Machine Learning Algorithms Based Prediction for Customer Churn in Banks

Chenyang Yang

Department of Data science and Big Data technology, Chendu University, Chendu, China

yangchenayang@stu.cdu.edu.cn

Abstract. Currently, competition within the service sector is intensifying rapidly. Acquiring new customers is becoming increasingly costly for banks compared to retaining existing ones. Consequently, it becomes imperative to forecast customer attrition in banks. This paper presents a method to predict customer churn in banks using machine learning. The study explored the possibility of customer churn in banks through a dataset of customer information on Kaggle. This study uses four kinds of machine learning such as ANN, Random Forest, Decision tree and Logistic regression. And the correlation analysis of the data set is carried out. Finally, the ROC curves of four kinds of machine learning were compared to get the model with the highest accuracy. Therefore, the accuracy of ANN is higher than that of the other three models. Then the importance is analyzed, and the most important feature of this experiment is the number of banking products owned by customers.

Keywords: Machine Learning, ANN, Random Forest, Decision Tree, Logistic Regression.

1. Introduction

In recent years, with the rise of the Internet, the market has become more and more dynamic, and the competition has become more and more fierce. This is because there are a large number of service providers. In the course of their competition, customers will be lost due to changing customer behavior and customer expectations. Also because of the internet, customers can more easily consult different service providers, and it is increasingly easy to switch to other service providers for better service. For the banking industry in particular, customer retention is critical to their success. The reason for this is because clients who remain loyal to the bank for an extended period of time are the most lucrative. Therefore, comprehending and averting customer attrition is a crucial aspect of a bank's triumph. However, even the most prosperous banks encounter client attrition. Indeed, it is observed that the typical bank or financial institution experiences a customer attrition rate ranging from 25 to 30 percent annually. Reichheld et al point out that reducing customer churn by 5% May increase banks' profits from these customers by 85% [1], while many studies have shown that preventing customer churn can save money because it can cost five times as much to acquire new customers as it does to satisfy and retain existing customers [2]. So it is crucial for banks to comprehend the reasons behind client attrition and devise strategies to mitigate it.

Bank churn analysis is a very broad field. One of the studies, some studies used k-Nearest Neighbors (KNN), Support Vector Machine (SVM), decision tree, and random forest classifiers to predict the probability of bank customer churn [3, 4]. In addition, this paper also proposes some feature selection methods to identify more relevant features and verify the performance of the system. Finally, it is concluded that the random forest model after oversampling is superior in accuracy. However, the dataset selected for this post has only 10,000 entries, which may result in the machine learning model not being trained adequately. This data set is also pre-2020 data set. However, after 4 years of development, the data and machine learning models have been greatly updated.

In another study [5], several supervised classification algorithms such as KNN, logistic regression, AdaBoost, Gradient Boosting, and random forest were applied. To achieve higher prediction accuracy, these classifier ensembles are voted uniformly. Through hyperparameter optimization and tuning, the model performance is improved. The most important features are defined and ranked using a random forest model. In the experimental case of 10K dataset, the model achieves 87% accuracy.

The study examined the prediction of customer turnover in the banking industry utilizing a distinct dataset at the customer level obtained from a major Brazilian bank [6]. Their primary contribution lies in the examination of this extensive dataset, encompassing past customer behavior traits, enabling them to record novel insights on the primary factors influencing the prediction of future customer churn.

In this research [7], a model including data mining is established by using decision tree technology. It is also concluded that bank managers can use the results of the decision tree to identify future losses.

This study aims to employ machine learning models for analyzing existing bank customer data in order to identify patterns of customer behavior and ascertain the primary factors contributing to customer attrition. The analytical efficacy of four distinct machine learning models, namely Artificial Neural Network (ANN), random forest, decision tree, and logistic regression, is presented in this study. The results are analyzed by comparing the area under the ROC curve between the prediction probability and the observed target. This study determines the model with the highest predictive accuracy and summarizes the key reasons for customer churn within the dataset.

2. Method

2.1. Dataset Preparation

The Kaggle dataset on bank churn from 2024 is used in this article [8]. This dataset contains 165,034 entries of information about bank users, and the goal is to predict whether a customer will continue to use their account or close it (e.g., churn). The target variable reflects a binary flag of 1 when a customer churns and 0 when a customer does not churn. The dataset contains 14 feature vectors. Fig 1 provides the details of these features.

