Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220570

Data analysis and machine learning in the context of customer churn prediction

Changran Jing[1,2]

[1]School of Naval Architecture, Ocean & Civil Engineering, Shanghai Jiao Tong University, Shanghai, China

[2]yuanjiekongchan@sjtu.edu.cn

Abstract. Due to the fierce competition in the market, customers are often faced with multiple choices when choosing products and services. So many industries, including banking, are now facing the problem of how to address customer churn. At the same time, in order to improve the quality of service for users, banks and other institutions need to conduct in-depth research on the characteristics of customers. This paper provides solutions to the above two problems by using data analysis and mining technology and machine learning technology in artificial intelligence. The study provides an in-depth exploration of customer churn likelihood by analyzing customer behavior and characteristics. This study used data analysis methods such as Chi-square test, Mann Whitney U test, linear regression, and machine learning methods such as logistic regression, random forest, and XGBoost. This research uses public datasets on Kaggle. This study uses data analysis techniques to provide recommendations on how the banking industry should improve service quality, and establishes 6 models with better performance to predict customer churn. In addition, this paper also uses a variety of evaluation indicators to compare the model performance, and selects the random forest model with high predictive ability as the most suitable model. In addition, the order of importance of the factors responsible for customer churn was successfully derived.

Keywords: customer churn in bank, hypothetical test, linear regression, logistic regression, random tree, naive Bayes, XGBoost, LightGBM, CatBoost.

1. Introduction

Customer churn is a problem that modern merchants in all fields need to carefully analyze. According to related research [1], in order to reduce costs, merchants should focus more on retaining existing customers rather than attracting new ones. At the same time, since customers often have a variety of similar choices when choosing goods and services, if merchants cannot provide customers with satisfactory services, there will be a great trend of customer churn. Therefore, it is one of the important topics in the field of business data analysis to study the characteristics of customers, analyze the main factors leading to customer churn, and establish a set of accurate and effective customer churn prediction models.

For the above problem, Benlan He et al. first used an SVM model to predict customer churn based on a Chinese commercial bank consumer dataset containing 46,406 valid data records [2]. Research by Abdelrahim Kasem Ahmad et al helps telecom operators predict the customers most likely to churn [3]. Meanwhile, the study by T. Vafeiadis et al. compares the performance of different machine learning

DOI: 10.54254/2755-2721/2/20220570

models on the same dataset, including SVM, artificial neural network (ANN), naive Bayes, decision tree learning and logistic regression, and successfully integrates augmented [4]. The accuracy of the model SVM (SVM-POLY and AdaBoost) is improved to nearly 97% [5]. However, the above-mentioned studies were limited by the time of publication, and failed to conduct comparative studies on some of the latest models. At the same time, they were lacking in researching user characteristics and giving relevant financial explanations.

This paper is organized as follows: Chapter 2 introduces the methods used in the study and their rationale, including data processing techniques such as Mann-Whitney U test (M-W U test), chi-square test Kolmogorov-Smirnov test (K-S test), Kruscal-Wallis test (K-W test), Pearson correlation, and Machine learning models like Logistic Regression, Naive Bayes, Random Forest, XGBoost, LightGBM, and CatBoost. The third chapter introduces the process of data analysis. Chapter 4 focuses on the construction of the model and compares the performance of the models. Chapter 5 summarizes the research topic and provides some financial interpretations of the experimental results.

2. Methodology

The purpose of this chapter is to introduce the methods or models used in the research. Due to the complexity of the statistical methods involved, they are distinguished in Table 1.

2.1. M-W U test

In order to perform nonparametric tests, especially for the test of the null hypothesis between discrete variables and continuous variables, we usually choose the Mann-Whitney u test. This test for statistical significance is the random selection of values X and Y from two populations such that the probability that X is greater than Y is equal to the probability that Y is greater than X.

2.2. Pearson simple correlation

When we are calculating whether there is a linear relationship between two variables, one of the most important and intuitive methods is to use the Pearson simple correlation. This method can be used to calculate the correlation coefficient between quantitative variables. It is generally believed that its absolute value close to 1 indicates a strong linear correlation, and close to 0, there is almost no correlation. When the covariance and standard deviation of two variables are known, we define the Pearson correlation coefficient as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{1}$$

2.3. K-S test

If you want to use a nonparametric test method to determine the distribution of a group of samples, the best solution is to use the K-S test. the test method can calculate the probability that the selected group of samples is drawn from the expected distribution (that is, to verify whether the sample belongs to the expected probability distribution), or calculate the selected two groups of samples from the same expected probability. The size of the probability drawn from the distribution (that is, to verify that the two sets of samples have the same probability distribution)

DOI: 10.54254/2755-2721/2/20220570

Table 1. Testing methods.

