CSCanada
Management Science and Engineering
Vol. 12, No. 3, 2018, pp. 51-56
DOI:10.3968/10816

ISSN 1913-0341 [Print]
ISSN 1913-035X [Online]
www.cscanada.net
www.cscanada.org

Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm

ZHUANG Yayun[a],*

[a]South China University of Technology, Guangzhou, China.
*Corresponding author.

Received 21 July 2018; accepted 6 September 2018
Published online 26 September 2018

Abstract

In recent years, with the development of Internet technology, the market competition is fiercer, the cost of acquiring new customers is increasing, and the cost of maintaining old customers is far less than the cost of acquiring new customers. Most companies are trying to market precisely through customer segmentation in order to reduce the rate of customer churn. Aiming at the customer characteristics of social network e-commerce, this paper builds a customer value model that integrates the value of social network to help companies subdivides the customer accurately. Then we use the machine learning algorithm XG-Boost to predict the churn of customers before and after the subdivision. The research found that the prediction accuracy is higher after customer segmentation. In addition, the XG-Boost algorithm is more advantageous than other algorithms.

Key words: Customer value; XG-Boost algorithm; Social network value

Zhuang, Y. Y. (2018). Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm. *Management Science and Engineering*, *12*(3), 51-56. Available from: URL: http://www.cscanada.net/index.php/mse/article/view/10816
DOI: http://dx.doi.org/10.3968/10816

INTRODUCTION

In recent years, with the development of the Internet technology, the market competition is fiercer, and the cost of acquiring new customers is increasing. Moreover, the cost of maintaining old customers is far less than the cost of acquiring new customers. Therefore, many companies have been paying attention to customer retention issues. In order to reduce the rate of customer churn, more and more companies are pursuing precision marketing through customer segmentation.

Customer segmentation refers to the classification of customers according to their attributes, behaviors, needs, preferences and values in a clear strategic business model and in a specific market. Dividing customers into different groups according to certain criteria and marketing them with targeted manners can improve customer loyalty and reduce customer churn. Yang believes that customer segmentation refers to the classification of customers based on a combination of customer value, needs and interests in a clear strategic business model and focused market (Su, 2016).

Customer churn refers to the company's original customers, who do not purchase of their goods or services any more, switch to the purchase of their competitor's goods or services. Retaining measures for valuable potential customers is called customer retention.

The study of customer churn prediction has gone through the following stages:

The first stage is based on traditional statistical predictions, including Bayesian classification, clustering, logistic regression and decision trees. Zhang used the C5.0 decision tree algorithm to establish a prediction model, and used the actual business data of China Post to conduct empirical research, which showed that the model has a high hit rate and coverage (Zhang, 2015). Ye established the customer churn model through the study of the structure and parameters of the Bayesian network, and obtained more accurate prediction results than the decision tree method (Ye, 2005). The second stage is based on traditional artificial intelligence prediction. The main methods are artificial neural network and self-organizing map. Tian et al. proposed a neural network-based customer churn prediction model, which selects the analysis variables according to the empirical value

of the industry experts, and calculates the weight of the analysis variables through the neural network to predict the customer churn trend (Tian & Qiu, 2007). The third stage is the prediction based on statistical learning theory. The main representative method is support vector machine (SVM). Qian et al. introduces cost-sensitive learning theory, establishes a telecom customer churn prediction model based on improved SVM, and incorporates different misclassification costs into the modeling process to effectively improve the predictive performance (Qian, He, & Wang, 2007). According to the characteristics of unbalanced data of customer churn, the CW-SVM algorithm with different weight parameters is proposed, which improves the classification accuracy through bank credit customer data test (Ying, 2007). The fourth stage is a prediction based on integrated learning and bionics algorithms. Wu first uses clustering to identify high-value customers, then applies improved SMOTE to process e-commerce customer imbalance data, and finally uses Ada-Boost algorithm to perform churn prediction. This paper get conclusion that after customer segmentation, the model predicts better results. (Wu, 2017)

However, in current study, most papers focus on the loss of contractual customers, such as the customers from telecommunications and finance. There are few papers focusing on non-contractual customers such as e-commerce customers. Moreover, these researches generally deal with the individual characteristics of customers but rarely consider the interaction effects between customers.

