Advances in Economics, Business and Management Research, volume 215

Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)

High Risk Bank Loan Recognition Based on Machine Learning

Zishen Zhao[1,*]

[1]*Department of economics and management, Beijing Jiaotong University, Beijing, 100091, China*
**Corresponding author. Email: 1144604228@qq.com*

ABSTRACT

This paper aims to compare the performance of C5.0 decision tree and random forest in high-risk bank loans' recognition. The author mine 1000 loan information data containing 16 variables and find our model's accuracy is around 0.71. Then the author try to do some change on model parameters to improve the model. The author finds that increasing the cost of false negative in error matrix can help bank better avoid the high risk loan. Compared different trials in decision tree, the author find that the model perform the best when trials equal to 45. Compared different mtry in random forest, the author find that the model performs the best when the mtry equal to 11. Then the author compares the random forest and decision tree according to ROC, sensitivity and specificity. The results show that random forest has strength in ROC and sensitivity, while decision tree has strength in specificity. And due to the random forest's bigger AUC value, the author conclude that random forest model slightly outperformed the tree model.

Keywords*: Bank Loan, Machine Learning, Random Forest Model, Decision Tree Model*

1. INTRODUCTION

1.1. Back Ground

Unregulated loans will lead to excessive overconsumption and speculation in the market, and eventually lead to a financial crisis. In 2008, the financial industry in the United States lacked supervision and widely lured people to borrow money and consume. A large number of subprime borrowers with "low credit and low repayment ability", who could not repay on time, and the houses could not sell at high prices. Finally, banks suffered serious losses, causing the financial crisis. People's consumption mode is gradually advanced, extravagant, blind optimistic about the future prosperity, the loan market. Advance consumption products such as Huabei, Huabei, zero down payment installment, as Ill as the pre-sale of Double 11 and Double 12, are all stimulating people's advance consumption. The essence of advanced consumption is insufficient demand, and advanced consumption is needed to guarantee credibility. Therefore, the society pays more attention to the credibility level of the people.

1.2. Related Work

In the analysis of bank loans based on random forest data, Liang Pei used the random forest model to analyze 1000 data including 21 characteristics of customer loans, the accuracy of his model is 0.777 [1].In the Assessment of Bank Credit Risk Based on Random Forest RFM Model Based on Random Forest, Cheng Yusheng found that the new RFM model using random forest can achieve 0.756 accuracy in predicting bank credit risk, which is higher than that of artificial neural network, KNN and C4.5 algorithm [2]. In the Research on Bank Risk Control Model Based on Random Forest Algorithm, Yuan Jing constructed a random forest model and analyzed 150,000 pieces of data. After the parameter tuning, the final model accuracy was 0.868, which was significantly higher than the accuracy of the decision tree model [3] . The accuracy in other relevant paper of solving similar problem is between 0.7 and 0.85[4-10], so that I can concluede that my model is useful if its accuracy is greater than 0.7.

1.3. Paper Framework

In this paper, my main work can be divided into three part, the research on decision tree model, the research on forest tree model and the comparison between them. For the first part, the author will firstly study the accuracy of decision tree. Then the author will make some change in “costs” and “trials” to improve my decision tree model. For the second part, the author will do similar research on random forest tree to find the best “trials” in ROC value. And for the third part, the author will compare the performance of these two model and make a conclusion. And lastly, the author will make a summary and outlook.

2. DATA

My data set take the public credit dataset on Packtpub, which has 16 variables describing loan and loan applicants, with a total of 1,000 records, named credit.csv. These variables including checking balance, months loan duration, credit history, loan purpose, loan amount, saving balance, employment duration, percent of income, years as residence, age, other credit, housing, existing loans count, job, dependents, phone, and the default. Then the author delete the record having missing value, and make some change on the value to fit in the model. the author set a series of intervals in checking balance, saving balance, and employment duration to transform the numeric data into categorical data. The author assume that checking balance can be divided into “0 DM”, “1-200 DM”, and “ >200 DM”(DM means Deutsche Mark), saving balance can be divided into “<100 DM”, “100-500 DM”, “500-1000 DM”, “>1000 DM”, employment duration can be divided into “<1 year”, “1-4 years”, “4-7 years”, “>7 years”.

3. METHODOLOGY

3.1. Decision Tree C5.0

Decision Tree is one of the most popular machine learning algorithm to solve classification problem. C5.0 is a version of decision tree which can do subtree raising and subtree replacement automatically. C5.0 has two parameter, trials and costs. Trails represents the times of self-help cycles, and its default value is 1. Costs represents the error cost matrix, describing the cost of different errors. In the first step, the author will first run the Decision tree in these two parameters default value. Then the author will figure out how to change the costs or trials to improve my model performance on accuracy.

The first step results of the test set including 100 data are shown in the Table 1, the accuracy is: (38 + 33) / 100 = 0.710, where the false positive probability is 38 / 51 = 0.745, and the false negative probability is 33 / 49 = 0.673.

