Prediction and Clustering of Bank Customer Churn Based on XGBoost and K-means

Tiansheng Zhang*

School of Economics, Fudan University, Shanghai, China

*Corresponding author: 19300680021@fudan.edu.cn

Abstract. Due to the fierce competition of commercial banks, customers are becoming more and more important to banks. Therefore, customer churn has become a major problem that banks need to face. In this paper, XGboost algorithm was used on a data set of customers of a US bank from Kaggle to predict customer churn, and grid search method was used to find the best hyperparameters. Moreover, K-means algorithm is adopted to further subdivide the lost customers. For predicting customer churn, XGBoost algorithm achieves 0.84 in accuracy, 0.83 in precision, 0.84 in recall and 0.84 in F1 score on the test set. And the most important score for features in the case of the algorithm adopted are customers' estimated salary, credit score and balance. For the segmentation of churn customers, K-means algorithm divides these customers into 5 groups. These five groups of customers have different values for banks, so this paper puts forward corresponding recovery suggestions for their respective characteristics..

Keywords: Bank Customer Churn, XGBoost, K-means.

1. Introduction

With the continuous development of Internet finance, retail finance business gradually occupies an increasingly important position in commercial banks. Retail finance business are increasingly diversified, providing individuals, families and small and medium-sized enterprises with comprehensive and integrated financial services, including deposits and withdrawals, loans, settlement, exchange, investment and finance. Therefore, as the source of bank profits, customers are regarded as the most valuable assets and wealth of banks. Many banks spend a lot of energy on finding new customers, hoping to increase the number of customers to expand profits, but often neglect the maintenance of existing customers. But research shows that it costs less to keep existing customers than to acquire new ones [1]. Customer churn has thus become a serious problem for many banks. Here customer churn is defined as the propensity of customers to cease doing business with a company in a given time period [2]. Therefore, it is very important for banks to predict customer churn, analyze the characteristics of customer churn and develop targeted strategies to retain customers by establishing models based on customer data.

Over the past decade, various machine learning algorithms have been widely used to build models for solving customer churn problem in banks. In [3], Back Propagation (BP) neural network algorithm was used to predict customer churn, but the accuracy of prediction needs to be improved. In [2], logistic regression, decision tree, neural nets, discriminant analysis and other methods were compared and classified into 5 categories, namely “logit”, “tree”, “practical”, “discriminant” and “explain”. After comparison, the authors found that Logistic and tree approaches perform relatively well, and these models have staying power. Besides, in [4], the authors adopted logistic regression and decision tree algorithms and further developed a new measure criterion called misclassification cost taking the economic cost into the evaluation of the model. Recently, in [5, 6], ensemble learning algorithms were applied to this problem such as Catboost, Lightgbm, Random Forest, XGBoost and Stacking ensemble learning method integrating XGBoost and logistic regression. These ensemble learning models are better in accuracy and recall of prediction. However, these literatures lack further research on the characteristics of churn customers. Some clustering algorithms have been used in customer classification, among which K-means algorithm is most popular [7-9]. In the field of bank customers, in [10], the authors compared K-means, DBSCAN and X-means, but this research focuses on the

division of all customers to segment market for banks, and the difference between churn customers remain unclear.

Therefore, this study hopes to further classify the churn customers so that banks can develop targeted recovery strategies for different types of churn customers, after predicting customer churn based XGBoost algorithm, the performance of which is relatively satisfying. In this regard, more attention was paid to churn customers. K-means algorithm was used in this study to classify churn customers. The result is that all the churn customers can be divided into five clusters, and corresponding recovery strategies was made for them according to the characteristics of the five clusters of churn customers.

The rest of this paper is organized as follows: Section 2 provides a discussion of this approach and dataset. Section 3 provides the experiment details and a discussion of the result. The conclusions of this work and future work are discussed in Section 4.

2. Methods

2.1 Dataset description and preprocessing

This study used a customer churn dataset of a U.S. bank from Kaggle [11]. This dataset provided customers' detailed information such as their customer ID, surname, credit score etc. The original dataset consists of 10, 000 customers' data and 13 features. The specific meanings of these features are shown in Table.1.

