ISSN: 2688-9323

Customer Churn Prediction based on Multiple Algorithms

Mengting Zheng*

School of Management, Shanghai University, Shanghai 200000, China

Abstract

Customer classification in bank customer relationship management is an important research problem in bank business field. At present, economic competition is strengthened, customer urn is serious and soaring. Therefore, it is of great significance for bank development to predict customer churn accurately. A study using the public bank customer churn data set. First, according to the literature research according to the bank customer churn reason will data set the properties of the divided into two categories, natural attribute, competitive and negligent attribute, the decision tree analysis method is used to identify the important attribute, and according to the actual classification to discuss important attributes. And then support vector machine algorithm and logistic regression classification algorithms are compared, which found that support vector machine algorithm is the best performance.

Keywords

Customer churn, decision tree, support vector machine, classification prediction.

1. Introduction

With the rapid economic development, the connotation and objectives of bank Customer Relationship Management have changed recently. For example, Liu Zhigang elaborated the evolvement of the content and form of the bank-customer relationship management strategy from the god-type to the strategic cooperative type[1]. Banks pay more attention to customer relationship management. Lin Bin et al. also indicated that the CRM of commercial banks should not only focus on the customer as the center but also use the CRM system to achieve the goal of reducing the bank's operating costs and maximizing the customer's income[2]. And customer classification is one of the basic principles of customer relationship management.

Moreover, with the strengthening of economic competition and the emergence of homogenization of banking services, the problem of customer churn of commercial banks is becoming more and more serious. At present, the current churn rate for bank customers can be as high as 20%, which is much higher to obtain a new customer than to maintain an old customer[3,4]. Therefore, it is very important to establish an efficient customer churn early warning system by mining the information that has an impact on churn from massive customer transaction data. However, the previous business model with products as the core and sales as the target is no longer suitable in the age of the Internet economy. Therefore, the customer demand-oriented marketing model has become an important content for banks to maintain customers and carry out fine management. And reasonable and accurate classification of lost customers is an important means and premise for commercial banks to maintain different levels of customers and improve profits.

Data mining is a method to discover and extract knowledge from many useful, practical and understandable data. With the advent of the era of big data, data mining methods have been applied in many industries, such as medical[5], financial[6]. One of the most important uses of data mining is to extract knowledge from data in a shorter time, less cost and more accurately, to obtain more comprehensive and complete results. These knowledge has been applied in many fields such as medical application, security, crime prevention, finance[7]. Among them,

ISSN: 2688-9323

some machine learning algorithms commonly used in data mining are often used to deal with the complexity and diversity of bank customer data attributes and the huge amount of data. In recent years, the classification machine learning algorithm is used to predict the types of bank customer churn in the financial field. These researches are of great significance to the development of bank business marketing and future business development.

2. Literature Review

2.1. Customer Churn

2.1.1. Definition and Classification of Bank Customer Churn

Customer churn refers to the fact that the customer does not repeat the purchase or terminate the service originally used. Bank customer churn refers to the bank's customers terminate all business in the bank and cancel the number. But in the actual operation, for specific business departments, bank customer churn can be defined as the current specific business termination behaviour, and choose another enterprise instead.

According to the characteristics of customer churn, customer churn types can be classified into two categories. One is progressive, which Banks can easily identify and focus on, and the other is interruption type. The progressive type mainly shows that the number of recent business transactions, transaction types, financial, deposit and loan balance of customers are gradually reduced. And the latter type is manifested in the cessation of trading but the retention of customer accounts, and the sudden termination of business due to certain emergencies.

According to the causes of customer churn, the types of customer churn can be divided into the natural churn, competitive churn and negligent churn[8]. Natural churn corresponds to the above-mentioned interruption type of churn, which is not inevitable due to human factors, but within the elastic range. The competitive churn refers to the loss of customers caused by the price war or service war of competitors. The negligent churn is caused by the improper service of service personnel or the decline of enterprise image in the industry. The latter two are the types that the industry needs to focus on.

