Prediction and Feature Importance Investigation for Bank Churn Based on Machine Learning

Yitong Zhan

Department of Accounting, Central University of Finance and Economics, Beijing, China

2021310688@email.cufe.edu.cn

Abstract. Since bank customers are one of the main sources of bank income, preventing the loss of bank customers has always been the primary and struggle problem for banks. This paper chooses three machine learning methods, random forest, decision tree and logistic regression should focus to predict the leaving customers. In order to more accurately determine the factors affecting the departure of bank customers, this paper grouped the data set according to the age of over and under 40. The results shows that the prediction performance of random forest is the best one in both groups, and the logistic regression is the worst one. The precision of this model is higher in younger group than in older group, the accuracy in each group is about 90% and 76% respectively. Then the random forest method is used to return the important features for two groups. For people older than 40 years old, whether to continue to stay in the bank to buy its products is greatly affected by their Balance and Age factors. Having more balance and being younger, the more possibility to keep purchasing. While for under 40 years old customers, their counterpart behaviors are more determined by the Estimated Salary and Credit Score. Thus, when banks managers tackle customer management, they should focus more on the above factors to better prevent the loss of customers.

Keywords: Bank churn, random forest, decision tree, logistic regression.

1. Introduction

The bank churn means that the customers of a bank stop buying financial products from the bank. Most of the bank churn situation is a time when bank's clients close their accounts or discontinue purchasing a particular bank service, including customers who have left recently and will leave. Bank churn can cause banks to lose customers, reduce revenue, affect brand reputation, and even lead to bankruptcy. As an increasing number of banking products emerge, banks face a higher risk of losing their customers to competitors. Therefore, in order to survive in such a competitive market, it is important to predict the bank churn and investigate the most influential factor, helping the bank take measures to decrease or prevent the loss.

Over recent years, machine learning technology has experienced rapid advancement and has effectively integrated into various industries and application domains [1-5]. For instance, Qiu et al. proposed a deep learning-based framework for effective evaluation of patients' rehabilitation training [1]. Sun et al. developed a reinforcement learning based on anchor graph hashing for autonomous cell activation in cloud-RANs [2]. Wu et al. consider the application of deep learning models in sentiment analysis [3]. In the early research of predicting churn, machine learning algorithms are also considered. Most researchers found the random forest was a powerful one. For example, Rahman et al. [6] used K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Decision Tree, and Random Forest classifiers to analyze bank churn data set and found the use of the Random Forest model after oversampling was better compared to other models in terms of accuracy. Besides, Vafeiadis et al. [7] analyzed a problem of telecommunications' customer churning prediction, which is similar to the bank churning. They found the clear superiority of the boosted versions of the models against the plain (non-boosted) versions. And the best classifier was the SVM-POLY using AdaBoost with accuracy of almost 97% and F-measure over 84%. Tran et al. [8] applied k-nearest neighbors, logistic regression, decision tree, random forest, and found the random forest was the best to predict customer churn, with an accuracy of about 97%, and the logistic regression was the least effective model with the lowest accuracy (87.27%). Teemu [9] also tried to predict a bank customer's churn or

no-churn decision by utilizing logistic regression based on personal retail banking companies' data. Moreover. Inkumsah [10] used descriptive research to detect the factors that impacted consumers' decision to repeat purchase product from same bank or not based on their current banks in the UPSA and found customer satisfaction and corporate image are important factors. Because, there are not many researcher pay more attention on find the most influential factor of bank churn, and are not divide the data set into different group to identity different results. Thus this paper will separate the data set into two groups according to 'age', and analyze the different performances and results between different groups.

In this regard, three machine learning methods, random forest, decision tree and logistic regression classifier, are chosen to predict the bank churn and the prediction performance of them on customer churn in two sets of data is analyzed and compared. The results show that the random forest is the most powerful method to predict bank churn for my data set, and the performance in group less than 40 years old is better than another. Finally, the feature ---- 'Balance' and 'CreditScore' are both important factors in two groups. Thus, the results of this paper could help banks adjust their future strategies to reduce their churn rate and become more competitive among others.

2. Method

2.1. Dataset Preparation

The source of the dataset used in this study is from Kaggle [11]. The original data set consisted of 10,002 data and 13 features such as gender, geography, age, balance. The purpose is to predict customers whether will exit by using this data set which is a binary classification problem.

The notation for features of this data set is as followed: 'CreditScore' is A numerical value representing the customer's credit score; 'EstimatedSalary' is the estimated salary of the customer; 'Balance' is the customer's account balance and 'Exited' is whether the customer has churned (1 = yes, 0 = no) et al.

