ISSN: 2664-9640

Customer churn prediction based on different algorithms

Peipei Liu

Shanghai University, Shanghai, 201800, China

Abstract

As an important part of customer relationship management, customer churn management is becoming more attentive by enterprises. As an effective management method of customer loss, early warning can effectively reduce unnecessary customer loss and reduce the loss of enterprises to a certain extent by constructing early warning model, forecasting potential factors that result customer loss, thus, timely forewarning and taking corresponding retention measures. Banks have large customer base and wealthy customer data, and at the same time, they have a strong demand for early warning management of customer loss. Previous studies have mostly used a single method to predict user churn, but as algorithms are optimized, more possibilities should be explored. Therefore, Logistic regression model was used in this paper for modeling and prediction, and compared with the prediction effect of C5.0 decision tree model, the results showed that C5.0 decision tree model could effectively improve the prediction accuracy of the model, and the prediction effect was better than Logistic regression model, and it could forecast more accurately. It is of great significance for banks to strengthen customer management and enhance core competitiveness.

Keywords

Customer churn; Decision tree; Logistic regression; Predict.

1. Introduction

In recent years, with the development of China's social economy and the constant improvement of the financial system, policy banks, stock commercial banks, regional local banks, foreign banks, etc. have sprung up all over the country, and various banks have even set up branches around the community[1]. At the same time, the rapid development of information technology and Internet technology has provided fertilized soil for the birth of the Internet financial industry. With the emergence of online banking, mobile banking, WeChat wallet, Alipay, and a large number of Internet financial products, the geographical concept of financial services is gradually blurred, regional differentiation is gradually reduced, customers' choices of financial services and financial products are increasingly free and diversified, customers' dependence and loyalty to banking institutions are becoming less and less, and the problem of customer churn of banks is increasingly prominent[2]. If the platform wants to be able to develop, it must have strong user stickiness, and the value created by an old user is much higher than that of a new user, so it is very crucial to precisely predict the possible churned users and dig out the key factors affecting their loss.

Churned customers refer to customers no longer continue to participate in the original business, no longer repeatedly purchase or terminate the original use of the service, performance in the Internet products usually for a long time inactive, a long time did not generate payment, etc., the specific time node business will vary. In today's highly homogeneous product brand marketing stage, the competition between enterprises is mainly in the competition for users. Yet in their efforts to develop new users, companies tend to ignore the churning of existing customers, resulting in an awkward situation where new customers are constantly being added, while on the other hand, hard-earned users are being quietly lost. Philip Kotler, a contemporary

ISSN: 2664-9640

marketing authority, has shown that if a company can reduce its churn rate by 5%, it will increase its profits by 25%-85%. Generally speaking, the value of one old customer is equivalent to the value of three new customers. the loss of old customers to the enterprise is huge, and the cost of a new user is usually higher than the cost of keeping an old customer, and the cost of recalling churned users is even higher. user churn is a double loss of enterprise economy and reputation, so enterprises gradually pay more and more attention to user churn analysis.

Therefore, it is necessary to know enough about the behavioral trends of customers, find users with a tendency to churn in a timely manner, analyze the factors that cause churn. Only then can we make effective retention measures to avoid or reduce the churn of customers, which in turn can ensure the stability of users. Internet user data is huge, and with the widespread use of big data in life, the massive amount of user data has become a valuable asset for enterprises. Therefore, it is very meaningful to dig out potential regulations from the seemingly disorganized data and make predictions on user churn.

2. Literature Review

The current approaches on bank customer churn prediction research are mainly divided into two categories: the first category is the traditional data mining classification methods. For example, Chanda et al. used CART, TreeNet and C5.0 classification methods to predict the churned customers of commercial banks, and the results showed that the CART algorithm could better predict the possible churned customers [3]； Popovic and Basic used the fuzzy clustering Fuzzy C-Means algorithm for customer churn prediction models of retail banks; Lei Gang and Ma Jie used K-nearest neighbor classification for customer churn prediction[4]; Prasad and Madhavi studied the customer churn behavior of commercial banks using two classification techniques, CART and C5.0, respectively[5]. The second category is customer churn prediction models by statistical analysis methods [6]. For example, Xiao Lian Meng et al, used logistic regression analysis to build a customer churn prediction model and selected the variables affecting bank customer churn[7].