2.1.2. The Factors Influencing Bank Customer Churn

Customer churn is a normal phenomenon, but it needs practitioners to dig out the deep reasons. The reasons can be divided into two categories, including the reasons for the bank itself and the reasons for the customer. And the reasons for the bank itself can include the following aspects. First, the banking business handling is not comprehensive, the business level is relatively low, do not pay attention to corporate image. Second, the bank's backward management mode and method can hardly meet the market development. Third, the bank staff service attitude is relatively poor, work attitude is not correct. Finally, the bank internal staff turnover. These are all due to the bank's shortcomings are easy to cause the loss of customers[8,9].

From the customer-specific attribute analysis, previous studies have found that Lei Gang et al. summarized 12 relevant variables including gender, age and length of account opening time, including customer loss[10]. And Wang Weiqing et al analyzed 12 predictive variables that may have a significant influence on customer churn through survival analysis method and Cox proportional risk model and found 7 significant influencing factors, including average monthly deposit, average monthly consumption, maximum deposit balance, account type, and average monthly transaction [11]. Furthermore, Meng Xiaolian et al summarized 58 relatively comprehensive customer attribute factors and selected 11 important factors based on objective reasons[12]. And they also used the cross-table analysis technique in statistics to select 3 most important factors from the 11 characteristics to establish a regression model for prediction. However, the above studies made a prediction analysis from a single factor, while no factor classification was used to make a prediction. Therefore, the innovation of the research is to

ISSN: 2688-9323

classify the influencing factors of users according to the reasons of the bank churn, to achieve better prediction effect and get a more realistic explanation.

2.2. Industrial Application of the Machine Learning Algorithm

Scholars believe that there are two main categories of classification algorithms to effectively predict potential churn customers. The first category of methods is traditional classification methods, such as decision tree[13], logistic regression[14], Bayesian classifier and cluster analysis[15]. The main feature of this method is that it can process categorical data and continuous customer data, and has strong interpretability for the model. However, the generalization ability of the model built by this method is low. The second method is artificial intelligence classification method, such as an artificial neural network, self-organizing map[16], and evolutionary learning algorithm[17]. This kind of method can make up for the deficiency of the first kind, but it is difficult to determine the structure of the model.

Specifically, in the research of customer churn, the commonly used algorithms mainly include decision tree[8,18-20], support vector machine[21-23], and logistic regression algorithm[12]. The commonly used evaluation indexes are accuracy, precision and recall and F_1. From the research of algorithm improvement, the commonly used algorithm improvement can be divided into two aspects: one is to preprocess the data imbalance problem through statistical method or other algorithms and then use a common machine-learning algorithm to carry out experiments. The second is to filter the attributes by statistical methods. When using the logistic regression algorithm, 3 to 5 most important attributes will be filtered out for many times, and then the regression model will be established to deal with the noise problems caused by attributes.

From the previous research, it is found that decision tree, support vector machine and logistic regression algorithm are commonly used machine learning algorithms for bank customer churn prediction two classification problem. Therefore, the research uses decision tree algorithm as the main algorithm to identify the important controllable factors affecting customer churn, and then through the accuracy, precision and recall rate and f_1 four index comprehensive evaluation of the decision tree algorithm and support vector machine algorithm and the logistic regression algorithm.

3. Model

Decision tree learning is an inductive learning algorithm based on examples. The main method is to compare the irregular and disordered data to get the tree model to construct the corresponding classification node judgment. The information entropy of a single attribute is used to measure the importance of the attribute and other corresponding indicators to build the model. The commonly used decision tree algorithms are ID3.0, C4.5, and CART algorithm. One of the criteria of the C4.5 algorithm test feature is the gain ratio, which is a modification of the information gain index used in ID3.0 algorithm to reduce its deviation. However, CART algorithm is like C4.5 algorithm, except that the measurement indexes of each node in the decision tree generated by both algorithms are different. Therefore, the C4.5 algorithm is selected to predict bank customer churn in this paper. The specific algorithm process is as follows:

(1) If the node satisfies the stop splitting condition (all records belong to the same category or the maximum information gain is less than the threshold), it is set as a leaf node.

(2) The feature with the largest information gain ratio is selected for splitting.