The preprocessing consists of five parts. First, some useless features including customer id and surname were dropped. Then this paper checked the missing values in the dataset and found that only four variables had one missing values. Since one missing value can cause very minimal impact compared to 10002 data, the missing values were just deleted. Third, the article visualized the bank churn distribution for the overall data, as well as the churn distribution of variables such as Age, Nationality, and Gender. As Fig. 1 and Fig. 2 shows most of customers in this data set not exited and 50-60 years old people, female and Germany are highly possible to exit. The next step was to convert text type variables such as Gender and Geography into dummy variables. Finally, the data set was divided into two groups with 40 years as the dividing line.

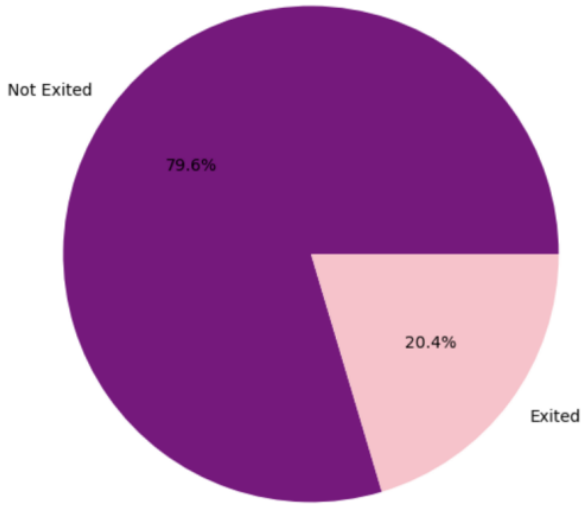

Figure 1. Exited proportion (Picture credit: Original)