The essence of user churn is a two-category problem. From the perspective of machine learning, the solutions for such problems can be divided into two main categories: basic learning models and integrated learning models. In this paper, decision trees in basic learning models and other machine learning methods are selected for user churn prediction, and their principles and prediction accuracy values are compared and analyzed. So as to get an optimal prediction model, find the causes of customer churn, and propose preventive measures.

3. Theretical Background

3.1. Customer churn and Factor Analysis

Customers are highly essential assets of an enterprise, and as enterprises attach importance to the management of customers, customer relationship management comes into being [8]. This management is achieved through better understanding of customer needs and improving product and service quality for the purpose of increasing customer satisfaction. Customer churn is a part of customer relationship management. In the process of cooperation with a company, a customer may stop working with the company due to certain interests or for other reasons, which may have a negative impact on the company's market operations, a phenomenon known as customer churn.

There are two types of customer churn: voluntary and involuntary churn [9]. Voluntary churn refers to the churn of customers for their own reasons. Involuntary churn refers to the churn of

ISSN: 2664-9640

customers who are unable to continue to purchase the company's products or services due to some external reasons.

The reasons for customer churn vary for different fields. In the financial area, China's financial institutions are undergoing great changes with the deepening of the economic reform and the liberalization of the opening-up policy, coupled with the rise of online banking and the popular use of third-party payment platforms such as WeChat and Alipay, making the competition among banks fierce[10]. Increased competition has led to a higher rate of customer churn. By referring to the analysis of bank customer data by international and related field experts, Wang zhuqing et.al. took the customer information of a domestic commercial bank branch as a sample and derived 12 factors affecting bank customer churn through single factor analysis, and further used the stepwise discriminant method as well as PHRGE to find eight valid indicators, part of which were positively correlated with customer churn; part of which were negatively correlated[11]. Lu meiqin,et.al. analyzed the factors affecting VIP customer churn using the correlation coefficient test for the phenomenon of continuous churn of VIP customers in commercial banks, and the results showed that 15 factors such as age and the number of active current accounts have a significant impact on customer churn[12]. Sun mingwei operated logistic regression and principal component analysis to explore the factor model of bank user churn and give suggestions to prevent user churn[13].

The above study analyzed the prediction from a single factor and did not classify the factors for prediction, so the innovation of the study is to classify the user influencing factors according to the reasons of churn, so as to achieve a better prediction effect and get a more realistic explanation.

4. Model Introduction and Evaluation Metrics

4.1. Logistic regression

Although logistic regression has the word regression, logistic regression is a classification algorithm [14]. Logistic regression can perform multiple classification operations, but by the nature of the logistic regression algorithm itself it is more commonly used for binary classification. The first step is to introduce the formula that gives logistic regression.

$$Y(X) = \frac{1}{1 + e^{-(\theta^T X)}}$$

Among them, Y is the output value of the function, x is the function of the characteristic value, the function is to become a Sigmoid function, is a shape as S-shaped curve, when the curve near 0.5 has a faster growth rate, and the closer to the ends of the rate of growth is slower. From the figure, it can be seen that the value region of Y is (0,1), so we can treat all the objects corresponding to x with function values greater than or equal to 0.5 as positive samples, and similarly, the objects corresponding to x with function values less than 0.5 are treated as negative samples. In this way, the sample data can be classified into two categories. The logistic regression model is thus obtained, and the discriminant function formula is as follows:

$$F(x) = \begin{cases} 1 & Y(x) > 0.5 \\ 0 & Y(x) < 0.5 \end{cases}$$

However, the model cannot be operated because the parameter θ is not determined, so the parameter θ has to be estimated. whose cost function is formulated as follows:

$$cost(Y(X), y) = \begin{cases} -\log Y(X) & y = 1 \\ -\log(1 - Y(X)) & y = 0 \end{cases}$$

The value of y is equal to 1 when the contemporary value is small. When the surrogate value is larger, the value of y is equal to 0. Therefore, the total cost function is shown as follows:

ISSN: 2664-9640

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(y^i \log(Y(x^i)) + (1-y^i)\log(1-Y(x^i)))$$

When the total cost function is confirmed, to make the obtained model more consistent with the real model, so the cost value is required to be as small as possible. The loss function can be solved by the gradient descent method, and when the loss function converges, the solution that is found is the most consistent with the real set of solutions. The update process of the gradient descent algorithm for the parameter θ is formulated as follows:

$$\theta_j = \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$

A represents the step size (also called learning rate) and parameter j represents the jth feature of the sample. learning rate is one of the hyperparameters that most affects the performance. Therefore, it is important to choose the appropriate learning rate for the model. In this thesis, the learning rate is finally selected as 0.01 after several trials.