In response to the above problems, this paper proposes a new value model based on the traditional model RFM. We firstly integrate the features of e-commerce customers under social networks into the RFM, in order to identify high-value customers. Then, we use the XG-Boost algorithm to perform the churn prediction. This approach provides new insights into studying customer behavior when it turns to non-contractual customer and provides a more accurate method of customer churn prediction.

1. RELATED WORK

1.1 RFM Model

In a certain competitive market, the expected benefits that created by company's target customers are called customer value. The traditional RFM model is an important tool to measure customer value and customer profitability. R (Recency) indicates the recently consumption. F (Frequency) indicates the consumption frequency. M (Monetary) indicates the consumption amount.

1.2 XG-Boost Algorithm

XG-Boost (Extreme Gradient Boosting) is a scalable machine learning algorithm for Tree Boosting. It uses the first derivative and the second derivative in the optimization. At the same time, the algorithm uses the tree model complexity as a regular term in the objective function to avoid overfitting.

Its objective function is as follows:

$$J(f_i)=\sum_{i=1}^{n} L(y_i, y_i^{t-1}+f_i(x_i)+\Omega(f_i)+C) \quad (1)$$

According to the Taylor expansion:

$$f(x+\Delta x)\approx f(x)+f'(x)\Delta x+\frac{1}{2}f''(x)\Delta x^2 \quad (2)$$

After derivation, we can get:

$$g_i=\frac{\partial L(y_i, y_i^{t-1})}{\partial y_i^{t-1}}, h_i=\frac{\partial L(y_i, y_i^{t-1})}{\partial y_i^{t-1}} \quad (3)$$

The decision tree complexity calculation formula is as follows:

$$\Omega(f_i)=\gamma T_i+\lambda\frac{1}{2}\sum_{j=1}^{T} w_j^2 \quad (4)$$

Substitute equations (2), (3), and (4) into equation (1):

$$J(f_i)\approx\sum_{i=1}^{n}[L(y_i, y_i^{t-1})+g_i f_i(x_i)+\frac{1}{2}h_i f_i^2(x_i)]+\Omega(f_i)+C=$$

$$\sum_{j=1}^{T}[\left(\sum_{i\in Ij} g_i\right)w_j+\frac{1}{2}(\sum_{i\in Ij} h_i+\lambda)w_j^2]+\gamma T+C \quad (5)$$

By solving the equation (5), we can get the optimal solution as follows:

$$w_j^*=-\frac{\sum_{i\in} I_j g_i}{\sum_{i\in} I_j h_i+\lambda} \quad (6)$$

$$J(f_t)=-\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in} I_j g_i)^2}{\sum_{i\in} I_j h_i+\lambda}+\gamma T \quad (7)$$

Use equation (7) to find the optimal tree structure.

The gain calculation formula of Decision tree is as follows:

$$Gain(\Phi)=\frac{1}{2}\left[\frac{(\sum_{i\in} I_L g_i)^2}{\sum_{i\in} I_L h_i+\lambda}+\frac{(\sum_{i\in} I_R g_i)^2}{\sum_{i\in} I_R h_i+\lambda}-\frac{(\sum_{i\in} I g_i)^2}{\sum_{i\in} I_R h_i+\lambda}\right]-\gamma \quad (8)$$

We can find the optimal separation scheme according to equation (8).

This paper also selects Logistic Regression (LR), Support Vector Machine (SVM), and BP Neural Network (BP), which are commonly used in customer churn prediction research, to compare with XG-Boost algorithm.

1.3 Model Evaluation Index

We use the confusion matrix of Table 1 as the evaluation index of the model.

Table 1
The Confusion Matrix

Status	Predicted as loss	Predicted as non-loss
Actual loss	TP	FN
Actual non-loss	FP	TN

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$Lift=\frac{TP/(TP+FP)}{TP+\frac{FN}{TP}+TN+FP+FN} \quad (10)$$

Accuracy is the ability of the model to identify the customer rightly. And the Lift rate indicates the improvement of predictive ability when using the model compared to not using the model.