Categorical & Categorical		**Categorical & Quantitative independent**		**Quantitative & Quantitative**	
Most conditions	Chi-square test	Normal, x variables	ANOVA	Equation Tests	Correlation test
Ordinal variables	Rank sum test	Normal, 2 variables	T test		Regression test
Paired Variables	Paired Chi-Square Test	Non-normal, x variables	M-W U test	Residuals Tests	Normality test Independence test Unbiased test
Others	Fisher's exact probability test	Non-normal, 2 variables	K-W test		Covariance test

2.4. Chi-square test

The chi-square test is a classic nonparametric test whose principle is to test whether observations belong to different classes of probability distribution models by dividing the target sample into mutually exclusive groups.

2.5. Kruskal-Wallis test

The Kruskal-Wallis test is a nonparametric test method for testing the probability distribution of two or more groups of samples, also known as "K-W test", "H test". The condition of this test method is that the samples must be independent or uncorrelated, but there is no requirement for sample normality, so it is very suitable for testing when the data is non-normal. Its null hypothesis is that the probability distribution obeyed by each sample has the same median, so it should be considered that the null hypothesis can be rejected when the median of the probability distribution of at least one sample is different from the other samples.

2.6. Logistic regression

Logistic regression is a statistical model that predicts the probability of an event by taking logarithmic and linear combinations of the odds of the event. The basic definition of logistic regression is as follows:

$$p(s) = \frac{1}{1 + e^{-(x-\mu)/s}} \tag{2}$$

The logistic function is in the form above, where μ is a location parameter indicating the midpoint of the curve and s is a scale parameter.

2.7. Naive Bayes

The naive Bayes classifier is a model that deals with the problem of binary classification prediction. Although it is a simple probabilistic classifier and one of the simplest Bayesian network models, the kernel density estimation can still make the model have high accuracy.

Using the chain rule for repeated applications of the definition of conditional probability, the fundamental equation of Naive Bayes classifier can be written as follows:

$$p(C_k \mid x_1, ..., x_n) \propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{3}$$

Then a classifier from the probability model can be constructed :

DOI: 10.54254/2755-2721/2/20220570

$$\hat{y} = \arg\max_{k \in \{1,\ldots,K\}} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{4}$$

2.8. Random forest

Random Forest is a decision tree based algorithm. The word "forest" in its name is because the algorithm uses ensemble learning to bring together a set of decision tree classifiers. The word "random" in its name comes from the fact that independent decision tree classifiers can sample different random subsets of the training data to test against those classes. When filtering eigenvalues to corresponding categories (that is, making predictions), the biggest difference between this model and logistic regression is that the latter can only perform linear processing, while the former's screening process can be nonlinear. In addition, random forests also Baggs its features, which means that each decision tree used will be trained with a randomly selected limited number of features.

2.9. XGBoost

XGBoost is an efficient machine learning algorithm based on the radial Boosting framework. If you need to experiment in different environments, XGBoost will be a good choice, because of its good portability, it can experiment in mainstream distributed platforms. Its ability to process data on the order of one billion is a very powerful machine learning framework.The latest version integrates naturally with the DataFlow framework.

XGBoost has both first-order and second-order partial derivatives, and can choose to perform classification or regression analysis as needed. When the second-order partial derivative obtained by Taylor expansion is selected as the independent variable, the leaf splitting optimization calculation can be automatically performed according to the input data value without selecting the specific form of the loss function. This feature enables XGBoost to perform gradient descent faster and more accurately.

2.10. LightGBM

LightGBM is an excellent, tree-based gradient boosting framework. Compared with the existing boosting framework, the advantages of LightGBM are not only higher efficiency and accuracy, but also lower memory consumption. To further improve the speed of the framework, people conduct learning experiments by setting specific parameters on multiple machines. LightGBM running on this basis completes the linear acceleration.

DOI: 10.54254/2755-2721/2/20220570

Table 2. Feature name, description, type.