4.2. C5.0 Decision Tree

Decision tree algorithm is an artificial intelligence machine learning technique, which is gradually used in data mining because of its mature core algorithm, good data analysis capability and intuitive and easy to understand results. The principle of decision tree algorithm is to explore and learn the input and output variables to obtain the distribution and classification laws of the data when the variables take different values, and use these laws to predict the class values of the corresponding output variables based on the values of the input variables of the new data. It can be seen that the decision tree algorithm requires the data to include both input and output variables, and is a kind of supervised learning.

C5.0 was created through two processes, ID3 and C45, proposed by J.R. Quinlan in 1986 and 1993, respectively[15]. ID3 was flawed in the selection of branching properties, and C4.5 improved on it, but could only handle a limited number of data types. C5.0, as a commercial version of C4.5, does not differ significantly from C4.5 in the core of the algorithm, except that it is better in terms of memory usage and execution efficiency. C5.0, as a classification algorithm, can build a multinomial classification tree for samples with a large amount of data, and requires that the output variables must be subtypes, but the output variables can make numeric or numeric [16]. C5.0 has advantages in processing non-numerical data, insensitive to missing data with good robustness, and the output rules are intuitive and easy to understand, but it is not strong in handling discrete continuous attributes. the core concepts in C5.0 are information entropy and information gain. The attributes of the input data include the attributes of the judgment pair i and the attributes of the classification object. Through the information entropy method, the information gain rate of the attributes of the judgment object is compared, and the attributes with large information gain rate are classified, and finally a judgment tree is formed recursively.

4.3. Evaluation Indicators

The confusion matrix reflects the prediction effect of the model and is the basis for constructing the model evaluation index [17]. The confusion matrix of the customer churn model prediction results is shown in Table 1. Table 1 shows the classification of the customer churn model prediction results according to the real situation and predicted situation on two dimensions. The true-case dimension refers to the actual churn of customers in the data used for verification, and the predicted-case dimension refers to the customer churn model's prediction of future churn. According to Table 1, the actual total number of lost customers can be calculated as P=TP+FN， Total actual non-churning customers can be calculated as N=FP+TN.

ISSN: 2664-9640

Table 1: Confusion Matrix

	Prediction	
	Churn	Non-Churn
Churn	TP	FN
Non-Churn	FP	TN

The recall rate represents the number of correctly identified churned customers as a percentage of the total number of actual churned customers and is calculated as:

$$Recall = \frac{TP}{(TP + FN)} = \frac{TP}{P}$$

The churn prediction accuracy rate represents the proportion of the number of correctly identified churned customers to the total number of customers predicted to be churned, and is calculated by the formula:

$$Precision = \frac{TP}{(TP + FP)}$$

Overall prediction accuracy of the model:

$$Accuracy = \frac{(TP + TN)}{(P + N)}$$

In the process of modeling data mining classification models, the above-mentioned model evaluation metrics are usually considered as the basis for model selection and judging criteria. However, in the case of unbalanced data sets, the classifier tends to predict most classes correctly, while the prediction accuracy of a few classes is poor, and the general classification evaluation rules such as accuracy alone can no longer effectively measure the model prediction ability. Therefore, many scholars have introduced evaluation criteria such as F-measure, accuracy, and ROC curve. Among them, the F-degree is obtained based on the confusion matrix, which is the summed mean of recall and precision, and is defined as follows:

$$F = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$

In this paper, we evaluate the prediction capability of the model by considering the churn prediction accuracy, recall rate, and F-measure.

5. Experimental design and analysis

5.1. Data Source

The dataset used in this paper is from a public dataset on kaggle, containing data from users in three countries. The dataset contains a total of 10,000 data and 12 feature attributes, of which three are numerical features (Surname, Geography, Gender) and the remaining nine are character features. The purpose of the study is to predict whether a customer will churn based on user attributes such as age, gender, credit, and card information, so the dependent variable is a dichotomous variable, with churn noted as 1 and non-churn noted as 0. This dataset is stored in CSV format.