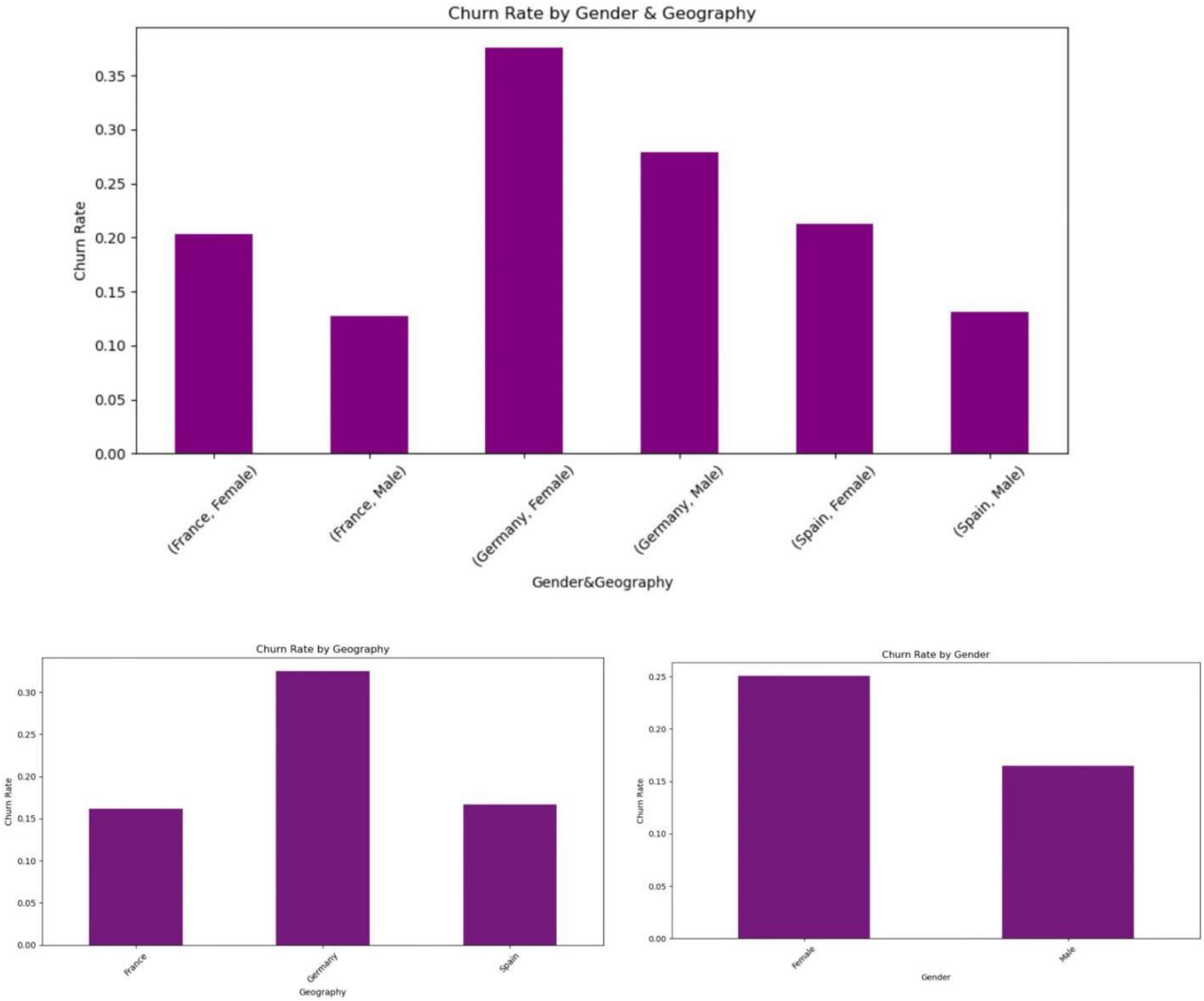

Figure 2. Exited ratio by age gender and geography (Picture credit: Original)

2.2. Machine Learning Models

This paper used logistic regression, decision tree and random forest classifiers to predict the bank churn on two groups and returned the important features. Evaluation of the prediction methods are based on the calculation of precision, recall and F1 score and the Receiver Operating Characteristic (ROC)and its Area Under Curve (AUC) Score are also used for this evaluation. More details are as followed:

2.2.1 Logistic regression

Logistic regression is a simple but effective machine learning algorithm. It is mainly used to solve classification problems, such as predicting whether something will happen or not, usually represented by 0 or 1. Specifically, logistic regression adds up all the factors that might affect the occurrence of an event and puts them into a special function called a “logistic function” or “Sigmoid function.” The output of this function ranges from 0 to 1 and represents the probability of the event occurring.

The logistic function Is showed as follows:

$$P(x) = 1 + e^{-z} \tag{1}$$

Where p(x) is the probability of the event occurring given the values of predictor variables x, and z is the linear function of predictor variables and their coefficients:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{2}$$

Here, β_0, β_1, $\beta_2 \ldots \beta_n$ are the coefficients estimated by the model, and x_1, x_2, $x_3 \ldots x_n$ are the values of predictor variables.

2.2.2 Decision Tree

Decision tree is a very frequently used machine learning method, which is useful to establish a prediction model, including linear prediction or binary classify associated prediction. It is a vividly tree-like model and shows a mapping relationship between its object attributes and its counterpart values. An object acts as a node of a tree, and then divides into different branches according to the attribute values of this object, and each leaf of the tree extends the value of the object from the root node.

The decision tree algorithm will select the best feature to split the data at each branch node, in order to maximize the homogeneity of the resulting subsets regarding the class labels. This recursive process repeats until certain stopping conditions are met, such as maximum depth.

2.2.3 Random Forest

Random forest is an algorithm of ensemble learning and can be visualized as a composition of many decision trees, which can better improve the accuracy of prediction model and reduce overfitting. It can assess the importance of variables when determining categories and are good at efficiently processing a large number of features and samples without excessive preprocessing or feature selection for features. It can handle data sets with hundreds or even thousands of features without compromising performance.

3. Results and Discussion

3.1. Model Performance

3.1.1 Under 40 Years Old Group Performance

The performance of three machine learning methods in under 40 years old group demonstrated that the accuracy for random forest is the highest at 90% approximately shown in Table 1 and Fig. 3. While the AUC Score for it is 0.58 and for decision tree is 0.59, the precision, recall and f1-score for random forest are 0.91,0.99,0.95 respectively, which are all higher than those for the rest of methods.

Table 1. Model Performance in under 40 years old group

	Accuracy		Precision	Recall	F1-score	Support
Random Forest	0.9034268	0	0.91	0.99	0.95	1149
		1	0.65	0.18	0.28	135
Decision Tree	0.8504673	0	0.91	0.92	0.92	1149
		1	0.28	0.27	0.27	135
Logistic Regression	0.89	0	0.89	1.00	0.94	1149
		1	0.00	0.00	0.00	135