Table.1. Detailed information of all features

Name	Description	Type
CustomerId	Customer's ID	Classification (string)
Surname	Surname	Classification (string)
CreditScore	Credit Score of the customer	Numeric
Geography	location of customer	Classification (string)
Gender	Gender whether male or female	Classification (string)
Age	Age of the customer	Numeric
Tenure	From how many years customer is in bank	Numeric
Balance	Average balance of customer	Numeric
NumOfProducts	Number of bank products customer is using	Classification
HasCrCard	Whether or not the customer has a credit card	Classification
IsActiveMember	Whether or not the customer is active	Classification
EstimatedSalary	Estimated salary of the customer	Numeric
Exited	Whether or not the customer will leave the bank	Classification

The preprocessing is consisted of four parts. First, irrelevant features were removed. The CustomerId and Surname have nothing to do with whether the customer is a churner or the characteristics of the churn customer, so they were removed. Second, string data such as Geography and Gender was transferred to numerical data. For example, for Gender, 'male' is now 1 and 'female' is now 0. But for Geography, different methods are used for different purposes of data. For prediction, Geography was replaced by two new features, inFrance and inSpain. For example, if Geography of a customer was 'France', then his inFrance feature is 1 and inSpain is 0. Or if Geography was 'Germany', then both inFrance and inSpain are 0. But for churn customer segmentation, in order to prevent increasing the weight of Geography feature, 'Spain' is now 0, 'France' is 1 and 'Germany' is 2. Third, for prediction purposes, the data set was randomly divided into just training set and test set in the 8:2 ratio. Later, the hyperparameters will be determined by 5-fold cross verification of the training set. However, the training set was not balanced, because there were 6365 existing customers and only 1635 churn customers. Several methods can be used to solve sample imbalance problems,

such as over-sampling and under-sampling. This article used Synthetic Minority Oversampling Technique (SMOTE) to solve this problem [6]. After SMOTE, there are 6365 customers in both groups. But for churn customer segmentation, allchurn customers were simply singled out. Fourth, Z-Score normalization was applied to all data sets. Sample data is shown in Table.2 and Table.3.

Table.2. Sample data for customer churn prediction

Credit Score	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	inFrance	inSpain	Exited
-0.148	1.164	0.234	0.076	-1.316	1.029	0.784	1.304	-0.916	1.115	-0.566	0
0.145	1.164	2.263	1.177	0.922	-0.725	0.784	1.304	-1.010	1.115	-0.566	0
0.493	-0.859	0.843	1.544	-1.316	1.029	0.784	-0.767	1.334	-0.994	1.948	0
-1.191	-0.859	0.437	-0.291	-1.316	-0.725	-1.276	-0.767	1.238	0.924	-0.338	1
0.102	-0.859	-1.287	0.076	-1.316	1.029	-1.276	-0.767	1.170	-0.525	1.389	1

Table.3. Sample data for churn customer segmentation

Credit Score	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
-0.263	-0.262	-0.888	-0.291	-0.999	-1.561	-0.593	0.656	1.331	-0.002
-1.429	-0.262	-0.888	-0.291	1.045	1.175	1.902	0.656	-0.751	0.215
-0.004	-1.594	1.126	-0.086	1.045	0.388	0.655	0.656	-0.751	0.834
-2.685	1.070	-0.888	-1.622	-0.318	0.410	3.150	0.656	-0.751	0.309
0.076	1.070	1.126	1.348	-1.339	0.711	-0.593	0.656	-0.751	-1.664

2.2 Proposed approach based on the machine learning

2.2.1 XGBoost

For customer churn prediction, XGBoost algorithm was adopted in this paper. XGBoost is a kind of Boosting algorithm which is a family of ensemble algorithms that can boost weak learners to strong ones. The idea of XGBoost is to keep adding trees, keep doing feature splitting to grow a tree. When the training is completed, k trees are obtained. In order to predict the score of a sample, according to the characteristics of the sample, the sample is corresponding to a leaf node in each tree, and each leaf node corresponds to a weight. The sum of these weights is the predicted value of the sample. XGBoost has several advantages: the addition of regularization terms to prevent overfitting, the use of second-order Taylor expansion in the objective function for high accuracy, and support for multiple base classifiers [6]. The object function of XGBoost is:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

where f_k is the Kth tree and $\Omega(f_k)$ is the regularization term, defined as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2 \quad (2)$$

where T is the number of leaf nodes and $\frac{1}{2}\lambda||w||^2$ is L2 regular term.

