Bank Direct Marketing Analysis Based on Ensemble Learning

Ruiting Hao[1], **Xiaoqian Xia**[1], **Siyi Shen**[1] **and Xiaorong Yang**[1*]

[1] Department of Statistics & Mathematics, Zhejiang Gongshang University Hangzhou, 310016, China

[*] Corresponding author's e-mail: xryang@zjgsu.edu.cn

Abstract. In the era of Internet and big data, the bank has gradually realized that the traditional data analysis cannot meet the demands of the existing marketing. So the bank direct marketing based on machine learning emerges. However, there are few references which are completely based on ensemble learning. As different banks have different structures of customer data, the existing model cannot be employed directly. Therefore, this article collects the marketing data of a Portugal's bank and compares the classification effects of six different models under three ensemble learning algorithms ---"Boosting", "Bagging" and "Stacking", respectively. Then we select the most appropriate model which has the best performance as the final classifier. Banks can use the classifier to judge whether a customer will order financial products and make direct marketing plans.

1. Introduction

With the trend of digital transformation, the bank has gradually realized that the traditional data analysis cannot meet the new marketing demands, and the bank urgently needs direct marketing driven by big data. Under this background, bank direct marketing based on machine learning emerges. Direct marketing mode can insight the potential requirements and preferences of customers and help banks obtain target customer groups. On one hand, the application of machine learning in bank direct marketing can improve the accuracy of the bank marketing, on the other hand, it can also increase the number of customers.

There are some references that apply machine learning algorithms to bank direct marketing. For example, Elsalamony et al.[1] used neural network and C5.0 models to predict the contact between marketing activities and the customer's behaviors of subscribing deposit. Sing'oei et al.[2] provided a comprehensive framework to achieve direct marketing success using data mining methods. Wisaeng[3] made a comparison of different classifiers for a bank direct marketing dataset. Apampa[4] improved the performance of classification algorithms used in the prediction of bank customer's marketing response by using the Random Forest ensemble. With imbalanced datasets, Marinakos et al.[5] improved the performance of classification algorithms on predicting potential depositors by rebalancing the datasets, and compared the performance of different classifiers. However, there is few reference of bank direct marketing that is completely based on ensemble learning. Inspired by the above studies, we compare the classification effects of six different models under three ensemble learning algorithms ---"Boosting", "Bagging" and "Stacking", respectively.(see[6], [7], [8]) then propose C5.0 model based on the Boosting algorithm to predict whether a customer will buy the bank's financial products.

2. Data analysis

In this section, we will preprocess the original data sets first, then perform feature extraction and train the classifiers. In the training process, we use ten-fold cross-validation to obtain the classification effects of six models under different ensemble learning algorithms. Finally, we select the model with the best classification effect to judge if customers will order financial products.

2.1. Data select and feature extraction

The data in this article is the marketing data of a Portuguese' s bank, which contains 18 variables including 9 class variables and 9 continuous variables. After eliminating the useless variables, we check the data sets for missing values, duplicate values, and outliers. There is no outlier, and subsequent analysis is performed.

After data preprocessing, there are 7 class variables remaining in the data set. But in machine learning algorithms such as XGBoost and LightGBM, the input data must be a numeric matrix. So we perform One-Hot coding on these variables, so that they can be converted into numerical forms.

In order to prevent the difference between variables' magnitude from affecting the classification effect, we normalize the continuous variables in the data set.

2.2. Ensemble learning algorithm

2.2.1. Boosting algorithm. Boosting algorithm is to give each sample the same weight initially. After each classification, the weight of the right result reduces, and the weight of the wrong result increases. Repeat the operation until reaching the threshold or the maximum number of cycles.

Under the Boosting algorithm, we select C5.0 model and GBM model. Then we use ten-fold cross-validation on the two models separately to obtain the classification effect. Table 1 shows the accuracy rate and Kappa values of the two models on the test sets. Figure 1 shows the ROC curves of the two models.

