axis. ROC area under the curve of ROC is called AUC. The value of AUC is between 0 and 1, the larger value of AUC, the higher the classification performance of the model.

   The point closest to the upper left of the graph is the critical value with higher sensitivity and specificity. Any subsequent model will select this threshold as one of the hyperparameters of the model. After that, the best model is filtered out according to Accuracy, **F₁-Score** and AUC.

3. **Modeling of default prediction based on random forest**
   The pre-processed data is balanced by SMOTE; the balanced training set is modeled by random forest model, using three criteria to measure the importance of variables and select the most critical features; traversing the hyperparameters combination model to obtain optical parameters, combining the results of the training set and the validation set to select the best parameter combination, and obtaining a complete random forest model; finally, use the test set to test the final model.

3.1. **Handling imbalanced data**
   Due to the strict review process of the P2P lending platform, the proportion of default records is much smaller than the proportion of repaid records. Therefore, the model tends to learn the characteristics of most types of samples, and predicts most of the test set samples as “repaid”, can not correctly identify the default record\cite{15}. Therefore, it is suggested to use Synthetic Minority Oversampling Technique (SMOTE), initiated by Chawla et al\cite{16}. A large number of empirical results show that the SMOTE can effectively reduce the over-fitting and improve the learning ability of the classification algorithm for a few classes. The simple data sampling method is used to divide the original data into Training Set, Validation Set and Test Set by 70%, 15% and 15% respectively. The training set is used to build the model, the validation set is used to parameterize and filter the model, and the test set is used to evaluate the model.

3.2. **Feature selection of random forest models**
   The variable importance scores of random forests are OOB and Gini. They are respectively listed for Mean Decrease Accuracy and Mean Decrease Gini. Set the number of trees large enough to establish a preliminary random forest model, get the convergence model error rate, and get the variable importance score, as shown in Fig.1.
In order to make the results of the model more stable, the idea of five-fold cross-validation is adopted to determine the optimal feature subset [17]. Finally, 12 features were obtained.

### 3.3. Parameter determination of random forest model

In addition to the features that need to be determined when establishing a random forest, it is also necessary to manually adjust the model hyperparameters, including the number of the tree in the random forest (ntree), the number of split point candidate feature (mtry), and the maximum value of terminal nodes used to control the size of decision tree (MaxNodes).

The selection criteria of ntree are to ensure that the model error converges, using the training set to determine ntree=80. The selection criteria of mtry are to minimize the out-of-bag error of the model and use the training set to determine mtry=5. The selection criteria of MaxNodes are to prevent overfitting.
In the training set, as the maximum value of terminal nodes increases, the accuracy of the model and F1-Score gradually rises and reaches 1, but in the validation set, as the maximum value of terminal nodes increases, the accuracy and F1-Score of the model rises rapidly and starts to fluctuate up and down. This also shows that the model that fits the training set too much does not perform well in the validation set.

In the training set, as the maximum number of terminal nodes increases, the AUC of the model gradually rises and reaches 1. In the validation set, as the maximum number of terminal nodes increases, the AUC of the model rises rapidly and then begins to rise slowly and converge gradually.

Integrating three evaluation criteria, when MaxNodes=80, the model already has a high accuracy, accuracy and recall rate.

3.4. Test set performance
The performance evaluation indicators of each classification are shown in the Tab.2.

| Evaluation index | Accuracy    | Recall rate | F1-Score | AUC     |
|------------------|-------------|-------------|----------|---------|
| Validation set   | 0.9053      | 0.973       | 0.70     | 0.958   |
| Test set         | 0.9157      | 0.955       | 0.725    | 0.933   |

3.5. Analysis of results
Compared to the validation set, the AUC of the test set decreased, but F1-Score has a certain degree of increase, which means that the model can be extrapolated outside the training data.

4. Prediction model based on LDA and random forest
In practice, the P2P lending platform pays more attention to borrowing hard information and borrower soft information, while there is a shortage of borrowing soft information. Adopting the LDA topic model to establish feature engineering sub-modules for feature description of borrowing in the dataset, which can significantly improve the performance of the classifier[18].