(3) Repeat steps 1-2 until the classification is complete.

If S is regarded as a group of training samples, and S is divided into S_1, S_2, S_3, …, S_n has n subsets, S can be expressed as {S_1, S_2, S_3, …, S_n}. Where X contains n attributes, n_i is the number of

ISSN: 2688-9323

instances of the decision attribute. H (x) is the information entropy of subset, and |S| is equal to the number of middle samples of training sample S.

If i = {1, 2, 3, …, n}, then the number of samples belonging to C_i is freq (C_i, S), and the probability of a sample belonging to C_i is $\log_2$ (freq (C_i, s)) over |S|. And the following is the calculation process of index information gain ratio.

$$info(S) = -\sum_{n=1}^{N}\left(\frac{freq(C_i,S)}{|S|}log_2\frac{freq(C_i,S)}{|S|}\right) \quad (1)$$

$$info_x(S) = -\sum_{j=1}^{n}\left(\frac{|S|}{S}log_2 \times info(S_i)\right) \quad (2)$$

$$Gain\ (X) = info(S) - info_x(S_i) \quad (3)$$

$$H(X) = -\sum_{i=1}^{n} P_i log_2 P_i\ , P_i = \frac{n_i}{|X|} \quad (4)$$

$$Split\ Info_x(S) = -\sum_{i=1}^{N}\left(\frac{|S_i|}{S} \times log_2\frac{|S_i|}{S}\right) \quad (5)$$

$$Gain\ ratio = \frac{info(S)}{Split\ Info} \quad (6)$$

4. Empirical Analysis

4.1. Data Sources

The data used in the study is from a European bank churn open data set on the Kaggle website. The data set mainly includes customers from France, Germany and Spain, including 12 attributes and 10000 sample sizes. The specific meanings of attributes are as shown in Table 1. The purpose of the study is to predict whether customers will be churn according to the user attributes such as age, gender, credit, card information and so on. Therefore, the dependent variable is binary (if bank customer churn is recorded as 1, otherwise, it is recorded as 0).

4.2. Two or More Data Preprocessing

(1) Data screening. After understanding the 12 attributes in the dataset, it is found that the user ID and name issued in the sequence have no effect on the customer churn, so these two attributes are deleted. Among them, the user credit score, age, service life, deposit and loan situation, and the number of products used are very important indicators. For example, regarding the attribute of age, there will be great differences in the selection of bank switching between young and old people, and young people are more likely to switch banks. The gender and the geography of users may have an impact on the prediction of customer churn.

ISSN: 2688-9323

Table 1: Description of the attributes of the bank customer churn data set

Attribute	Description	Type
CustomerID	The user ID	String
Surname	The user name	String
CreditScore	The user credit score	Numerical
Geography	The user country or region	String
Gender	If the gender is male, the code is 1, otherwise, it is 0	String
Age	The user age	Numerical
Tenure	The total number of days used in the bank	Numerical
Balance	The customer deposits and loans	Numerical
NumOfProducts	The number of products used by users	Numerical
HasCrCard	If the user has a bank credit card, the code number is 1, otherwise, it is 0	String
IsActiveMember	If the user is active, code 1, otherwise 0	String
EstimatedSalary	User estimated revenue per year	Numerical

(2) Data transformation. Before the data transformation, this study first finds that there is no missing value in the data set through data query, so the data transformation is carried out directly. The country and region and gender of users in the data set are categorical variables. The categorical variable can be divided into binomial, ordered and unordered categorical variable. Different kinds of the categorical variable have different coding methods, and numerical feature preprocessing is also different. The gender variable can be considered as a dichotomous variable, and it is coded 0,1. It is considered that the geography is an unordered classification variable, and the dummy variable is used for coding. First, the three countries are represented by a combination of three numbers. For example, France changed from 0 to (1, 0, 0), Germany changed from 1 to (0, 1, 0), while Spain changed from 2 to (0, 0, 1), from the value of category to the combination of multiple variables. At the same time, to avoid falling into the "virtual variable trap", column 0 should be deleted.