1. Customer ID: A unique identifier for each customer
2. Surname: The customer's surname or last name
3. Credit Score: A numerical value representing the customer's credit score
4. Geography: The country where the customer resides (France, Spain or Germany)
5. Gender: The customer's gender (Male or Female)
6. Age: The customer's age.
7. Tenure: The number of years the customer has been with the bank
8. Balance: The customer's account balance
9. NumOfProducts: The number of bank products the customer uses (e.g., savings account, credit card)
10. HasCrCard: Whether the customer has a credit card (1 = yes, 0 = no)
11. IsActiveMember: Whether the customer is an active member (1 = yes, 0 = no)
12. EstimatedSalary: The estimated salary of the customer
13. Exited: Whether the customer has churned (1 = yes, 0 = no)

Figure 1. The feature description in the dataset used (Photo/Picture credit: Original)

This study first explores that the dataset does not have null values, so no further processing is required. Second, this study discards features that are not relevant to the topic, such as id, CustomerId, Surname. Finally, some more complex features are transformed, so as to facilitate the training model later. Transformed in this study are: Geography——Geography_France, Geography_Germany, Geography_Spain. Gender——Gender_Female, Gender_Male.

Furthermore, the correlation between each feature vector is calculated, as shown in Fig 2. It is concluded that customer churn has a strong positive correlation with age and a strong negative correlation with the number of bank products used by customers.

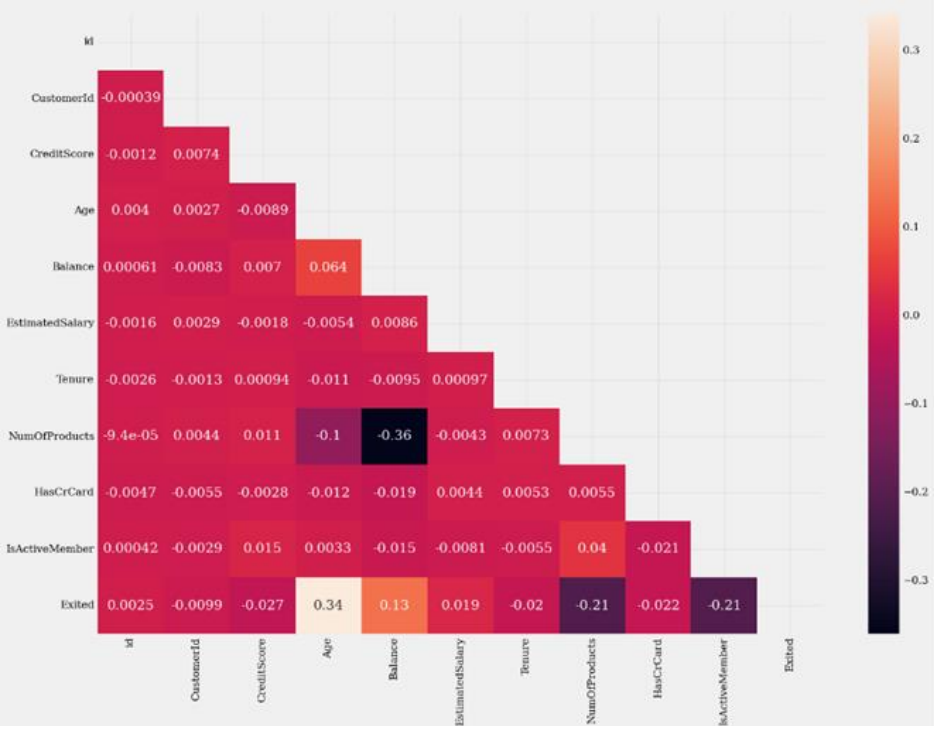


Figure 2. Correlation matrix for features (Photo/Picture credit: Original)