The importance of the model variables after adding new features is shown in the Fig.3.

![Fig.3 Variable importance of the improved model](image_url)

| Evaluation index | Accuracy    | Recall rate | F1-Score | AUC     |
|------------------|-------------|-------------|----------|---------|
| Original model   | 0.9157      | 0.9550      | 0.7250   | 0.9330  |
| New model        | 0.9297      | 0.9911      | 0.7668   | 0.9564  |
| Growth           | 1.54%       | 3.77%       | 5.77%    | 2.51%   |
In each classification performance index, the improved model has a certain degree of improvement than the original model, indicating that the feature screening under the LDA theme model has a significant effect on the performance improvement of the classifier.

5. Conclusion
Random forest is a typical combination classifier, which can overcome the problem of over-fitting and local optimization of single classifier, and has better classification effect and robustness. After applying the random forest model in the field of P2P personal credit risk assessment, a classification model with good effect is obtained. Then the LDA feature engineering sub-module is introduced on the basis of the traditional random forest model, and the reference value of borrowing description is explored. The empirical results show that the classification performance indicators are significantly improved.

References
[1] Zhang Chunxia, Cai Yanhong, Liu Wei. Research on pricing strategy of P2P online lending platform under competitive conditions[J]. Journal of Tsinghua University(Science and Technology), 2015(04): 100-104(in Chinese).
[2] Wang Xiuhua, Meng Lu, Ouyang Hui. Analysis of P2P Network Lending Problem Platform Characteristics and Investor Identification——Evidence from 222 Platforms[J]. Finance & Trade Economics,2016(12)(in Chinese).
[3] Feng Guo, Jiang Shasha. On the Alienation and Supervision of China's P2P Network Loan Platform[J]. Law and Business Research, 2013(05): 31-39(in Chinese).
[4] Pang Sulin. Application of Logistic Regression Model in Credit Risk Analysis[J]. Mathematics in Practice and Theory, 2006(09): 131-139(in Chinese).
[5] Liu Hong. Research on Loan Default Identification Based on Neural Network[J].Statistical Research,2008(07):63-67(in Chinese).
[6] Yang Jie, Wu Zhongru. Multi-collinearity problem in fitting data fitting analysis[J].Journal of Sichuan University (Engineering Science Edition),2005(05):22-27.
[7] Tao Ran. Diagnosis and Improvement of Multicollinearity Problem in Logistic Model[J].Statistics & Decision,2008(15):24-27(in Chinese).
[8] Quinlan,J.R. Induction of decision trees [J]. Machine Learning 1986,1:81-106.
[9] Breiman L. Random forests [J]. Machine Learning, 2001,45(1):5-32.
[10] Breiman L, Friedman J, Olshen R, etal. Classification and Regression Trees [M]. New York: Chapman & Hall, 1984.
[11] Nitesh V. Chawla , Kevin W. Bowyer , Lawrence O. Hall , W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research 16 (2002) 321–357.
[12] Breiman L. Bagging Predictors [J]. Machine Learning, 1996,24(2):123-140.
[13] Guan Peng, Wang Yufen. Scientific Document Mining Based on LDA Topic Model and Life Cycle Theory[J].Journal of the China Society for Scientific and Technical Information,2015,34(3):286-299(in Chinese).
[14] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27,861-874.
[15] Chen C, Liaw A, Breiman L. Using Random Forests to Learn Imbalanced Data. Tech. rep., University of California, Berkeley(2004).
[16] Chawla NV, Bowyer KW, etal. SMOTE: Synthetic Minority Over-Sampling Technique [J]. Journal of Artificial Intelligence Research, 2002,16(1).
[17] Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research, 5(Sep):1089–1105.
[18] Guan Peng, Wang Yufen. Scientific Document Mining Based on LDA Topic Model and Life Cycle Theory[J].Journal of the China Society for Scientific and Technical Information,2015,34(3):286-299(in Chinese).