Table 1. The ten-fold cross-validation results of C5.0 and GBM

	model	Accuracy	Kappa
Boosting	C5.0	0.9611	0.9170
	GBM	0.9491	0.8909

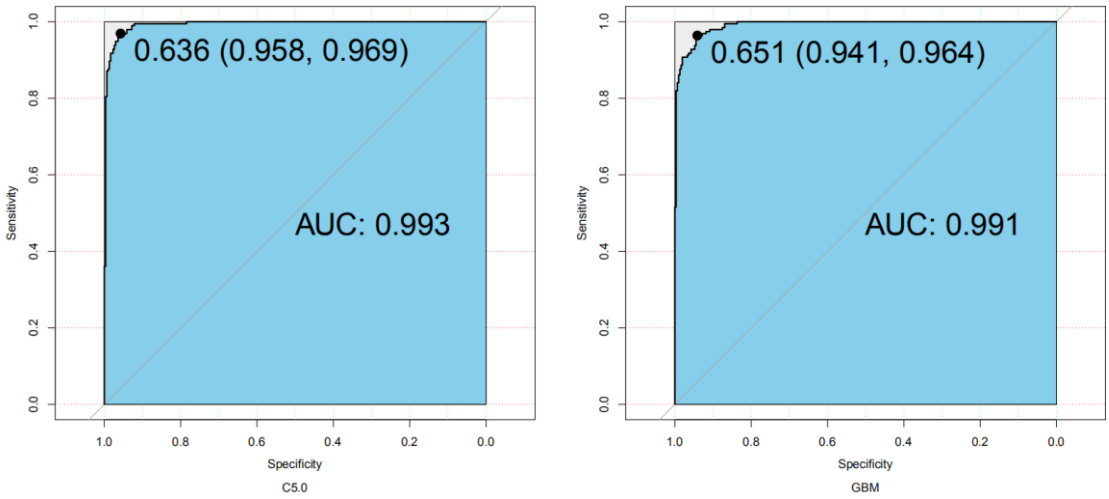

Figure 1. The ROC curves of C5.0 and GBM

From Table 1, we can see that under the Boosting algorithm, the accuracy rate and Kappa value of the C5.0 model are higher than the GBM model's. Figure 1 shows that the AUC value of the C5.0 model is 0.993, the AUC value of the GBM model is 0.991. Therefore, the C5.0 model is superior to the GBM model.

2.2.2. Bagging algorithm. Bagging algorithm is to generate data sets through replacement sampling, and train classifiers using these data sets separately. Then we combine them with the same weight to obtain the final classifier.

Under the Bagging algorithm, we select Bagged CART model and Random Forest model. Then we use ten-fold cross-validation on the two models separately to obtain the classification effect. Table 2 shows the accuracy rate and Kappa values of the two models on the test sets. Figure 2 shows the ROC curves of the two models.

Table 2. The ten-fold cross-validation results of CART and RF

	model	Accuracy	Kappa
Bagging	Bagged CART	0.9341	0.8592
	Random Forest	0.9409	0.8739

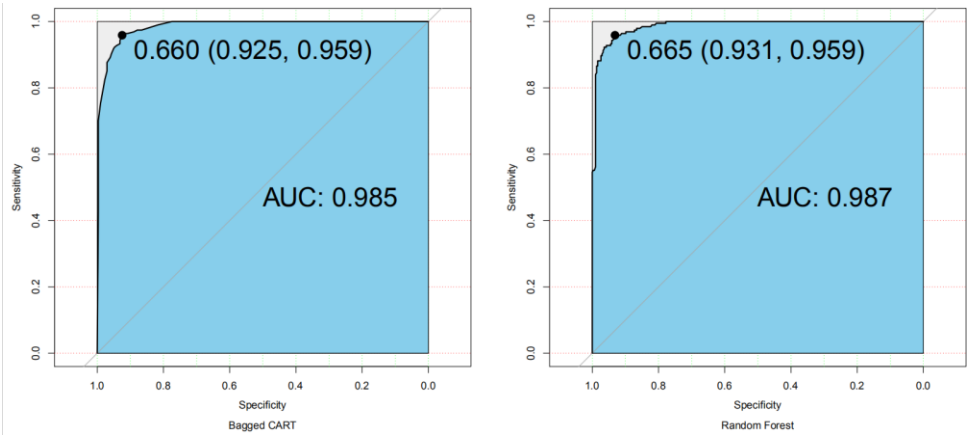

Figure 2. The ROC curves of CART and RF

From Table 2, we can see that under the Bagging algorithm, the accuracy rate and Kappa value of Random Forest are higher than Bagged CART's. Figure 2 shows that the AUC value of the Random Forest model is 0.987, the AUC value of Bagged CART is 0.985. Therefore, Random Forest has a better classification performance.

2.2.3. Stacking algorithm. Stacking is to train a multi-layer model structure. The first layer (learning layer) contains different models, and the obtained prediction results are the input of the next layer model. The next layer model is trained again according to the data label to obtain a complete frame.

Under the Stacking algorithm, we select Stack.glm model and Stack.rf model. Then we use ten-fold cross-validation on the two models separately to obtain the classification effect. Table 3 shows the accuracy rate and Kappa values of the two models on the test sets. Figure 3 shows the ROC curves of the two models.