ISSN: 2664-9640

5.2. Data Process

Data pre-processing is one of the essential operations before model training, and although this process can be time-consuming and tedious, it is a very important part of the process. The quality of the data directly affects the prediction and generalization ability of the model. The main point of data pre-processing is to analyze some intrinsic characteristics of the original data. In real data, the data is often imperfect and may contain many missing values, outliers and other factors that are not conducive to model training, and data preprocessing is to analyze and process various abnormal data in a corresponding way to get clean and useful data for model training, which is conducive to improving the performance of the model.

5.2.1. Data Type Conversion

By looking at the specific type of data and converting the inappropriate data types, the final processing result is shown in table 2:

Table 2. Data Type

	Colum	Non-Null Count	Dtype
0	RowNumber	10000 non-null	Int64
1	CustomerId	10000 non-null	Int64
2	Surname	10000 non-null	object
3	CreditScore	10000 non-null	Int64
4	Geography	10000 non-null	object
5	Gender	10000 non-null	object
6	Age	10000 non-null	Int64
7	Tenure	10000 non-null	Int64
8	Balance	10000 non-null	Float64
9	NumOfProducts	10000 non-null	Int64
10	HasCrCard	10000 non-null	Int64
11	IsActiveMember	10000 non-null	Int64
12	EstimatedSalary	10000 non-null	Float64
13	Exited	10000 non-null	Int64

5.2.2. Process of missing-value data

The missing value cases of all variables of the dataset used in this paper were visualized and analyzed as shown in table3 3. The statistical missing value information reveals that the dataset used in this paper has no missing value cases, so no operation is performed.

Table3.Missing value processing

Colum	Missing-value Number
RowNumber	0
CustomerId	0
Surname	0
CreditScore	0
Geography	0
Gender	0
Age	0

ISSN: 2664-9640

Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0

5.2.3. Data normalization

Since character-based data cannot be recognized in the algorithms of machine learning to the extent that the data cannot be transmitted, it is necessary to change their data type to numeric for the character-based features of the data in this paper in order to continue the training. Gender The variable is a dichotomous variable, which is encoded with 0, 1. The variable of user's country or region is an unordered discrete categorical variable, and one-hot encoding is used to convert it using the one-hot function that comes with the pandas-library, and the results are shown in table 4 below.

Table4.Results of data normalization

Geography	[0 2 1]
Gender	[0 1]

5.2.4. Data Balancing

In most of the datasets used to predict user churn, there is a distinctive feature - class imbalance. Because the vast majority of machine learning algorithms are based on probabilistic and statistical studies, they often result in these classification algorithms favoring classes that are numerically overwhelming in the dataset. On the data set, the proportion of churned users was counted using the visualization tool matplotlib, and a pie chart was drawn as shown in Figure 1. It can be found that the proportion of churned users is only 20.37%, while the proportion of non-churned users is 79.63%. The distribution of normal and churned users in this sample is unbalanced (as shown in Figure 2), and it is easy to ignore the small number of churned customers when modeling, and the model appears to be lopsided.