Feature Name	Feature Description	Feature Type
CreditScore	Customer's credit score.	Quantitative
Geography	Customer's nation, including Spain, Germany and France.	Nominal
Gender	Male and Female.	Nominal
Age	Customer's age.	Quantitative
Tenure	Number of years as a bank customer.	Quantitative
Balance	Bank balance for each customer.	Quantitative
NumOfProducts	Number of bank products the customer is utilizing(savings account, mobile banking, internet banking etc.).	Quantitative
HasCrCard	Binary flag for whether the customer holds a credit card with the bank or not.	Nominal
Estimated Salary	Estimated salary of the customer in Dollars.	Quantitative
Exited	Binary flag 1 if the customer closed account with bank and 0 if the customer is retained.	Nominal
IsActiveMember	Binary flag for whether the customer is an active member with the bank or not.	Nominal

2.11. CatBoost

CatBoost is an open source software library that can solve classification problems well, and it provides a gradient boosting framework compared to traditional algorithms. It is suitable for a variety of systems and languages, showing good portability and compatibility.

3. Data analysis

The purpose of this chapter is to describe, test and analyze the data, obtain the statistical laws hidden behind the original data, and explain them with appropriate financial principles, which will help to deepen the understanding of commercial banks to their customers.

3.1. Dataset description

This public dataset from Kaggle contains 10,000 customer information of commercial banks. The most critical variable "Exited" is the customer churn information. The value of the variable "1" means that the customer has lost, and "0" means that the customer has not lost. Some important variables and their meanings are explained in Table 1.

3.2. Descriptive statistics

About the dataset, we can derive some important and basic statistics and generate some plots of variables.

- The frequency of each variable is shown in Figure 1.

DOI: 10.54254/2755-2721/2/20220570

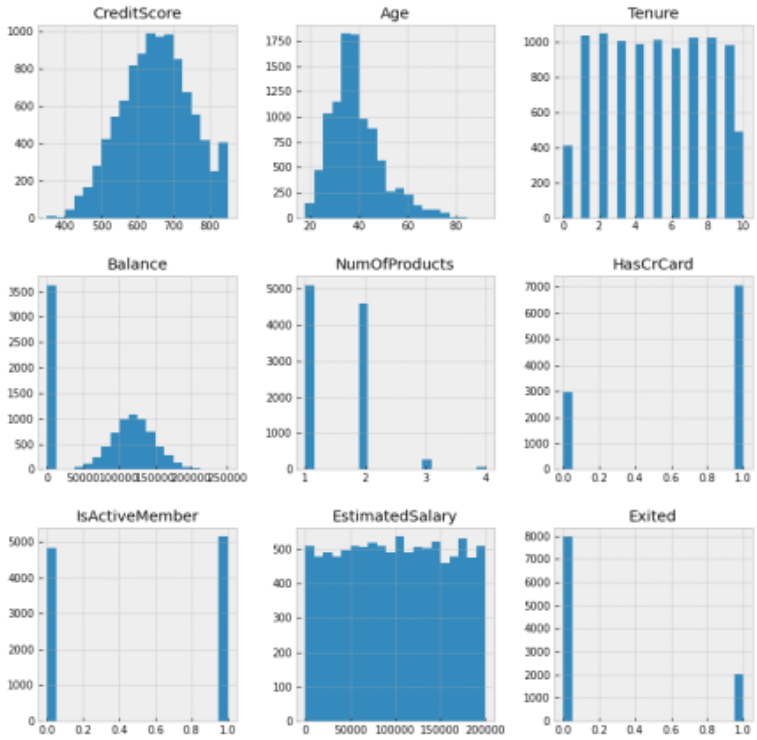

Figure 1. Frequency histogram for each variable.

- Nationality and gender are two types of categorical binary variables. By studying the churn of customers of different nationalities and genders and conducting comparative analysis, a preliminary classification of churn can made, as is shown in Fig. 2. and Fig. 3.

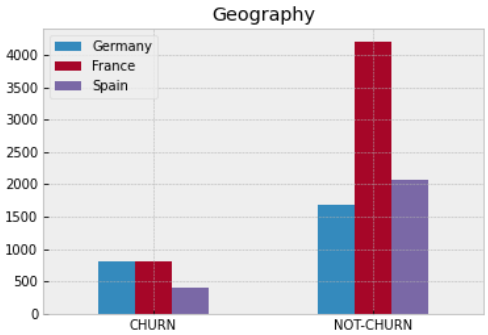

Figure 2. Customer churn of different nationalities.

