Prediction for the Customer Behavior in the Financial Institution Based on the Machine Learning Algorithms

Hongyu Deng [1, *]

[1]Department of Economics, Boston University, MA 02215, United States

*Corresponding author: joeydeng@bu.edu

Abstract. Investigation for the customer behavior is an emerging field that many studies fouces on in the past. In the field of the bank, predicting whether the customr will leave or not is an important question. This study mainly tries to figure out what is the most influencing factor of a customer when they choose to exit tone financial institution. It is important that the company get the rational prediction, which could help them to make the right decosison to maintain and retain the customers. We use the random forst and Neuralnetwork algorithm to do the process. The factor age is the most influencing factor which world affect the exiting decision. This paper firstly predicted that the factor age is the most influencing factor, since the age is a factor which could make the customer change their decision based on their experiences and savings. Generally speaking, the results fit the prediction in this study. The results of this work is that the factor age is the most influencing one, and the factor gender matters the least. Other factors such as CreditScore and Salary are also play an important role when the customers make the exiting decision. Besides, there is also an interesting finding. The German customers seem to be more possible to exit one financial institution.

Keywords: Customer behavior, neural network, feature importance, random forest.

1. Introduction

In the banking industry, there is no shortage of options for customers when choosing where to put their money. As a result, the data analyzer needs to figure out what are the key factors which cause the customers leave. With increased competition from both existing players who are revamping their digital offering and newcomers threatening to disrupt the industry, customer experience is an increasingly important battleground for financial institutions, which means that the prediction based on the customers data could help the financial institution make the right decision on two things. Firstly, the rational prediction could help the financial institution find the possibility of whether one customer leaves. Secondly, on account of the former reason, the institution could edit the current policies to persuade the customers to stay rather than leaving.

In the last two decades, many studies have been focused on the predicting customer behavior based on machine learning algortihms [1-5]. For instance, Ezenkwu et al. employed the K-Means algorithm to do the customer segmetation to find the different properties of customers in the company [1]. Qiu et al. employed a optimized K-Means algorithm called K-Means++ to segment the customer in the field of telecommunication. The result suggests that which part of the customer is the silent customer, that is, easy to be lost [2]. Therefore, the company can consider making some proper decisions to win back this segment of customers. In addition, some other studies [3-5] also considered how to use machine learning algortithms to predict the behavior of the customer. However, a lot of studies mainly focus on the customer related to the field of telecommunication or similar fields, little attention is paid to the analysis for the bank customer. In addition, these studies mentioned above mainly use unsupervised learning algorithm to the prediction. Supervised learning algorithms are rarely considered in this case. Therefore, these limitations should be addressed properly.

In this study, two typical machine learning algorithms called the random forest [6-8] and neural network [9-11] was chose in this study in finding the relationship between properties of bank customers and their corresponding behavior. Each of the methods will come out the scores the data got during the processing. Then, this study can further do the comparison between the results of these two methods, which could help us make a rational prediction.

2. Datasets and Methods

2.1 Data sources

The data is mainly coming from a website called Kaggle [12], which contains the information of all different kinds of banking customers. The information in this dataset contains the CustomerId, the Surname, the CreditScore, the Geography, the Gender, the Age, the Tenure, Balance, the NumOfProducts, the HasCrCard, whether the customer is an active member (marked by 0 or 1), the EstimatedSalary, and whether such customer exits such financial institution. All of these elements can influence the customer's decision. The sample data for this dataset can be found in the Table 1.

Table 1. Sample data in the collected dataset.