(3) Selection of modelling data. According to the analysis of the reasons for the churn, the competitive churn and negligent churn caused by the preferential policies of competitors and the dissatisfaction of customers with the current service are the real concerns of the bank. So, identifying the customer churn caused by these two kinds of reasons is the bank's customers with retention value. For example, Li Xia[8] divided whether it is fixed-term, the number of deposits, the monthly business frequency and whether it is an investment into two attributes that lead to c competitive churn and negligent churn. Therefore, according to the cause of the churn and attribute static-dynamic analysis, the remaining 10 attributes after deletion are divided into two types: natural attributes, including Geography, Gender, Age, CreditScore and EestimatedSalary. And the other is competitive churn and negligent churn, which includes Tenure, Balance, NumOfProducts, HasCrCard and IsActiveMember. First, the important attributes in the data set can be identified through the decision tree model. Secondly, if it is found that the attribute belongs to the natural attribute category, this kind of attribute cannot be controlled by the bank but has a great reference effect. And if attribute belongs to the category of a competitive or negligent attribute, the bank staff should pay attention to the customer churn caused by these attributes, analyze the hidden deep reasons, and retain the customer through corresponding measures.

4.3. Result Analysis

First, this paper mainly uses decision tree algorithm to identify the important attributes that affect bank customer churn and classifies the important attributes according to the previous classification and then analyzes them according to different types of attributes. Second, because

ISSN: 2688-9323

support vector machine algorithm has a good prediction effect in small sample model prediction, this paper focuses on comparing the model effect of decision tree algorithm and support vector machine algorithm. At the same time, logical regression is also a common prediction algorithm for bank customer classification. Therefore, this algorithm is compared with a decision tree algorithm and support vector machine algorithm.

According to the literature research and experimental results, the ratio of the training set to test set is 2:8, that is, the training set length is 8000, and the test set length is 2000. Through the experiment, the depth of the tree is set to 3. According to the above data, the C4.5 decision tree algorithm is used to classify and get a decision tree as shown in Figure 1. According to Figure 1, the algorithm can identify four important attributes. And the most important attribute is CreditScore after data preprocessing. Since attribute Geography and Gender will be adjusted to the top of the data set list after data preprocessing, the most important attribute feature is CreditScore, followed by Balance, HasCrCard and Tenure. According to the previous classification, it can be found that the categories of the four important attributes are shown in Table 2. CreditScore is the most important attribute to judge the bank customer churn, but it is not the focus of the bank. While the bank customer churn caused by the other three attributes is the focus of the bank, and further analysis and corresponding retention measures should be taken.