XGBoost has a lot of hyperparameters to set: number of trees, minimum leaf node sample weight sum, maximum depth of tree, minimum loss function drop value required for node splitting, L2 regularization term, L1 regularization term, learning rate. In this paper, grid search algorithm and k-fold cross validation method are used to determine the above hyperparameters on the training set, and the results are 160,1,100,0,1,0.2,0.1 respectively.

2.2.2 K-means

K-means, one of the most representative and popular unsupervised clustering algorithms, has the advantages of simple principle, fast operation speed and relatively good effect. Therefore, this paper adopted K-means to subdivide the churn customers. The thought of k-means is that for the given sample set, according to the distance between samples, the sample set is divided into K clusters so that the points in the cluster are closely connected together, and the distance between clusters is as large as possible. In other words, the purpose of k-means is to minimize the squared error E:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - u_i||_2^2 \quad (3)$$

Where u_i is the mean-valued vector of cluster C_i. The working process of k-means is as follows: firstly, K samples are randomly selected as the initial clustering center, and then the distance from each sample to the clustering center is calculated. Put the sample into the class where the nearest cluster center is located, recalculate the cluster center of the adjusted new class, and repeat this process until there is no change in the cluster center of the adjacent two times. At this time, the sample adjustment is finished, and the algorithm has converged.

Therefore, a difficulty in adopting k-means algorithm is that the number of clusters, K, needs to be determined first. In this paper, the value of K was determined by finding the distortion score elbow for K-means clustering. As shown in Figure 1, after comparing the scores of K range from 2 to 12, the value of K was determined to be 5.

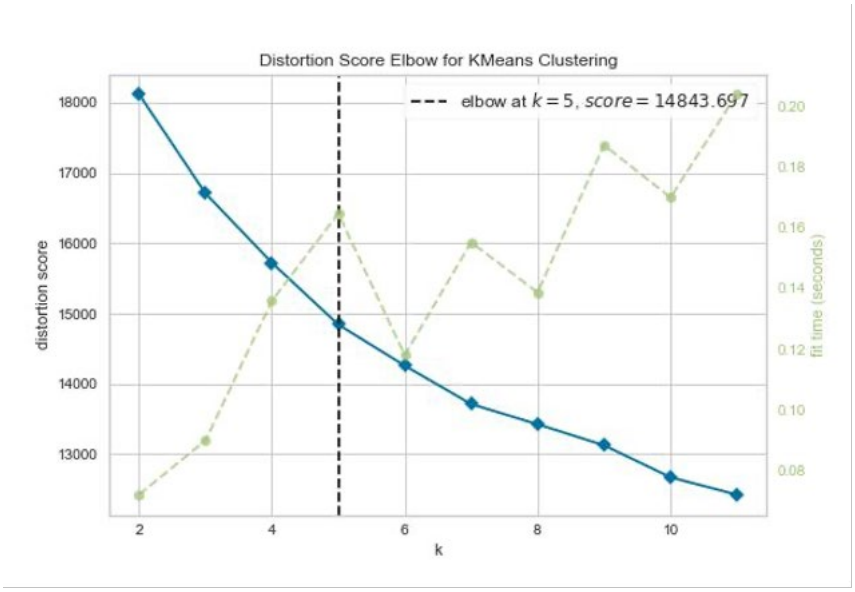

Figure 1. Distortion score elbow for K-means clustering

3. Results and discussion

3.1 Customer churn prediction using XGBoost

After adjusting the superparameters, the optimum result is shown in Table.4, with an accuracy of 83.65%. To be more specific, for predicting customers that won't leave, the precision, recall and F1 score are all pretty high, while for predicting churn customers, the performance is not very good. Of all the customers judged as churner, 60% are true churn customers, and of all the actual churn customers, only 58% are correctly predicted. However, the overall performance can be considered satisfactory. From Figure 2, the most important score for features in the case of the algorithm adopted are customers' estimated salary, credit score and balance.

This relatively poor performance for churn customers may be due to the low rate of customer attrition. Although the imbalanced dataset was processed by SMOTE algorithm, it may still have a negative impact on the results. Besides, the bank should pay more attention to the three most important customer features. Through the observation of the sample, it can be found that the average estimated income and balance of churn customers are higher. This suggests that the bank has yet to develop an

incentive strategy for high-earning customers and does not offer benefits to high-balance customers. Therefore, banks should provide more professional and personal services to customers with high income or high balance to enhance their wealth.

Table.4. The performance of XGBoost.