Table 1. the results of the initial model

Actual¥ predicted	no	yes	Row Total
No	38	13	51
yes	16	33	49
Column Total	54	46	100

Further analysis, from the perspective of risk, banks prefer false positive (accept the loan behavior of users who actually do not default, and are not willing to bear the risk of false negative (loan to users who actually default). However, the above model test set errors are more false negative, so it is better to adjust the error cost matrix of the decision tree C5.0 model cost, error_cost which default value is the matrix(0,1,1,0), nrow =2) takes the matrix (c (0,1,5,0), nrow =2), to increase the cost of true default prediction errors and reduce the number of model making false negative errors. The results are shown in the Table 2, false negative errors appear only four times, and the accuracy of default results was 45 / 49 = 0.918, significantly higher than the previous 0.673. But at the expense, the accuracy is obviously down to 0.52.

Table 2. the results after adjusting error cost matrix

Actual¥ predicted	no	yes	Row Total
No	7	44	51
yes	4	45	49
Column Total	11	89	100

The sampling method of the decision tree can be changed by the caret package afford by R to improve the prediction accuracy of the model. The next experiment the author will try 10-fold cross-validation method to observe the model performance of different values of trials according to kappa value. Figure 1 shows that when the trials takes 45, kappa is 0.343 and the model perform the best, and the corresponding accuracy value is 0.678.

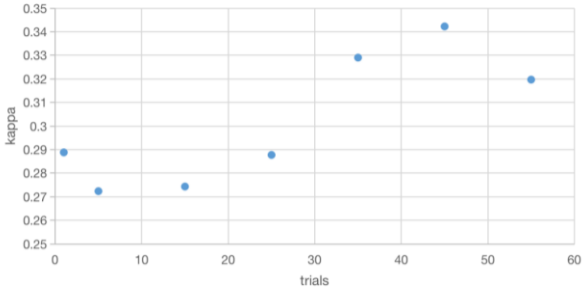

Figure 1. comparison among the kappa value in taking different trials

3.2. Random forest model

Random forests can handle a large amount of data, and also have good adaptability to high-dimensional data. It combines the bagging and the selection of variables to increase the model's diversity. Random forest have 3 parameters, ntree, mtry and metric. Ntree means the node in the model. Mtry means the number of selected variables. Metric is used to measure the model's performance. The random forest model is also available in caret package, the author choose repeatable-10-fold-cross-validation as my model's sampling method and set the value of ntree 500. Then the author will observe model performance in different mtry value according to the ROC value.

The results are shown in the Figure 2, one blue point represents one experiment, and you can get its mtry and ROC value from the coordinate axis. When mtry takes 16, the highest ROC value is 0.722, but when the mtry takes 11, the ROC has firsty reached 0.721. When mtry takes 11,12,15,16, the ROC value almost the same. Then considering the model complexity, my answer to the best value of mtry is 11.

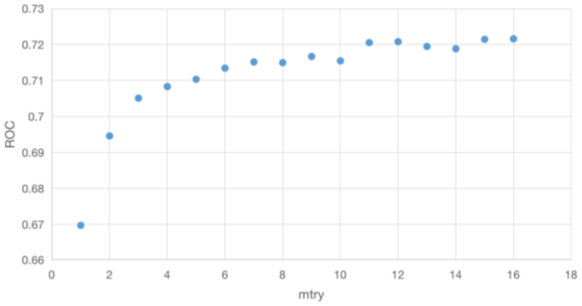

Figure 2. comparison among the ROC value in taking different mtry

3.3. Comparing the performance of random forests to decision trees

The confusion matrix of the random forest describes the out-of-bag error, while the confusion matrix of decision tree is different, so it can't simply compare their performance though the accuracy computing by actual default and predict default. So the author design another experiment to figure out which model is better in this dataset. the author assume that the random forest model with mtry 2,4,8,16 was separately equal to the C5.0 decision tree with trials 10,25,50,100 due to they share the time complexity. Then the author is going to record their performance on ROC value, Sensitivity value and Specificity value. Lastly the author will make a ROC curve to help me make a comprehensive comparison of their performance according to AUC value which means the area under ROC curve. The results are shown in the Table 3 and the Table 4.

(1) Random forest model:

Table 3. the results of Random forest model

mtry	ROC	Sens	Spec
2	0.7041862	0.8742414	0.3609420
4	0.7095348	0.8031149	0.4900181
8	0.7190377	0.7800805	0.5294928
16	0.7233096	0.7597586	0.5579529

(2) C5.0 Decision tree model:

Table 4. the results of C5.0 Decision tree model

trials	ROC	Sens	Spec
10	0.6737436	0.6979655	0.5475543
25	0.6796407	0.7072874	0.5636775
50	0.6862809	0.7210115	0.5702536
100	0.6900807	0.7274598	0.5719384

Compare the corresponding result, all the ROC value and Sensitivity value in random forest model is bigger than decision tree model, while Specificity value is smaller, which mean the random forest is more sensitiy to recognize the high risk bank loan record, but it performs worse in accuracy. To make these results visualization and easy to compare, the author use the plot function to make two ROC curve showing in the Figure 3, the red one describes the random forest and the blue one describes the decision tree. The Area under the curve can decide which model is better by comprehensive considerate the ROC, Sensitivity and Specificity. Apparently the random forest AUC is slightly higher than the C 5.0 decision tree model, which means it has a higher AUC value in analytic geometry. Further from the data, the random forest ROC is 0.732 at the optimal and the C5.0 decision tree is 0.690, the random forest model ROC value is slightly higher than the decision tree. So, the author can conclude that the random forest model's performance is slightly better than the decision tree model .