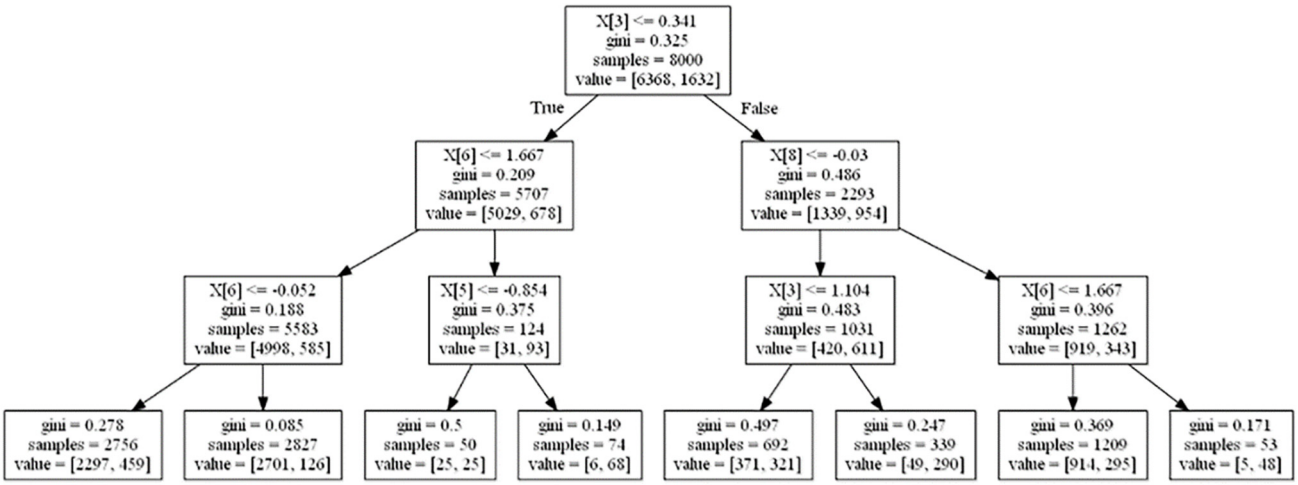

Figure 1: A decision tree obtained by the C4.5 algorithm

Table 2: Classification of identified important attributes

	Natural quality	Competitive and negligent attributes
Important attributes identified	CreditScore	Balance, HasCrCard, Tenure

The evaluation index is based on the confusion matrix and is expanded based on the original accuracy index which is often used. The precision, recall and comprehensive evaluation index F_1 score are added. In the 10000 samples of the actual public dataset, the proportion of customer churn is higher than that of the past. But in the construction model, the sample size of customer churn is 2037, and the sample size of customer non-churn is 7963, and the ratio of the former and the latter is about 1:4. Therefore, in the prediction results of the model, the cost of misidentifying bank customer churn is higher than that of misidentifying no customer churn. Therefore, the recall rate in the experiment is a very important indicator. Also, this study takes F_1 as the most important measurement index through a comprehensive evaluation of accuracy and recall index. As shown in Table 3, the values of each index of the three algorithms are relatively close, and the accuracy of the model constructed by the three algorithms is above 0.8,

ISSN: 2688-9323

which has high prediction performance. Among them, the support vector machine algorithm has the best performance in accuracy, recall and F_1, followed by the C4.5 decision tree algorithm, and finally the logical regression algorithm. But the C4.5 decision tree algorithm has the highest precision. However, the index recall and the index F_1 values predicted by the three algorithm models are all low, even for the model constructed by the best performance support vector machine algorithm, the value of the index F_1 is only 0.69. The index F_1 is a comprehensive indicator of the accuracy rate and recall rate, and is an important indicator of the performance of bank customer churn prediction. Therefore, the practical application ability of the model constructed in the study is still insufficient. Based on the analysis of the data set, the attributes may not be comprehensive enough to cover the important factors affecting the bank customer churn, and there may be some correlation between the attributes, From the analysis of the experimental process, the imbalance of data categories may have a certain impact on the experimental results. Future research can further classify the tag data to build a model.

Table 3: Three models of basic prediction results

Algorithm	Accuracy	Precision	Recall	F_1 score
C4.5	0.84	0.85	0.63	0.72
Support vector machine	0.86	0.83	0.65	0.73
Logistic regression	0.81	0.70	0.52	0.60

5. Conclusion

Based on the previous research, the attributes are divided into natural attributes, competitive and negligent attributes according to the reasons for churn. In this paper, the decision tree algorithm is used to identify the important attributes affecting the bank customer churn, in which the banking practitioners can focus on identifying the latter important attributes and specify the corresponding retention strategies according to this type of customers. At the same time, the decision tree algorithm, support vector machine and logistic regression algorithm are compared and analyzed. The overall comparative analysis shows that the support vector machine algorithm has good prediction effect in the construction of the model. Therefore, in further research, researches can first identify the important attributes by decision tree algorithm and then construct the model with a support vector machine algorithm. However, there are also some deficiencies in the research, such as incomplete data attributes and unbalanced processing of data categories, which have a great impact on the experimental results, which are the aspects that need to be improved in the future.

Acknowledgements

This work was partially supported by the grants from the National Natural Science Foundation of China (NSFC) (71802126), and a grant from the Shanghai Pujiang Program(18PJC060).

References

[1] Z.G. Liu: Bank customer relationship management and implementation in the network era, Financial BBS, (2002) No.9, p. 57-60. (In Chinese)

[2] B. Lin, B. Li: The current situation and strategy of customer relationship management in commercial banks, Financial Economy, (2008,) No.16, p. 82-84+90.