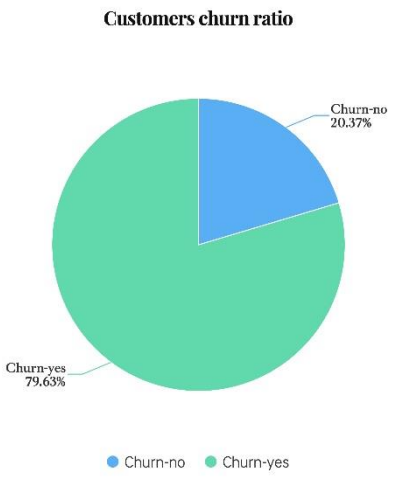

Figure 1. Customers churn ratio

ISSN: 2664-9640

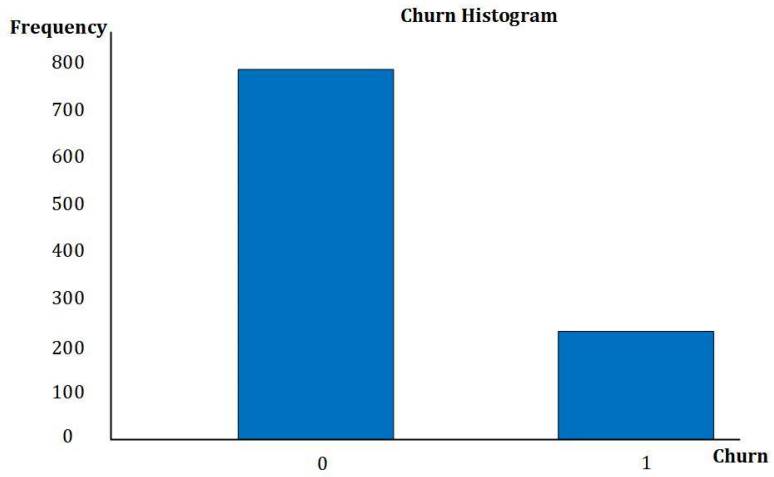

Figure 2. Customers churn number

When data sets with imbalances are used in the paper for classification algorithms, it can lead to inaccurate results, and to eliminate this effect, it is necessary to perform class equalization on the data to facilitate model training and improve the accuracy of the model. In this paper, a downsampling scheme is used for processing [13], the results are shown in table 5.

Table 5. Results of balancing

Percentage of Churn_no	0.5
Percentage of Churn_yes	0.5
Percentage of Churn in resampled data	4074

5.3. Feature Selection

In general, when performing a machine learning task, the resulting prediction value is based on the features of the data sample. If the data has too few features, the result may cause the model to be ineffective, and then it is necessary to consider adding features to the original features. In fact, there are often too many features, and the features need to be filtered to reduce the number of less relevant or irrelevant features. The variables with weak correlation were eliminated, and variables that had no effect on the dependent variable were also removed in this paper. For example, serial number, user name attribute, and after the above data preprocessing, the structure of the final data is shown in Table6 below.

Table 6.Results of feature selection

RowNumber	Surname	CreditScore	IsActiveNemer	EstimatedSalary	Exited
1	Hargrave	619	1	101348.88	1
2	Hill	688	1	112542.58	0
3	Onio	582	0	113931.57	1
4	Boni	699	0	93826.63	0
5	Mitchell	858	1	79884.10	0

5.4. Modeling and experimental results

In this paper, according to the related literature, the training set and test set are divided in the ratio of 7:3 by using sklearn's train_test_split method, that is, 7000 data are used for training and 3000 data are used for testing.

ISSN: 2664-9640

5.4.1. Logistic Regression

According to the parameter penalty principle of regularization, the optimal parameter of the final obtained model is 0.1. Then the logistic regression method in sklearn.linear_model is used to model the logistic regression model, and the accuracy of the logistic regression model finally verified on the test set is 0.807.

5.4.2. Decision Tree

Based on the pre-processing of the data, a decision tree model is established here, and the depth of the tree is set to 4 through the experiment, and a decision tree is obtained after classification by C5.0 decision tree algorithm for the above data[18].

According to the decision tree, the algorithm identifies 4 important attributes. The most important attribute is the credit score, followed by the "deposit and loan status" and "possession of the bank's credit card", and finally the "usage period". The most important attribute is credit score, which is the most important attribute for determining customer churn, but not the most important one for banks, while the other three attributes are the most important ones for banks to focus on and require further analysis and appropriate retention measures.

Table 7 shows the accuracy, precision, recall, and overall evaluation metrics under two different algorithms $f1$[19]， The comparison shows that the decision tree algorithm outperforms the logistic regression algorithm. Also Figure 3 shows the ROC curves under different algorithms, and observing the two curves also reveals that the decision tree algorithm outperforms the logistic regression algorithm.