2.2. Machine Learning Models

2.2.1 ANN

Contemporary computers have the capability to do intricate numerical and symbolic calculations at an astonishing rate. However, human brains far surpass them in terms of performance when it comes to carrying out perceptual tasks like language comprehension and image identification. Computers rely on accurate input data and execute instructions in a sequential manner, but human brains complete tasks through a distributed and parallel approach. The development of artificial intelligence based on biological neural networks leads to the creation of an Artificial Neural Network (ANN). When compared to a conventional regression method, the ANN has the ability to accurately represent intricate nonlinear connections. It is a network of interconnected nodes that is similar to the biological neural network, resembling neurons. The three essential components of a neural network are node characteristics, network structure, and learning rules. The node's character governs the processing of signals, including the number of inputs and outputs, the weight assigned to each input and output, and the activation function. Network topology refers to the arrangement and interconnection of nodes inside a network. The learning rules dictate the process of initializing and adjusting the weights [9].

In this paper, a 4-layer network is built with keras. There are 128, 64,32, and one neuron. And the first three layers use the relu function. In the fourth layer, the sigmoid function is used to output whether users are lost. The model's loss function uses binary_crossentropy. The optimizer uses adam.

2.2.2 Random Forest

Random forest is a popular machine learning algorithm that combines the power of multiple decision trees to make accurate predictions. It is an ensemble learning approach that uses decision tree collections to improve overall performance and reduce overfitting. The basic idea of Random Forest is to build a collection of multiple decision trees, and then let these decision trees classify input samples, and determine the final classification result by voting. During the construction of each decision tree, an independent random vector with the same distribution is used to guide the growth of the tree. The workflow of random forest can be summarized as follows: Data sampling: Randomly select a subset of training data replacement (bootstrap sampling). This creates multiple subsets of the

data, each the same size as the original dataset. Feature sampling: Randomly select a feature subset from the feature set. This subset is used to train each decision tree. The number of features selected is usually the square root of the total number of features. Tree construction: For each data subset and feature subset, a decision tree is built using specific algorithms such as Classification and Regression Trees (CART). The data is segmented according to the selected features and their corresponding thresholds, and the tree is grown recursively. Integration creation: Perform iterations of steps 2 and 3 in order to generate a predetermined quantity of decision trees. Every each tree undergoes training using a distinct selection of data and characteristics. Prediction: Each decision tree in the integration should generate predictions for new inputs. The ultimate forecast is ascertained through the consolidation of individual forecasts through voting (in the case of classification) or averaging (in the case of regression). Random forests possess numerous benefits, such as their capacity to handle data with a large number of dimensions, handle missing values, and offer an estimation of the significance of features. However, it has a lower susceptibility to overfitting compared to a solitary decision tree [10]. In this paper, the number of subtrees in the random forest is 100. The maximum selection feature number is sqrt.

2.2.3 Decision Tree

The decision tree approach is a widely used data mining technique that employs several covariates to construct a classification system or to create prediction algorithms for target variables. The proposed approach involves the categorization of populations into distinct segments resembling branches, forming inverted trees consisting of root nodes, inner nodes, and leaf nodes. The algorithm has non-parametric characteristics and demonstrates fast handling of extensive and intricate data sets, without the need for elaborate parameter structures. Once the sample size reaches a sufficient magnitude, it becomes possible to partition the study data into distinct training datasets and validation datasets. The training data set is utilized to construct a decision tree model, while the validation data set is employed to ascertain the optimal tree size required to attain the most effective final model [11]. In this paper, the maximum depth of the decision tree is 4, and min_samples_split is 4. The random state is 42.

2.2.4 Logistic Regression

Logistic regression aims to create a mathematical model that describes the connection between an independent variable and a dependent variable. This model is then used to forecast the probability of the dependent variable. The process of logistic regression consists of the following stages: 1. Data preprocessing: Gather and arrange the data for both the independent and dependent variables. 2. Model building: Choose suitable independent variables and employ logistic regression models to construct mathematical models. 3. Model fitting: Statistical techniques are employed to estimate the parameters of the model in order to optimize the fit between the model and the data. 4. Model evaluation: Assess the quality of the model by examining its level of accuracy in fitting the data and its ability to make accurate predictions. The interpretation of results involves analyzing the parameter estimates and statistical significance of the model to understand how the independent factors affect the dependent variables. Logistic regression is a statistical method that establishes a mathematical model to predict the likelihood of the dependent variable based on the independent variable. The workflow consists of several stages: data preparation, model construction, model fitting, model evaluation, and interpretation of the results [12].