AUC (Area Under Curve) is defined as the area enclosed by the ROC curve and the coordinate axis. When

you randomly select a positive sample and a negative sample, the probability that the current classification algorithm ranks the positive sample before the negative sample based on the calculated Score value is the AUC value. Therefore, the larger the AUC value, the more likely the current classification algorithm is to rank the positive samples in front of the negative samples, that is, to be better classified. This paper uses Accuracy, Lift coefficient and AUC value as indicators to measure the performance of the algorithm.

2. PROPOSED METHOD

2.1 Social Network Value

Currently, traditional e-commerce companies are gradually transforming to adapt to the market environment. This new type of e-commerce model based on social network is called social e-commerce. In this situation, traditional models that measure customer value are no longer appropriate. Studies have shown that the value of customers under social networks is not economic utility or functional value, but social value and emotional value.

More and more companies have discovered that customer self-propagation and word-of-mouth communication between themselves can bring more flow and revenue. Therefore, most companies are willing to stimulate customers' social behavior with low price. In this model, customers with low purchases and low profit margins are not necessarily low-value customers for the company. Companies also need to consider the impact of this customer on other customers, namely the social network effect. Therefore, companies need a new set of models for identifying high-value customers that are suitable for the current online shopping environment.

This paper takes a domestic e-commerce platform as the research object. We draw on the quantitative model of social network influence to quantify the network influence of social e-commerce customers. Then, we integrate the customer's social network influence I to build a new customer value model RFMI based on the RFM model. Finally, we use the segmented customer as input to the forecasting model to finish the churn prediction.

When customers interact in a social network, a lot of data is generated. By analyzing these behavioral data, we can measure the impact between customers. Goya Amit et al. used the log information to calculate the influence of the user and the behavior itself.

$$\text{infl}(u) = \frac{|\{a|\exists v,\Delta t: prop(a,v,u,\Delta t) \wedge 0 \leq \Delta t \leq$$ (11)

In detail, 'u' and 'v' represent different customers, 'a' represents action, Δt represents the time interval between actions, $\tau_{v,u}$ is a time constant, $prop(a,v,u,\Delta t)$represents the propagation of actions between customers, and A_u represents the number of actions generated by customer 'u'. The formula describes the influence of the customer 'u' as the proportion of behaviors that are propagated within a certain time interval to the total number of behaviors generated by customer 'u'.

Due to the limitations of data acquisition, this paper focuses on customer sharing behavior. We define that, when the links shared by customer u are opened by other customers, the sharing behavior of customer u is spread. Therefore, A_u indicates the number of sharing actions generated by customer u during within a certain time. Considering that the customer's personal credit rating will affect his or her sharing effect, we add the customer's personal credit rating factor to the above formula. The higher the customer's credit rating, the greater the impact.

$$\text{infl}(u)' = x\text{infl}(u)$$ (12)

2.2 Overview of the Proposed Method

According to the RFMI value model proposed in this paper, the customer data can be processed. Then the k-means clustering algorithm can be used to subdivide the customer, in order to identify customers with high network value. Next, segmented customer data is used as input to the forecasting model for churn prediction. The overall predictive model framework for this paper is shown below:

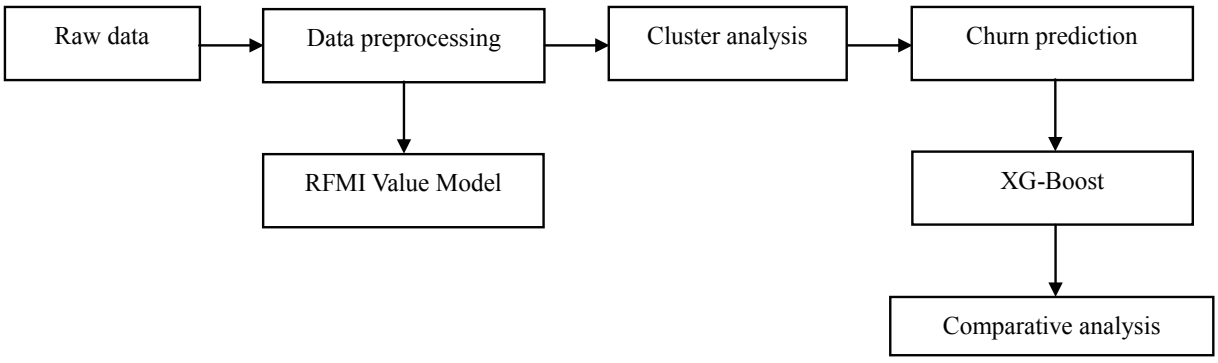

Figure 1
Overview of the Proposed Method

3. EMPIRICAL RESEARCH

3.1 Datasets

The research data comes from an e-commerce platform in China. The empirical research time range is from January 1, 2018 to October 30, 2018. The observation period is from January 1 to June 30, and the verification period is from July 1 to October 30. We define customers who have purchased during the observation period and haven't purchased during the verification period as a lost customer. Firstly, select the users who have purchasing behavior during the observation period. And then extract other relevant data based on this condition. After initial cleaning, 3,792 effective customers were obtained.

High Value Customer Identification

According to the value model RFMI proposed, the corresponding data is extracted and calculated:

'R' is the number of days since the customer's last purchase from January to June (the current time is July 1, 2018);

'F' is the cumulative number of orders purchased by customers from January to June;

'M' is the cumulative consumption amount of the customer from January to June;

'I' is the network effect value of each customer, which is calculated by formula (12).

Each attribute is normalized by the Min-Max method.

The above four attribute values are used as input to the clustering algorithm (K-means). Refer to Kaufman et al. for the outline width of the product, take the best cluster number 5 for k.

The result of clustering is shown in the following table 2 and figure2.

Table 2
The Result of Clustering

kind	R	F	M	I	count
1	0.0426	0.1894	0.2195	0.0255	156
2	0.3260	0.0110	0.0163	0.0875	1022
3	0.8703	0.0022	0.0075	0.1199	431
4	0.0863	0.0234	0.0277	0.0683	1525
5	0.5988	0.0058	0.0116	0.1214	658