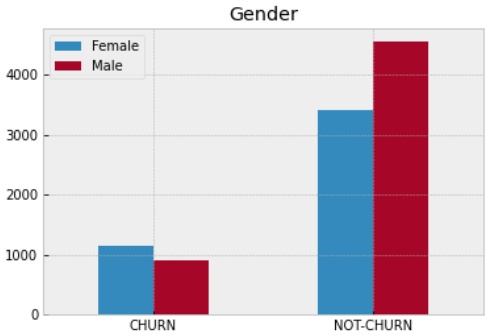

Figure 3. Customer churn of different genders.

DOI: 10.54254/2755-2721/2/20220570

3.3. Inspection and analysis

As one of the most important parts of data analysis, the purpose of hypothesis testing is to make assumptions about the characteristics and laws of variables, use statistical methods to test them, and draw conclusions through significance analysis, that is, to reject the original hypothesis or not to reject the original hypothesis. Suppose. In general, we consider that for p-values greater than or equal to 0.05 (significance level), we consider that the null hypothesis cannot be rejected, and for p-values less than 0.05, the results should be considered not statistically significant and the null hypothesis should be rejected .

In our study, we first used the K-S test to determine which distribution the samples belonged to. The advantage of using it is that as a nonparametric test, he does not need to know what distribution the sample belongs to. The null hypothesis (H0) is that the sample distribution is not significantly different from the expected distribution. The samples involved in the test are other variables except last name, customer ID, last name, and nationality, and the test distribution is normal distribution, uniform distribution, exponential distribution and Poisson distribution. After testing, it is only assumed that the p-value of "Estimated Salary conforms to a uniform distribution" is greater than 0.05, and the remaining variables do not conform to the above four classical statistical distributions.

Before examining the relationship between categorical variables and categorical variables, we need to discuss the test method in more detail. The chi-square test is applicable except for the following situations: the sample size is too small (the total sample size is less than 40), the low-frequency samples are too many (the number of cells with the theoretical frequency less than 5 exceeds 20%), a certain cell appears If the frequency is 0, the tested variable (dependent variable) is an ordinal variable and paired samples.

By observing the sample data, we can draw these two relatively macro conclusions: the sample size is sufficient (10,000 samples), and there are no ordered variables in the sample.

For analysis of variance, three conditions must be met, that is, the population obeys a normal distribution, the samples are independent of each other, and the homogeneity of variance is satisfied. Since the K-S test showed that the samples in each group did not conform to the normal distribution, and the kurtosis and skewness errors of the data were both greater than 1.96 times the standard error, the original data could not be analyzed by variance analysis.

Through Mann-Whitney u test (in the case of independent variables with 2 groups) or Kruskal-Wallis test (in the case of independent variables with multiple groups), we can judge for different groups of the same categorical independent variable, non-normal dependent variables distribution is the same. In the experiment, with each nominal variable as the independent variable and the quantitative variable as the dependent variable, the Mann-Whitney U test or the Kruskal-Wallis test was performed. The test results are shown in Table 2.

Table 3. P-value Of M-W U test & K-W test.

	CreditScore	***Age***	***Balance***	***Estimated***	***NumOfProduct***	***Tenure***
Exited	0.020	0.000	0.000	0.227	0.000	0.162
Gender	0.763	0.003	0.177	0.408	0.199	0.131
Geography	0.769	0.000	0.000	0.569	0.045	0.918
HasCrCard	0.704	0.127	0.325	0.315	0.700	0.025
IsActiveMember	0.015	0.000	0.250	0.251	0.103	0.004

In the same way, we can study the relationship between various categorical variables by cross-tabulation and chi-square test, as shown in Table 4.

DOI: 10.54254/2755-2721/2/20220570

Table 4. P-value Of Chi-square test.