Category	Precision	Recall	F1-score	Accuracy
total	0.83	0.84	0.84	83.65%
0	0.89	0.90	0.90	-
1	0.60	0.58	0.59	-

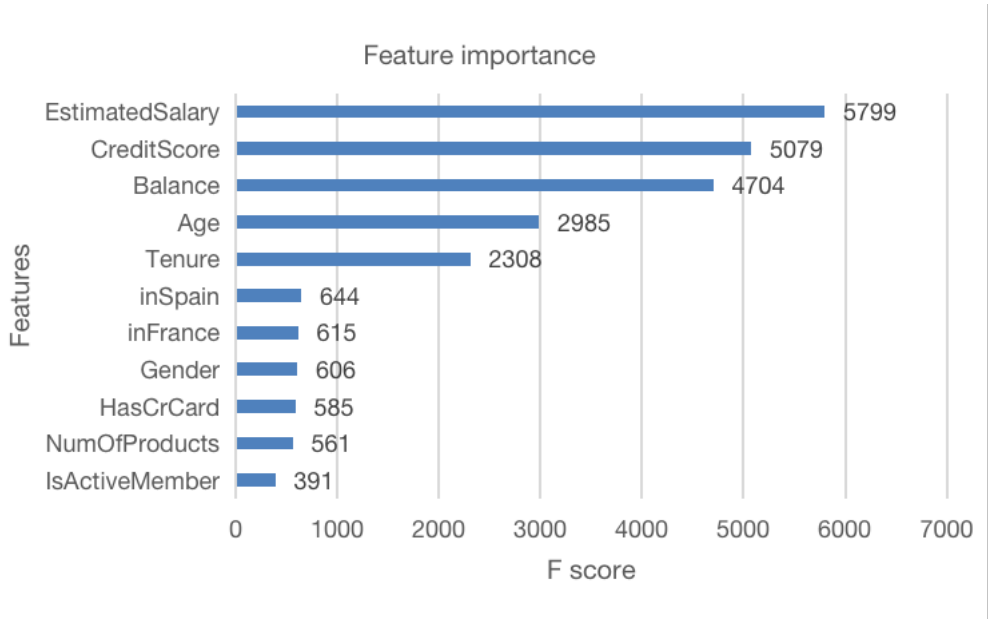

Figure 2. Feature importance in predicting customer churn

3.2 Churn customer clustering using K-means

Table.5. Cluster centers

Cluster	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Number
1	646.1	1.41	0.42	44.64	4.68	122138.91	1.22	1	1	98879.01	305
2	642.92	1.48	0.47	44.88	5.08	125216.5	1.17	1	0	100843.5	564
3	643.62	0.62	0.42	44.85	4.86	8706.41	1.21	0.73	0.31	101386.16	489
4	650.12	1.35	0.47	45	4.9	122988.55	1.21	0	0.41	100945.75	400
5	645.64	1.19	0.4	44.72	5.1	86956.16	3.22	0.71	0.41	106435.92	279

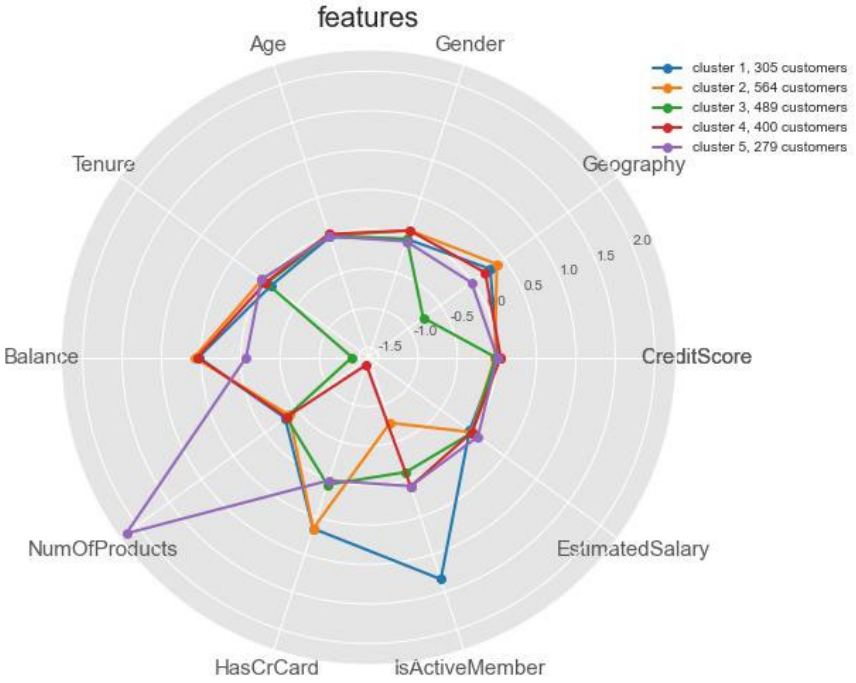

Figure 3. Cluster centers based on the K-means.