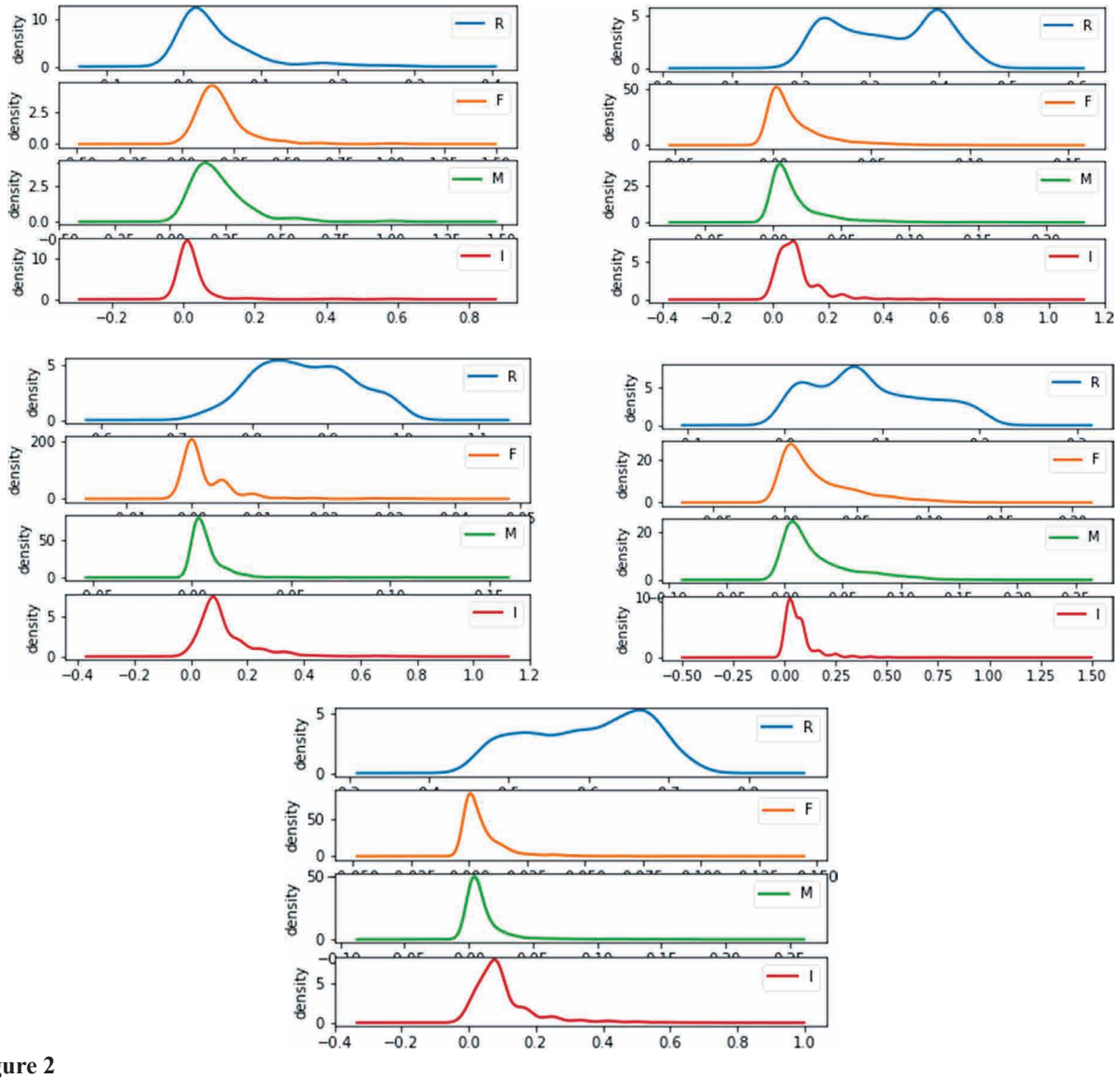

Figure 2
Overview of the Proposed Method

According to data and images from table2 and figure2, each customer segment has significantly different characteristics. We can see that, the type 1 is the smallest on the on R and I attributes and the largest on F and M attributes; the type 3 is the largest in the R attribute, and the F and M attributes are the smallest; the type 5 is the largest on the I attribute and smaller on the F and M attributes; the values of the type 2 of and the type 4 of are in the middle of all groups; compared with the type 4, the type 2 of has higher R and I attributes and smaller F and M attributes.

Analyzing the numbers and the contribution of consumption amount of these 5 groups (Table 3), it can identify that type 1 is an important customer group of the company, with its customer accounted for only 4%, but contributed 30% of the purchase amount; the type 5 is social high-value customer with the low purchase amount but the greatest influence on social network; the type 3 is the low value customer of the company; the type 2 and the type 4 customers are ordinary customers of the company.

Table 3
The Proportion Details

Kind	Number	Proportion of People	Proportion of Amount
1	156	4%	33%
2	1022	27%	16%
3	431	11%	3%
4	1525	40%	41%
5	658	17%	7%

3.3 Customer Churn Forecast

Conduct customer churn prediction research on data before and after segmentation. According to relevant researches and actual data, we determine the key attribute set {gender, region, membership level, purchase amount, number of repeated purchases, last purchase time} as input to the forecasting model. The customer churn status is the target variable:1 means churning customers, and 0 means non- churning customers. Due to the small sample data set, we use cross-validation for model prediction.

Firstly, we make churn predictions for customers before the segmentation. The results are shown in Table 4. We can see that XG-Boost performs better than other three types of algorithms. For the subdivided customers, we select customer type 1, 2, and 5 to make predictions. The results are as shown in the figure 3. Compared with the pre-segment, the major indicators have improved. Customer segmentation can identify the high-value customer segments effectively. It improves the accuracy of forecasting for lost customers and non-loss, so as to improve the overall forecast.