	Exited	***Gender***	***Geography***	***HasCrCard***	***IsActiveMember***
Exited		0.000	0.000	0.475	0.000
Gender	0.000		0.031	0.564	0.024
Geography	0.000	0.031		0.327	0.070
HasCrCard	0.475	0.564	0.327		0.235
IsActiveMember	0.000	0.024	0.070	0.235	

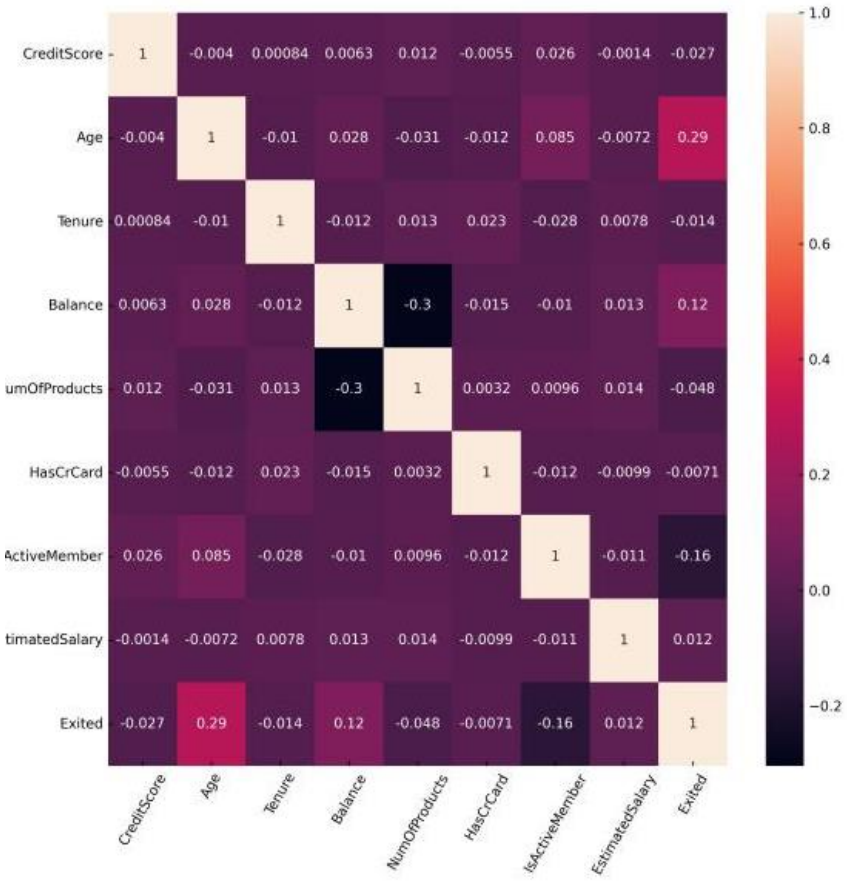

Figure 4. Pearson's correlation heatmap.

In the table, p-values below 0.05 indicate the dependent variable has a statistically significant difference among its different groups.

For the correlation between continuous variables, it is necessary to use linear regression to test. A linear regression test usually consists of two parts: an equation test and a residual test. The specific inspection content can be found in Table 1.

By calculating the Pearson correlation coefficient of each variable, we can obtain the linear correlation between each variable, and the closer the value is to 1 (or -1), the stronger the positive (or negative) correlation. Correspondingly, if the correlation is close to 0, it means that there is no obvious linear correlation between the two. The results are shown in Figure 4, which is a correlation heat map.

3.4. Conclusion and further explanation

Through the above test and analysis results, we can further discuss the issues we are most concerned about, and give some financial explanations and operational suggestions for commercial banks.

DOI: 10.54254/2755-2721/2/20220570

• For our most concerned customer churn data, we can see that Gender, Geography and IsActiveMember have a significant impact on whether or not to churn. Combined with the Pearson correlation coefficient, it can be seen that banks can reduce the operating risks caused by the easy loss of specific groups by providing customers from different regions and genders with as personalized services as possible. At the same time, due to its easy risk of churn, customers with low credit scores should be marked to some extent. On the contrary, banks should try their best to provide more satisfactory services to customers with high credit scores, or take certain rewards or rewards. mechanism to better retain these customers.

• By comparing the significance levels of each test, we found that the variables with the most significant impact on Exited are Age, Geography, Balance, NumOfProduct, IsActiveMember. The results will help us to judge the importance of factors in the subsequent machine learning modeling process, which will be discussed in detail in Chapter 4. For other metrics that banks may be concerned about, such as NumOfProduct, we test the results and find that whether the customer quits and the nationality of the customer have a significant impact on the number of products customers buy or use. This also suggests that banks can increase customers' desire to purchase products by providing customized services to customers in different regions.