Table 3. The ten-fold cross-validation results of Stack.glm and Stack.rf

	model	Accuracy	Kappa
Stacking	Stack.glm	0.8830	0.7468
	Stack.rf	0.8664	0.7112

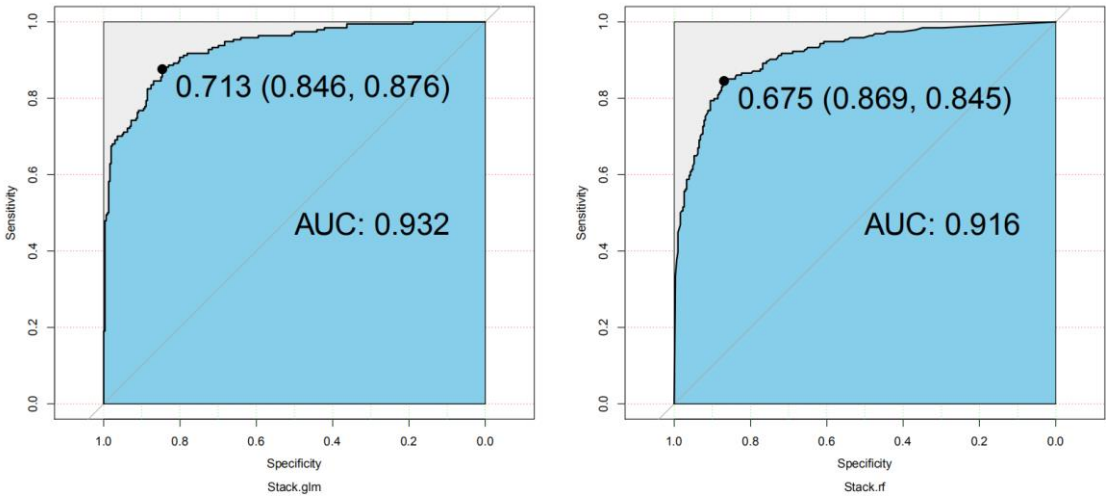

Figure 3. The ROC curves of Stacking.glm and Stacking.rf

From Table 3, we can see that under the Stacking algorithm, the accuracy rate and Kappa value of Stack.glm are higher than Stack.rf. Figure 3 shows that the AUC value of Stack.glm is 0.932, the AUC value of Stack.rf is 0.916. Therefore, Stack.glm is better than Stack.rf.

3. Conclusion

From the above analysis, we can conclude that among the six models under the three ensemble learning algorithms, C5.0 model based on Boosting algorithm has the best classification effect on the test sets. The accuracy rate, Kappa value and AUC value are the highest among six models. Therefore, we select the C5.0 model as the classifier to classify whether customers will order financial products. The classification results can provide a certain reference for the bank to formulate direct marketing plans.

Acknowledgments

This article is supported by National Social Science Foundation of China (No. 17BTJ027), Graduate Research Innovation Fund of Zhejiang Gongshang University, first-class disciplines of Zhejiang Province, characteristic disciplines of Zhejiang Province and Statistics of Zhejiang Gongshang university.

References

[1] Elsalamony, H. A., & Elsayad, A. M. (2013). Bank direct marketing based on neural network and C5. 0 Models. International Journal of Engineering and Advanced Technology (IJEAT), 2(6).

[2] Sing'oei, L., & Wang, J. (2013). Data mining framework for direct marketing: A case study of bank marketing. International Journal of Computer Science Issues (IJCSI), 10(2 Part 2), 198.

[3] Wisaeng, K. (2013). A comparison of different classification techniques for bank direct marketing. International Journal of Soft Computing and Engineering (IJSCE), 3(4), 116-119.

[4] Apampa, O. (2016). Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction. Journal of International Technology & Information Management, 25(4), 85-100.
[5] Marinakos, G., & Daskalaki, S. (2017). Imbalanced customer classification for bank direct marketing. Journal of Marketing Analytics, 5(1), 14-30.
[6] de Menezes, F. S., Liska, G. R., Cirillo, M. A., & Vivanco, M. J. F. (2017). Data classification with binary response through the Boosting algorithm and logistic regression. Expert Systems with Applications, 69, 62-73.
[7] Roshan, S. E., & Asadi, S. (2020). Improvement of Bagging performance for classification of imbalanced datasets using evolutionary multi-objective optimization. Engineering Applications of Artificial Intelligence, 87, N.PAG.
[8] Shi, F., Liu, Y., Liu, Z., Li, E., Li, C., & de Oliveira, J. V. (2018). Prediction of pipe performance with stacking ensemble learning based approaches. Journal of Intelligent & Fuzzy Systems, 34(6), 3845-3855.