RowNumber	1	2	3
CustomerId	15634602	15647311	15619304
Surname	Hargrave	Hill	Onio
CreditScore	619	608	502
Geography	France	Spain	France
Gender	Female	Female	Female
Age	42	41	42
Tenure	2	1	8
Balance	0.00	83807.86	159660.80
NumOfProducts	1	1	3
HasCrCard	1	0	1
IsActiveMember	1	1	0
EstimatedSalary	101348.88	112542.58	113931.57
Exited	1	0	1

However, these categories need to be edited since some of those cannot be utilized in the following progress, and they do not influence the customer to make the existence decision.as well, such as category CustomerId, Surname and CreditScore. Then, for the process be easier, we need to transfer the column Geography and Gender from strings to numbers. This study also need to figure out the 'y' and 'x'. Y represents the column exited, and the x is the function of y, which means that how the other factors, such as Geography and Gender, influence the exiting decision. The data scale of this learning is contained 10000 pieces of all kinds of customers and thirteen different factors. As a result, we need to allocate twenty present of data for the test set and eighty present of data for the training set.

2.2 Random forest

Firstly, random forest as a famous ensemble machine learning was employed in this study to find the feature importance and predict for this dataset. Actually, random forest can complete both classification and regression tasks. But in this case, the random forest was only used to achieve the classification function since the aim of this dataset is to predict whether the customer will leave this bank. In addition, this paper let the Characteristic score be 100, which is the scale of the Decision tree.

2.3 Artificial Neural network

The second method this paper chose is neural network algorithm. Neural network is commonly used in different domains in the last decade due to the rapid development of the machine learning. In general, neural network consists of three parts, namely the input layer, hidden layer and output layer. Additionally, activation function that can provide the non-linear ability is also usually added to the hidden layers. In this study, the architecture of the neural network can be found in the Table 2.

Table 2. The architecture of the designed neural network.

Layer	Out shape	Parameters
Dense	(None, 100)	1400
Dense	(None, 50)	5050
Dense	(None, 25)	1275
Dense	(None, 1)	26

2.4 Implementation details

Some hyperparamters (i.e. learning rate, loss function, optimizer and metrics) are set to 0.0001, binary crossentropy, Adam and accuracy, respectively, before training the neural network. Additionall, the epochs are set to 30 during the whole training process. After training the neural network, then the testing dataset was employed to test the generalizatino ability of the neural network.

3. Results and Discussion

3.1 Feature importance

To observe the effects of different features on user behavior, the random forest was firstly trained on the training dataset. Next, the importance score of features are presented as shown in Table 3 and Figure 1. It can be observed that age is the most important factor that determines whether the customer leave the bank or not. In addition, four features namely EstimatedSalary, CreditScore, Balance and NumOfProducts are also important in this case.

From this process, it could be observed that the factor age is the most influencing reason, which could be rather rational, since the characteristics of people in different age change. Most of the young customers are more willing to take the risk, or choose to maintain their decision when the policy is not that profitable than the people older do, since the youth could wait until the policy is profitable again. However, the willingness of middle-aged people to take the financial risk decreases because of the debt. Loads of middle-aged people are not able to bear risks, which means that whenever the financial institutions' policies cannot be as profitable as they predict, they will make the exiting decision immediately. The willingness of taking risks is so different, which could be vital enough to be the most influencing factor when customers make the decision.

Table 3. The specific important score of features.

Feature name	**Score**
Age	0.24100931348849036
EstimatedSalary	0.14679552365547746
CreditScore	0.14584568078335533
Balance	0.14200731032395986
NumOfProducts	0.12691740731438017
Tenure	0.08352570491515163
IsActiveMember	0.039786727772151385
Geography_France	0.021128470231091583
HasCrCard	0.017764617494203785
Exited	0.009716844643706963
Geography_Spain	0.008851621966405083
Geography_Germany	0.008582747483032638
Gender_Female	0.00806802992859383