In a study on bank customer analysis [6], the authors conducted a cluster analysis of bank customer groups by using indicators such as age, expected income, and credit score. This study shows that customers of different ages have significant stratification and differences in expected income, and it is appropriate to use age as an indicator to distinguish different types of customers. Due to the use of different research methods from this paper, the experimental results provide evidence for the above conclusions of this paper.

4. Machine learning

After the data analysis, we can start building a machine learning model to predict customer churn. The purpose of this chapter is to introduce the process of modeling, and to compare and summarize the prediction results in various aspects.

4.1. Data preprocessing

Due to the following shortcomings of the original dataset, we cannot directly use it for machine learning modeling:

• The original data does not have a unified dimension, such as the unit of age is years, and the dimension of EstimatedSalary is pound sterling. If you build probabilistic models, you don't need normalization, because they don't care about the value of the variable, but about the distribution of the variable and the conditional probability between the variables, such as decision trees and random forest. Optimization problems like logistic regression, KNN need normalization. In this experiment, we use the sklearn.preprocessing.MinMaxScale() function for normalization.

• According to the descriptive statistics of the sample data, the number of samples with a value of 0 in Exited is about 4 times the number of samples with a value of 1. If you directly use this as a label for supervised learning, the model will be more diverse in prediction. This judgment is a false negative, which affects the accuracy and practicability of the model. Therefore, the samples must be unbalanced. The samples are oversampled using the SMOTE object of the imblearn.over_sampling library. The principle is to generate new data points based on the features of the neighbors.

• For the Geography variable, we use one-hot encoding to convert it into three variables, namely "customer is German", "customer is French", and "customer is Spanish".

4.2. Model training

Before training starts, the normalized and oversampled dataset is divided into training and test sets. Due to the many features of the sample, it is inconvenient to carry out the run-length test, and it is impossible to determine whether the data generated after oversampling is random, so the K-fold cross-validation method is used for data and division. The principle is: the original data is randomly divided into K parts, K-1 parts are selected as the training set each time, and the remaining 1 part is used as the test set. The

DOI: 10.54254/2755-2721/2/20220570

cross-validation is repeated K times, and the average of the accuracy of the K times is taken as the evaluation index of the final model. In this experiment, the k value is 5, that is, the division and training are repeated 5 times and the average of the results is taken. Similar data preprocessing techniques were also used in the study led by Dudyala Anil Kumar D and V. Ravi [7], proving their scientificalness and effectiveness.

4.3. Result analysis

For classification problems, there are many measures to evaluate a model's predictive power and usefulness. Usually, we will use accuracy, recall, F-Score, ROC curve, confusion matrix and other methods to comprehensively evaluate the pros and cons of the model.

- Confusion Matrix, is also known as likelihood matrix or error matrix. In order to be able to visualize various types of values when dealing with the results of classification problems, we introduce confusion matrices. Each row of the confusion matrix represents the number of samples of the true classification, that is, the number of samples of each label in supervised learning. Correspondingly, each column of the confusion matrix represents the result obtained by the model prediction. As shown in Table 5, the labels in the actual classification are named true and false, while the predicted results are denoted by positive and negative.

Table 5. Typical confusion matrix.

		Prediction	
		Positive	***negative***
Reference	True	True Positive(TP)	True Negative(TN)
	False	False Positive(FP)	False Negative(FN)

- Accuracy refers to the percentage of the total number of samples that the model predicts correctly in a classification problem. Although it is simple and intuitive to use the accuracy rate as the evaluation index of the model, it is not correct in many cases. For example, considering that some samples should be given higher weights, or that certain types of samples are the purpose of our research in special cases, such as earthquake prediction, financial transaction fraud, cancer prediction and other problems, correctly detecting the value of 1 is more practical than a higher global accuracy.

$$precision = \frac{TP}{TP + FP} \tag{5}$$

- Recall is the ratio of true positives to the sum of true positives and false negatives. In problems such as earthquake prediction, cancer diagnosis, fraud prediction, etc., it is more meaningful and practical to predict all true positives than the global accuracy, because these true positive samples and what we really care about.

$$recall = \frac{TP}{TP + FN} \tag{6}$$

- With the above two concepts, we can know that accuracy and recall have their own application scenarios, and are two different dimensions for evaluating the quality of machine learning models. The problem derived from this is that in more complex practical problems, we may not be able to clearly judge which of the two indicators of accuracy and recall is more important to us, and for larger data sets, these two indicators are often negatively correlated. So we introduced F-score, an indicator that comprehensively considers precision and recall.
- In the ROC curve graph, the abscissa is the false positive rate, and the ordinate is the true positive rate. It is generally considered that the intersection of the ROC curve and the -45 degree point is the equilibrium point, which can be used to measure the merits of the model without considering the cost factor.We use AUC to represent the area under the ROC curve.