[3] R.G. Javalgi, T.W. Whipple, A. K. Ghosh, et al. Market orientation, strategic flexibility, and performance: implications for services providers, Journal of Services Marketing, Vol.19 (2005) No.4, p. 212-221.

ISSN: 2688-9323

[4] W.J. Reinartz, V. Kumar: The impact of customer relationship characteristics on profitable lifetime duration, Journal of Marketing, Vol.67 (2003) No.1, p. 77-99.

[5] L. Luo, L. Yu, H. Chen, et al. Deep mining external imperfect data for chest X-Ray disease screening, IEEE Transactions on Medical Imaging, Vol.39 (2020) No.11, p. 3583-3594.

[6] A. Nazari, M. Mehregan and R. Tehrani: Evaluating the effectiveness of data mining techniques in credit scoring of bank customers using mathematical models: a case study of individual borrowers of Refah Kargaran Bank in Zanjan Province, Iran, International Journal of Nonlinear Analysis and Applications, Vol 11 (2020), p. 299-309.

[7] M. Abdar, M.M. Zomorodi, R. Das, et al. Performance analysis of classification algorithms on early detection of liver disease, Expert Systems with Applications, Vol 67 (2017), p. 239-251.

[8] X. Li: ID3 applying to loss of bank clients, Computer Technology and Development, Vol 19 (2009) No.3, p. 158-160+167.

[9] L. Gao. Research on bank customer churn based on customer characteristics clustering, Finance, (2016) No.18, p. 346-346. (In Chinese)

[10] G. Lei, J. Ma: Customer churn prediction of commercial banks based on K nearest neighbour, Information and Computers (Theoretical), (2010) No.20, p. 122-123. (In Chinese)

[11] W.Q. Wang, R. Yao and C. Liu: The factors to influence the customer running-off of commercial banks—a study based on survival analysis method, Financial BBS, (2014) No.1, p. 73-79.

[12] X.L. Meng, S.Q. Cai, K.Q. Du, et al. Research on customer churn prediction model of commercial banks, Systems Engineering, (2004) No.12, p. 67-71. (In Chinese)

[13] J. Zhang, L. Ye: Local aggregation function learning based on support vector machines, Signal Processing, Vol 89 (2009) No.11, p. 2291-2295.

[14] H.S. Kim, C.H. Yoon: Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market, Telecommunications Policy, Vol 28 (2004) No.9-10, p. 751-765.

[15] M.C. Lee: Using support vector machine with a hybrid feature selection method to the stock trend prediction, Expert Systems with Applications, Vol 36 (2009) No.8, p. 10896-10904.

[16] A. Este, F. Gringoli and L. Salgarelli: Support vector machines for TCP traffic classification, Computer Networks, Vol 53 (2009) No.14, p. 2476-2490.

[17] Q. Wu, S.Y. Liu and L.Y. Zhang: Adjustable entropy function method for support vector machine, Journal of Systems Engineering and Electronics, Vol 19 (2008) No.5, p. 1029-1034.

[18] Y. Shi, J.J. Yue: Bank customer churn decision tree prediction algorithm based on data mining technology, Computer Knowledge and Skills, Vol 10 (2014) No.11, p. 2533-2536. (In Chinese)

[19] T.R. Wang: A review of the application of the decision tree method based on SAS, Financial Times, (2017) No.29, p.18-19. (In Chinese)

[20] Y.H. Yang, B.X Liu and Y.C. Wan: An analysis method of mobile communication customer churn based on decision tree, China Management Informatization, Vol 19 (2016) No.22, p. 70-71. (In Chinese)

[21] B.L. He: A study of the application of SVM in prediction about a decrease in bank's customers, Financial BBS, Vol 19 (2014) No.9, p. 70-74.

[22] T. Wang, Q. Cong, Y.L Shang, et.al. Customer churn prediction of the product-service system based on improved support vector machine, Modular Machine Tools and Automatic Processing Technology, (2018) No.5, p. 181-184.

[23] Z. Qin, Y. Zhao, B. Li, et.al. Support vector machine and its application in customer churn prediction, System Engineering Theory and Practice, (2007) No.7, p. 105-110.