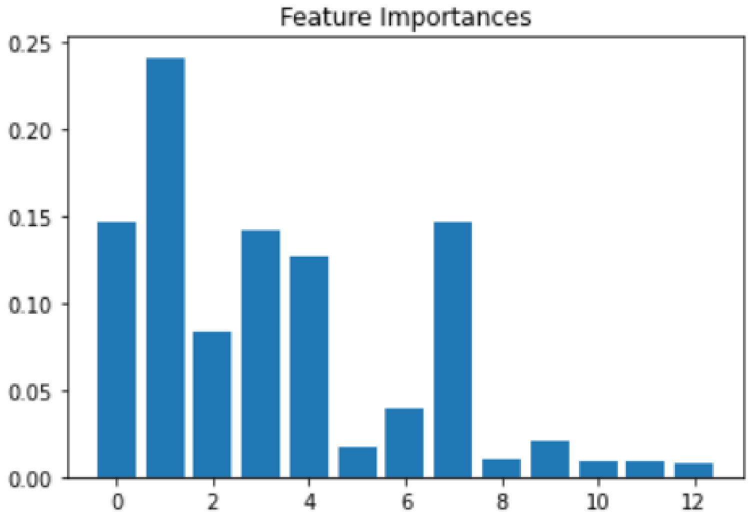

Figure.1 The feature importance based on the random forest.

3.2 Performance for the predicition

Apart from the importance score, the other task in this study is to predict the customer behevior (i.e. leave or do not leave). Random forest and artificial neural network mentioned above are employed here to predict this dataset. The experimental results for these two algorithms can be observsed in Table 4.

Table 4. Accuracy based on the random forest and artificial neural network.

Model Name	**Accuracy**
Random forest	0.8685
Artificial neural network	0.7950

Table 4 indicates that the performance of random forest is more satisfactory compared to the artificial neural network. The main reason may be that the neural network is easy to be overfitting due to the a large number of training parameters in the network.

4. Conclusions

This study aims to figure out what kind of customers are more likely to exit one financial institution for making a prediction, which could help such institution make the former edition to prevent customer churn. The random forest and neural network algorithm was used to get the prediction. The results of this work is that the factor age is the most influencing one, and the factor gender matters the least. Besides, the German customers seem to be more possible to exit one financial institution. However, only two machine learning algorithms are considered in this study and experimental results have a space to be further improved. In the future, more advanced algorithms will be also used to improve the prediction performance

References

[1] Ezenkwu, Chinedu Pascal, Simeon Ozuomba, and Constance Kalu. Application of K-Means algorithm for efficient customer segmentation: a strategy for targeted customer services [J]. (2015).

[2] Qiu, Yuhang, et al. Clustering Analysis for Silent Telecom Customers Based on K-means++ [C]. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Vol. 1. IEEE, 2020.

[3] Syakur, M. A., et al. Integration k-means clustering method and elbow method for identification of the best customer profile cluster [C]. IOP conference series: materials science and engineering. Vol. 336. No. 1. IOP Publishing, 2018.

[4] Kansal, Tushar, et al. Customer segmentation using K-means clustering [C]. 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS). IEEE, 2018.

[5] Zare, Hamed, and Sima Emadi. Determination of Customer Satisfaction using Improved K-means algorithm [J]. Soft Computing 24.22 (2020): 16947-16965.

[6] Oshiro, Thais Mayumi, Pedro Santoro Perez, and José Augusto Baranauskas. How many trees in a random forest? [C]. International workshop on machine learning and data mining in pattern recognition. Springer, Berlin, Heidelberg, 2012.

[7] Probst, Philipp, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest [J]. Wiley Interdisciplinary Reviews: data mining and knowledge discovery 9.3 (2019): e1301.

[8] Probst, Philipp, and Anne-Laure Boulesteix. To tune or not to tune the number of trees in random forest [J]. J. Mach. Learn. Res. 18.1 (2017): 6673-6690.

[9] Abiodun, Oludare Isaac, et al. State-of-the-art in artificial neural network applications: A survey [J]. Heliyon 4.11 (2018): e00938.

[10] Wang, Sun-Chong. Artificial neural network [J]. Interdisciplinary computing in java programming. Springer, Boston, MA, 2003. 81-100.

[11] Hecht-Nielsen, Robert. Theory of the backpropagation neural network [J]. Neural networks for perception. Academic Press, 1992. 65-93.

[12] Kaggle, bank customer churn prediction, https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction, 2022