In this article, solver uses saga and elasticnet is selected for penalty. class_weight is balanced. And l1_ratio=0.5, C=1.0, tol=0.001

3. Results and Discussion

3.1. The Performance of Different Models

The results are analyzed by comparing the area under the ROC curve between the prediction probability and the observed target shown in Table 1.

Table 1. The ROC-AUC score

ANN	Random forest	Decision tree	Logistic regression
0.8830439924308268	0.8723369528721169	0.8609298172461637	0.6671946755625617

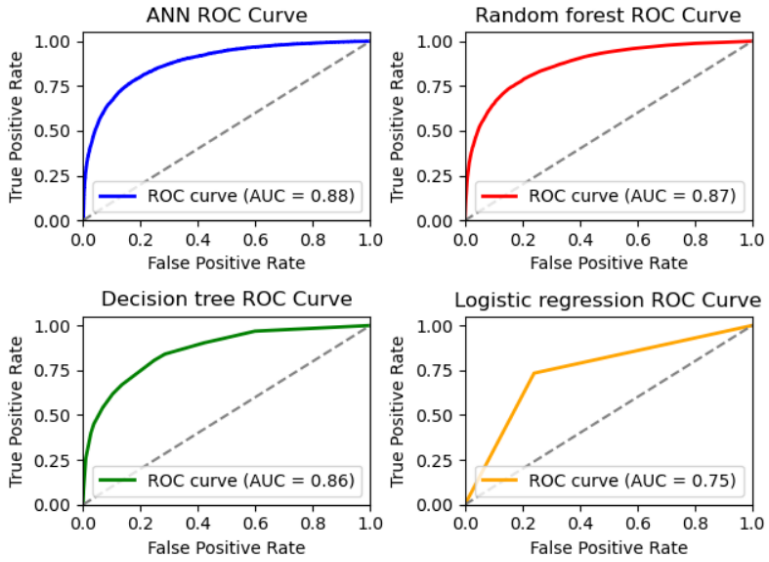


Figure 3. The ROC of different models (Picture credit: Original)

The model was evaluated using the ROC-AUC score. The ROC-AUC metric will evaluate the performance of the four created classifiers on a test dataset that represents the full dataset. The objective is to identify the best model with a fair baseline score, as depicted in Fig. 3. The significance of the ROC-AUC scoring system is in its ability to provide a score that takes into account all potential thresholds, rather than perceiving a single threshold as the average accuracy score. The artificial neural network (ANN) classifier achieved a ROC-AUC score of 0.88 and the decision tree classifier achieved a score of 0.87 on the validation dataset. Other classifiers, such as logistic regression and decision tree, do not perform well on validation sets.

3.2. The Shap Importance of Different Features

This article uses the importance of SHAP to understand which features are most important for prediction shown in Fig 4.