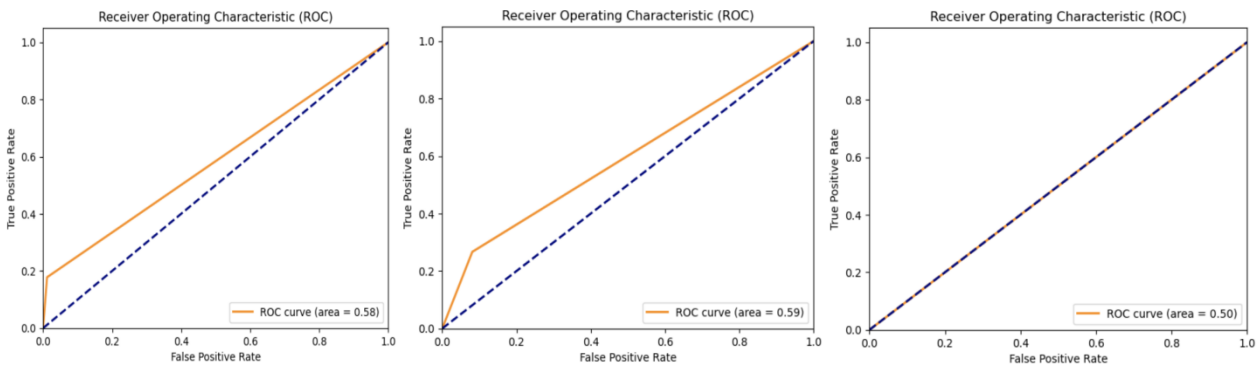

Figure 3. AUC Score in under 40 years old group (Picture credit: Original)

3.1.2 Older Than 40 Years Old Group Performance

The performance of three machine learning methods in older than 40 years old group shows that the accuracy for random forest is the highest at 76% approximately shown in Table 2 and Fig. 4. The AUC Score, precision, recall and f1-score for random forest are 0.74, 0.77,0.85 and 0.81 respectively, which are all higher than those for the rest of methods.

Table 2. Model Performance in older than 40 years old group

	Accuracy		Precision	Recall	F1-score	Support
Random Forest	0.7597765	0	0.77	0.85	0.81	430
		1	0.73	0.63	0.68	285
Decision Tree	0.7025139	0	0.76	0.74	0.75	430
		1	0.62	0.64	0.63	285
Logistic Regression	0.64	0	0.64	0.91	0.75	430
		1	0.64	0.24	0.35	285

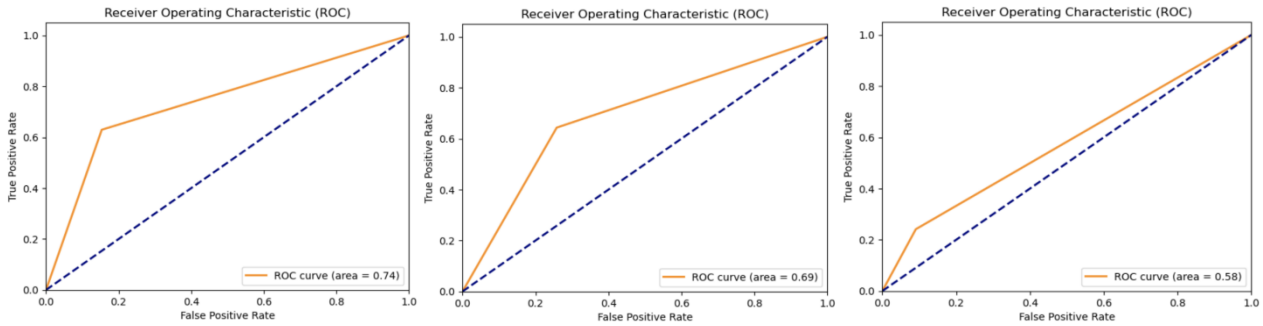

Figure 4. AUC Score in older than 40 years old group (Photo/Picture credit: Original)

According to the results, the random forest classifier is the best model in both groups and the performance of same model in the younger is better than the older one. It may be because different age groups may have different importance for different characteristics. There are some distinct behavioral patterns or characteristics in the younger age group that make it easier for the model to pick up on their exit trends. In the group older than 40, there may be more individual differences, which are less related with the features in this data set, making it difficult for the model to predict accurately. The reason why random forest is the best method maybe its ensemble learning advantage, the random forest model simultaneously trains multiple decision trees and combines their results for predictions. This ensemble approach reduces overfitting when handling a large number of features and samples, enhances the model's generalization capability, and consequently performs better on unseen data. Additionally, bank customer churn problems may involve various nonlinear relationships, and decision tree-based methods such as random forests, decision tree can effectively capture these nonlinear relationships. In contrast, logistic regression is more suitable for problems where linear relationships are more evident.