Table 7. Results of prediction

	Accuracy	Precision	Recall	f1
C5.0	0.836	0.8041	0.7994	0.8016
logistic regression	0.807	0.7890	0.7836	0.7863

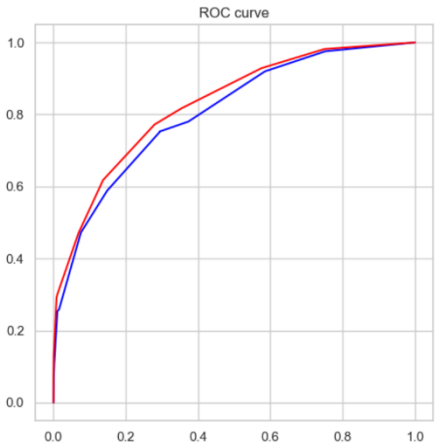

Figure3.Comparison Results of ROC

ISSN: 2664-9640

6. Conclusion

In this paper, 2 prediction models are used, and the same training set is selected and predicted on the software respectively, and the prediction results of each prediction model are compared to derive the corresponding prediction accuracy, and the final comparison results show that the C5.0 decision tree model has the best accuracy and better prediction effect. It can help to predict bank customer churn more accurately and take measures to retain in advance. Compared with the traditional manual identification methods, the machine learning-based model has higher prediction efficiency and brings more gain to the bank, so it is necessary for bank staff to check regularly based on users' consumption information to ensure the stickiness of their important customers.

In online era, more people are getting used to using the Internet. This phenomenon sends a signal to enterprises that they cannot ignore the influence brought by the Internet and should focus their attention on the study of online customer behavior in the future. First, network customers belong to non-contractual customers, usually the relevance of network customers is weak, the churn rate is high, and it's difficult for enterprises to accurately determine the potential churn behavior of customers, so it's necessary to determine which factors play a decisive role in the identify potential churn customers. Secondly, network customer data is generally large and mixed. With the continuous development and maturity of big data, the analysis of customer behavior has become easier, and with the advantages shown by deep learning methods in recent years, deep learning methods can be effectively used in the process of predicting network customer churn in the future.

References

[1]. Xia, G.E., Ren. , A Review of Customer Churn Research. Academic Forum, 2018.

[2]. Jin Xin, Qian. , An empirical study of retail customer churn prediction in commercial banks. Entrepreneur World, 2010.

[3]. Wadikar, D., Customer Churn Prediction. Computer Sciences Commons, 2020.

[4]. Wadikar, D., Customer Churn Prediction. Computer Sciences Commons, 2020.

[5]. Madhavi, V., et al., An overview on research trends in remediation of chromium. Research Journal of Recent Sciences, 2013. 2277: p. 2502.

[6]. Rad, H.M.a.A., Customer Churn Prediction in Irancell Company Using Data Mining Methods. EasyChair Preprint, 2020.

[7]. Yang, S.K., Gui, J.. Duan. Duan, A comparative analysis of data mining methods for customer churn prediction. Computer Learning, 2010.

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

[9]. Amin, A., et al., Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research, 2019. 94: p. 290-301.

[10]. Osmanoğlu, Ö.Ç.U.Ö., Comparing to Techniques Used in Customer Churn Analysis. Journal of Multidisciplinary Developments, 2019.

[11]. Wang Weiqing; Yao Yao; Liu Cheng, Factors influencing customer churn in commercial banks - A study based on survival analysis method, in Financial Forum. 2014. p. 73-79.

[12]. Lu Meiqin; Wu Chuanwei, Study on the prediction of VIP customer churn in commercial banks, in Journal of Fujian Business School. 2018. p. 31-36.

[13]. Sun Pengwei, Churn early warning analysis of commercial bank customers, in Chongqing University. 2014.

ISSN: 2664-9640

[14]. Zhang Zongji, Building a model to predict credit card application results based on logistic regression algorithm. Mechatronics Applications, 2021.

[15]. Quinlan, J.R. Combining instance-based and model-based learning. in Proceedings of the tenth international conference on machine learning. 1993.

[16]. Shirazi, F. and M. Mohammadi, A big data analytics model for customer churn prediction in the retiree segment. International Journal of Information Management, 2019. 48: p. 238-253.

[17]. Höppner, S., et al., Profit driven decision trees for churn prediction. European Journal of Operational Research, 2020. 284(3): p. 920-933.

[18]. Feng-Ying Dai, Data mining based early warning analysis of telecom customer churn, in Dalian University of Technology. 2020.

[19]. Ying Li, Wu, A logistics regression algorithm-based churn prediction model for securities customers and its application, in Financial e. 2013. p. 65-67.7.