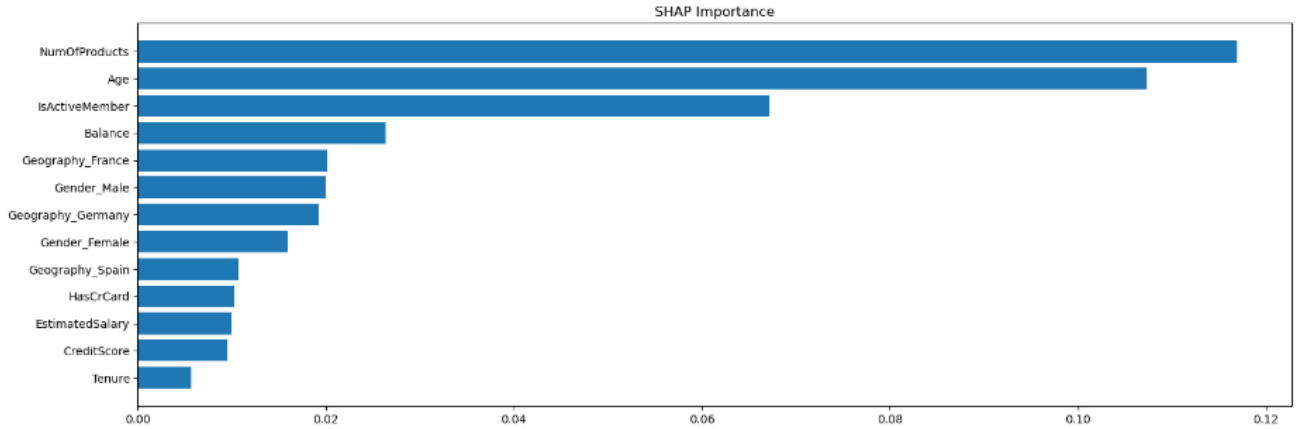


Figure 4. SHAP Importance of different features (Picture credit: Original)

A key finding of the study was the identification of the attributes with the highest ability to predict lead churn. The study found that the number of banking products used by customers (NumOfProducts) had the highest predictive power. So, it is important for banks and financial institutions to understand why NumOfProducts has such a strong influence on customer churn. One possible explanation could be that customers who use multiple banking products have higher expectations and demands from their financial institution. If these expectations are not met or if they find better offers elsewhere, they may choose to switch providers. To address this issue and reduce churn, banks can focus on strengthening their relationship with customers through personalized product offerings and exceptional customer service. By understanding each customer's unique needs and preferences, banks can tailor their product recommendations accordingly. This approach not only increases customer satisfaction but also enhances loyalty, making it less likely for them to consider switching providers.

In addition, t the study shows that customer age is also highly predictive of lead churn. By analyzing the data on customer age, banks can gain valuable insights into their target demographics and tailor their services accordingly. For instance, younger customers may have different financial needs and preferences compared to older customers. Banks can develop personalized products and services that cater to these specific age groups. Moreover, understanding the relationship between customer age and lead churn allows banks to implement targeted marketing strategies. They can create tailored advertising campaigns that resonate with different age segments, effectively reaching out to potential customers who are more likely to be interested in their offerings. This not only increases the chances of acquiring new leads but also enhances overall customer satisfaction.

Surprisingly, the research also shows that the number of years a customer has served in a bank (Tenure) has the worst predictive power. This illustrates that customers with higher years of use do not necessarily have higher retention rates. It also demonstrates the need for banks to strengthen their contact with each and every customer.

4. Conclusion

In this paper, four machine learning models are used to predict the customer churn of banks. Four kinds of machine learning models include ANN, random forest, decision tree, and logistic regression. The final experimental results demonstrate that ANN has the best performance on prediction, and the analysis shows that the number of products owned by bank customers is the most significant feature in this experiment. The parameter setting in machine learning is not very detailed, which may have a slight impact on the result. In the future, I plan to further optimize the machine learning model to achieve higher accuracy.

References

[1] Ascarza E. Retention Futility: Targeting High Risk Customers Might Be Ineffective. Social Science Electronic Publishing, 2016.

[2] Sharma A, Kumar Panigrahi P. A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. International Journal of Computer Applications, 2013.

[3] Hasselmo ME, Schnell E, Barkai E. Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. Journal of Neuroscience the Official Journal of the Society for Neuroscience, 1995.

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

[5] Galal M, Rady S, Aref M. Enhancing Customer Churn Prediction in Digital Banking using Ensemble Modeling. 2022 4th Novel Intelligent and Leading Emerging Sciences Conference (NILES), Giza, Egypt, 2022.

[6] Lemos RAdL, Silva TC, Tabak BM. Propension to customer chum in a financial institution a machine learning approach. Neural Computing and Applications, 2022.

[7] Keramati A, Ghaneei H, Mirmohammadi SM. Developing a prediction model for customer churn from electronic banking services using data mining. Financial Innovation, 2016.

[8] Reade W, Chow A. Binary Classification with a Bank Churn Dataset. Kaggle, 2024.

[9] Zou J, Han Y, So SS. Overview of Artificial Neural Networks. Methods in Molecular Biology™, 2008.

[10] Breiman L. Random Forests. Machine Learning, 2001, 45.

[11] Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry, 2015.

[12] Kleinbaum DG, Dietz K, Gail M, et al. Logistic regression. New York: Springer-Verlag, 2002.