3.2. Feature Importance

The feature importance for both two groups shown in Table 3 shows that EstimatedSalary, CreditScore and Balance are the top three correlated features in predicting younger people exit behavior. The CreditScore reflects the customer's credit status and repayment ability. Thus, customers with higher credit scores are more likely to be reliable bank customers who tend to pay on time and maintain a good credit history, so the risk of churn is lower. Because young people's economic strength and resources are limited, many financial behaviors are greatly affected by income. The expected income is the most influential factor, and the higher the expected income, the more strength and possibility of the customer to continue to buy products in the bank.

Table 3. Feature importance in under 40 years old group

Feature	Importance
EstimatedSalary	0.209569
CreditScore	0.198769
Balance	0.186708
Age	0.124197
NumOfProducts	0.115780
Tenure	0.107047
Geography_Germany	0.017243
Geography_France	0.010928
Gender_Female	0.010383
Geography_Spain	0.010109
Gender_Male	0.009268

The top three features in the older than 40 years old group shown in Table 4 are Balance, Age and CreditScore. While the most influential feature for older people is 'Balance' which is the third influential factor for younger group. The reason for this is possible that for older people, the economic source is mainly the balance saved in the first half of life, and there is basically no new large amount income source at this period, so at this stage, the amount of their own savings will greatly affect whether they will continue to buy products in the bank. Besides, for those people, the age factor plays more roles than for youngers, they are less likely to be marketed to the corresponding bank products or are less adept and willing to continue to manage money in the bank. As a result, the more they get older the more they leave.

Table 4. Feature importance in older than 40 years old group

Feature	Importance
Balance	0.175592
Age	0.170688
CreditScore	0.169605
EstimatedSalary	0.167346
NumOfProducts	0.149397
Tenure	0.093809
Geography_Germany	0.027664
Geography_Spain	0.012050
Gender_Female	0.011614
Geography_France	0.011205
Gender_Male	0.011029

4. Conclusion

In this work, the purpose is to find a way to predict bank churn and detect the important features, in order to help bank to identify potential losing customers and retain existing customers. Three machine learning methods, including random forest, decision tree and logistic regression classifier, are used in this paper to predict bank churn. The results show that the random forest method has the best performance among others in both age groups, and the performance in younger group is better than the older. Besides, Estimated Salary is the most influential factor affecting whether banks lose young customers, while the balance is the most powerful factor for older people. In the future, some other algorithms, such as neural networks, deep learning and other methods will be applied to the research of this problem to see whether there is a better prediction method.

References

[1] Qiu Y, Wang J, Jin Z, Chen H, Zhang M, Guo L. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control. 2022 Feb 1; 72: 103323.

[2] Sun G, Zhan T, Owusu BG, Daniel AM, Liu G, Jiang W. Revised reinforcement learning based on anchor graph hashing for autonomous cell activation in cloud-RANs. Future Generation Computer Systems. 2020 Mar 1; 104: 60-73.

[3] Wu Y, Jin Z, Shi C, Liang P, Zhan T. Research on the Application of Deep Learning-based BERT Model in Sentiment Analysis. arXiv preprint arXiv:2403.08217. 2024 Mar 13.

[4] Zhou Y, Osman A, Willms M, Kunz A, Philipp S, Blatt J, Eul S. Semantic Wireframe Detection, 2023.

[5] Wang H, Zhou Y, Perez E, Roemer F. Jointly Learning Selection Matrices for Transmitters, Receivers and Fourier Coefficients in Multichannel Imaging. arXiv preprint arXiv:2402.19023. 2024 Feb 29.

[6] Rahman M, Kumar V. Machine learning based customer churn prediction in banking. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020.

[7] Vafeiadis T, et al. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 2015, 55: 1-9.

[8] Tran H, Le N, Nguyen V-H. CUSTOMER CHURN PREDICTION IN THE BANKING SECTOR USING MACHINE LEARNING-BASED CLASSIFICATION MODELS. Interdisciplinary Journal of Information, Knowledge & Management, 2023, 18.

[9] Mutanen T. Customer churn analysis–a case study. Journal of Product and Brand Management, 2006, 14 (1): 4-13.

[10] Inkumsah WA. Factors That Impacted Customer Retention of Banks. A Study of Recently Acquired Banks in the UPSA Area of Madina, Accra (Specifically Access Bank). Journal of Marketing and Consumer Research, 2013, 1.88: 103.

[11] Kaggle. Binary Classification with a Bank Churn Dataset. Available at: https://www.kaggle.com/competitions/playground-series-s4e1, 2024.