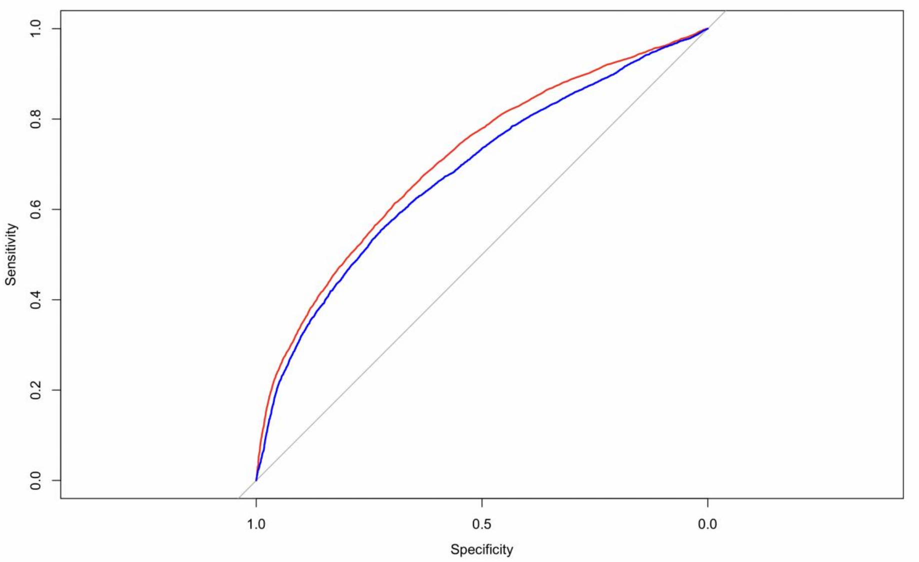

Figure 3. ROC curve of the C5.0 model and the Random forest model

4. CONCLUSION

This paper uses the machine learning algorithm to identify high-risk bank loans, which can reach the level of similar prediction accuracy of 0.7 compared to other's machine learning model on similar subject. So the author think my work is useful for the bank to evaluate of the loan risk. Under this dataset, the random forest model slightly outperformed the decision tree model for the identification of high-risk bank loans. The decision tree model can change the error cost matrix at the cost of increasing the number of false positives to reduce the number of false negatives, essentially making the bank sacrifice part of the loan profit to reduce the risk of loan default; Besides, the decision tree C 5,0, the trials have the highest accuracy of 45. The random forest model, considering the prediction performance and model complexity, takes the model is optimal when the number mtry is 11.

In the future study, it is expected to further optimize the algorithm. the author will try the apply parallel computing method or federated computing method; what's more, the author will try to establish more accurate and specific models, adopting better features and algorithms to better assess the risk of bank loans.

REFERENCES

[1] Liang Pei. Random forest-based bank loan data analysis. Modern Trade Industry, 2020

[2] Cheng Yusheng, and Zou Huan. Evaluation of bank credit risk by random forest based RFM model. Journal of Anqing Normal University (Natural Science edition) , 2018

[3] Yuan Jing. Research on Bank Risk Control Model Based on Random Forest algorithm. Software, 2022

[4] Lv Shuang. Research on the identification of credit risk factors of small and micro enterprises. Chinese market, 2022

[5] Hu ChanJuan, Yu Lianzhi, and Xue Zhen. Iighted random forest algorithm for financial credit risk control based on the Spark framework. Small Microcomputer System,2020

[6] Ding Decchen. Research on the financial distress prediction of commercial banks integrating random forest and support vector machine. The Practice and Understanding of Mathematics, 2020

[7] Yang Xiaowei, Liu Qianqian, and Yu Fang. Empirical analysis of the random forest prediction model of online lending platform data. Journal of Yibin College, 2019

[8] Zhang Jiaqian, Li Ii, and Ruan Sumei. Machine learning-based loan default risk forecast. Journal of Changchun University of Science and Technology (Social Science edition), 2021

[9] Yao Shanshan, and Leng Xiaopeng. Evalation of creit by Random Forest algorithm based on combined feature selection. Computer Systems & Applications, 2022

[10] Cao Taoyun. The study on the improtacne of variables based on Random Forest algorithm. Statistics & Decision, 2022