After the application of K-means algorithm, lost customers are divided into 5 categories, as shown in the radar chart in Figure 3. After restoring the original data scale, the results are shown in Table 5. Cluster 1 has 305 customers and has the highest customer activity. Cluster 2 has the largest number of customers, 564. Cluster 3 has 489 customers and the lowest balance per capita. Cluster 4 has 400 customers, most of whom do not have credit cards. Cluster 5 has only 279 customers and has the largest variety of products per capita.

Among these 10 features, balance and number of products are the most important for banks. Customers with a high balance and a high number of products can actually bring more profit to the bank, because the profit model of the bank is based on interest rate spread and various service fees as profit sources. So the bank should pay different attention to the customers who can bring different profits to the bank. As a result, the bank should make the most effort to retrieve cluster 5 customers, such as promoting various products and services and recommending appropriate products to each customer according to their risk tolerance, because that's the most efficient. Because customers of this cluster have the highest number of products and the average balance is not low, retrieving customers of this cluster can bring higher returns than retrieving customers of other clusters. However, this group of customers is relatively small, accounting for only a small fraction of all churn customers. Secondly, the bank should focus on the customers in clusters 1, 2 and 4, because these customers have similar average balance and number of products and therefore have similar value to the bank. These customers have a smaller number of products than cluster 5, so the bank should advertise more products to them more frequently. In addition, the bank should strongly recommend credit cards to cluster 4 customers, since only a very small percentage of cluster 4 customers have credit cards, which are very helpful in keeping customers engaged. At the same time, the bank should increase contact and access to cluster 2 customers to increase their activity. Thirdly, although the balance and number of products of cluster 3 customers are small, the bank should not ignore these customers because according to the above discussion, the estimated salary of churn customers is relatively high, and the estimated salary of cluster 3 customers is also at the average level of all churn customers. These customers have the potential and ability to bring more profits to the bank. Therefore, for these customers, the bank need a more long-term strategy, that is, to enhance its brand awareness and professionalism so that these customers can trust the bank more, so that they can deposit more and buy more products.

4. Conclusions

This work not only predicted customer churn and found the most important features, but also further subdivided churn customers to give corresponding recovery strategies. XGBoost algorithm is used to predict customer churn, and K-means algorithm is used to cluster the churn into 5 categories. The results showed that predicting who won't churn outperforms predicting who will churn. And the bank should devote three different levels of effort to recovering five categories of churn customers to maximize its profit. In the future, more research should be focused on improve the algorithm, other than SMOTE, to deal with sample imbalance, so that the performance of predicting customers who will leave.

References

[1] Roberts J H. Developing new rules for new markets [J]. Journal of the Academy of Marketing Science, 2000, 28(1):31.

[2] Neslin S A, Gupta S, Kamakura W, et al. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models[J]. Journal of marketing research, 2006, XLIII (2):p.204-211.

[3] Zhao S X, Tai Q Y. Applied Research on Data Mining in Bank Customer Churn[J]. Applied Mechanics and Materials, 2014, 687-691:5023-5027.

[4] Nie G, Rowe W, Zhang L, et al. Credit card churn forecasting by logistic regression and decision tree[J]. Expert Systems with Applications, 2011, 38(12):15273-15285.

[5] Y Deng, Li D, Yang L, et al. Analysis and prediction of bank user churn based on ensemble learning algorithm[C]// 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA). IEEE, 2021.

[6] M Zhang. Bank credit card customer churn Prediction based on Logistic regression and XGBoost [D]. Shandong University.

[7] Hruschka H, Natter M. Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation [J]. European Journal of Operational Research, 1999, 114(2):346-353.

[8] 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.

[9] Y. Qiu, 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.

[10] Hua H Y, Zhao H C. Application of Clustering Algorithms in Bank Customer Segmentation [J]. Computer Engineering, 2008, 34(24):37-39.

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