DOI: 10.54254/2755-2721/2/20220570

In the experiment, we calculated and obtained the indicators of the above six models, as shown in Table 6. At the same time, we also obtained the ROC curves of each model, as shown in Figure 5.

Table 6. Results for models.

	Logistic Regression	***Naive Bayes***	***Random Forest***	***XGBoost***	***LightGBM***	***CatBoost***
Precision	0.887	0.816	0.919	0.915	0.915	0.918
Recall	0.765	0.781	0.868	0.866	0.859	0.858
F-score	0.821	0.798	0.893	0.890	0.886	0.887
AUC	0.914	0.879	0.958	0.956	0.957	0.958

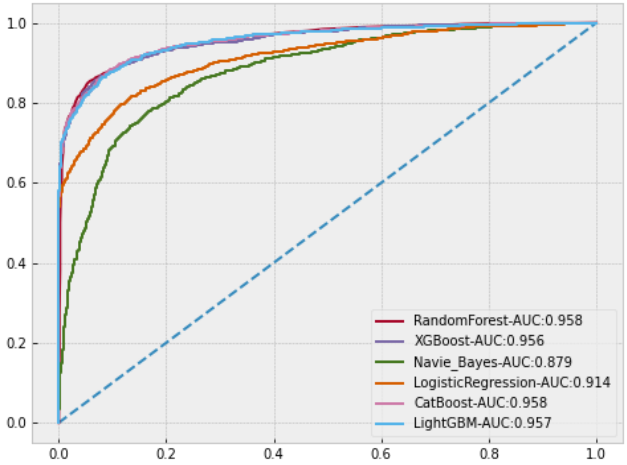

Figure 5. ROC curve for each model.

From the experimental results, we can conclude that Random Forest has the highest precision and AUC values, so it performs best for this customer churn prediction problem. In another study [8], it was also confirmed that random forests can significantly improve prediction accuracy compared to other algorithms such as artificial neural networks, decision trees, and class-weighted core support vector machines (CWC-SVM). By the way, we can notice that the performance of XGBoost, LightGBM, and CatBoost models are all good, and they are also suitable as ideal solutions for this problem. Although the predictive ability of SVMs was not compared in this study, numerous studies have analyzed the performance of this traditional algorithm on classification problems, especially customer churn. For example, in a study [9] led by Kristof Coussement et al., it was shown that support vector machines can only outperform logistic regression if the optimal parameters are selected, and the random forest algorithm generally outperforms both algorithms.

In addition, we use random tree model to draw to study the importance of each factor in predicting customer churn. The results are shown in Figure 6.

DOI: 10.54254/2755-2721/2/20220570

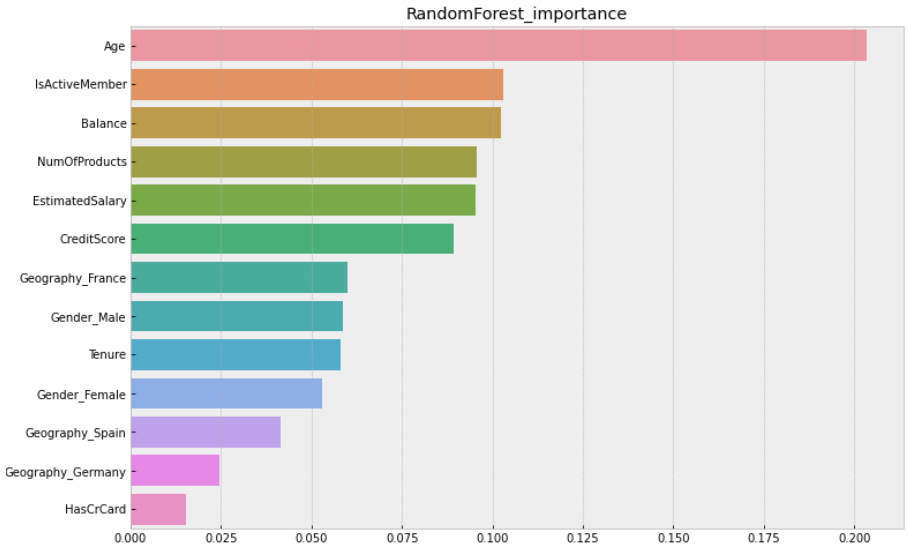

Figure 6. Importance of variables.

5. Conclusion

The purpose of this research is: to select the factors that have the most significant impact on customer churn through data analysis, to obtain some suggestions that are helpful for banks to establish user profiles and improve service quality through data analysis, and to establish the most suitable model to predict the bank's early customers churn.

This study uses only a small amount of data (10,000 samples), and the sample data is highly unbalanced, reflected in non-normality, non-uniform number of target features, and so on. But this is not a problem to worry about. In the experiments we solve this problem by oversampling, and in the real environment, the huge data volume of commercial banks can overcome this difficulty.

The data analysis part studies the statistical characteristics of the data, and studies the distribution characteristics of the data population. Based on the non-normality of the data, M-H U test, K-W test and chi-square test were selected to test and analyze the significance degree (p value) of the mutual influence between each variable. Using the method of linear regression analysis, the linear correlation between the data is analyzed, and some suggestions are obtained that are beneficial to the bank to improve service quality and reduce customer churn, such as the age, nationality, etc. between lost customers and non-churned customers. As a result, there are significant differences in characteristics such as credit scores, and banks should design individualized and differentiated measures to meet the needs of customers with different basic backgrounds.

In the part of machine learning modeling, this study uses logistic regression, naive Bayes, random forest, XGBoost, LightGBM, CatBoost for modeling and comparative analysis. Compared with Manas Rahman et al.'s research [10] on the classification problem of the same data set, this study adopts more diversified results comparison indicators, and the performance of each model has been improved to a certain extent. The experimental results show that random forest has the best prediction performance. The three new high-performance algorithms, XGBoost, LightGBM, and CatBoost, have similar predictive capabilities to random forests. However, some classical methods such as logistic regression and naive Bayes have relatively poor performance results in this problem, and the performance on the AUC value is about 5% and 8% different from the previous methods, respectively.

At the same time, we selected the best performing random forest model for factor importance analysis. The results show that the age of the user, whether he is an active user of the bank, and Balance has a relatively high impact on customer churn. This result also has a certain degree of coincidence with the detection results obtained in the data analysis part.

DOI: 10.54254/2755-2721/2/20220570

References

[1] Kim M K, Park M C, Jeong D H. The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services[J]. Telecommunications policy, 2004, 28(2): 145-159.

[2] Mihelis G, Grigoroudis E, Siskos Y, et al. Customer satisfaction measurement in the private bank sector[J]. European Journal of Operational Research, 2001, 130(2): 347-360.

[3] He B, Shi Y, Wan Q, et al. Prediction of customer attrition of commercial banks based on SVM model[J]. Procedia computer science, 2014, 31: 423-430.

[4] Vafeiadis T, Diamantaras K I, Sarigiannidis G, et al. A comparison of machine learning techniques for customer churn prediction[J]. Simulation Modelling Practice and Theory, 2015, 55: 1-9.

[5] Ahmad A K, Jafar A, Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform[J]. Journal of Big Data, 2019, 6(1): 1-24.

[6] Zakrzewska D, Murlewski J. Clustering algorithms for bank customer segmentation[C]//5th International Conference on Intelligent Systems Design and Applications (ISDA'05). IEEE, 2005: 197-202.

[7] Anil Kumar D, Ravi V. Predicting credit card customer churn in banks using data mining[J]. International Journal of Data Analysis Techniques and Strategies, 2008, 1(1): 4-28.

[8] Xie Y, Li X, Ngai E W T, et al. Customer churn prediction using improved balanced random forests[J]. Expert Systems with Applications, 2009, 36(3): 5445-5449.

[9] Coussement K, Van den Poel D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques[J]. Expert systems with applications, 2008, 34(1): 313-327.

[10] Rahman M, Kumar V. Machine learning based customer churn prediction in banking[C]//2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020: 1196-1201.