Table 4
The Operation Result (before subdivision)

Algorithm	Accuracy	AUC	Lift
BP	0.86	0.84	2.31
LR	0.85	0.87	2.13
SVM	0.87	0.86	2.55
XG-Boost	0.88	0.89	3.12

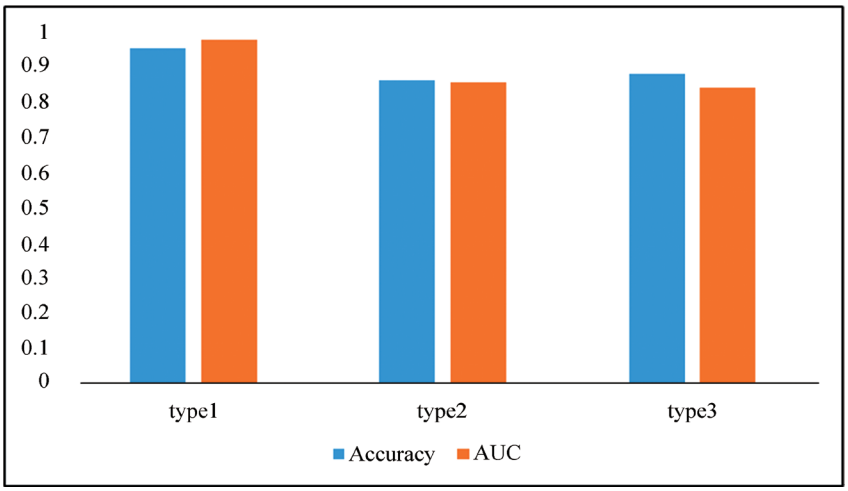

Figure 3
The Operation Result (after subdivision)

CONCLUSION

Nowadays, in order to tap the value of customers' social networks, e-commerce companies use various marketing methods such as low prices to attract customers. Considering the value of the customer's social network, this paper proposes the RFMI value model based on the RFM in order to segment customers. This value model can help companies identify customers with high social network value. For these high-value network customers, companies can conduct targeted marketing and increase their spending amount and frequency. Then we use the machine learning algorithm XG-Boost to predict the churn of the customer groups before and after the segment. Finally, we compare the algorithm proposed with other algorithms.

The experimental results show that the forecast results are more accurate after segmenting the customers. At the same time, the XG-Boost algorithm performs better than

other prediction algorithms. The amount of data generated by the e-commerce platform is large. And the XG-Boost algorithm can improve the efficiency of the algorithm in big data processing through multi-threading and data compression. The new customer value model proposed in this paper, which integrates the value of social network, can help social network e-commerce companies identify customers with high social influence. Simultaneously, predicting the customer's churn status helps companies to build a better strategy of precision marketing.

REFERENCES

Cheng, H., & Fan, C. J. (2018). Application research of neural network in risk prediction of e-commerce enterprises' customer loss in China. *Chinese Collective Economy,* (17), 54-56.

Cheng, Y. S., & Zou, H. (2018). Evaluation of bank credit risk based on random forest RFM model. *Journal of An Qing Normal University, 24*(03), 34-37.

Du, K., Deng., J. W., & Chen., J. H. (2018). Research and application of improving RFM model in real estate customer segmentation. *Computer Knowledge and Technology, 14(*19), 243-245+251.

Feng, X., Wang, C., Liu, Y., Yang, Y., & An, H. G. (2018). Research on customer churn prediction based on comment emotional tendency and neural network. *Journal of China Academy of Electronics Science, 13*(03), 340-345.

Huang, X., Liu, X., & Liu, J. (2019). A new research on impact evaluation algorithm of Weibo user. *Computer Engineering,* 1-7.

Huang, J. (2018). A Comparative Study of Social E-Commerce and Traditional E-commerce. *Economic and Trade Practice,* (23), 188-189.

He., Y., J (2017). Social network user discourse influence value fractal dimension method. *System Engineering, 35*(02), 145-152.

Liao, J. F., Chen, T. G., & Chen, X. (2018). Research on social media user flow based on customer churn theory. *Information Science, 36*(01), 45-48.

Liu., D. (2018). R*esearch on civil aviation passenger value and travel prediction model based on social network*. China Civil Aviation University.

Lu, N., Liu, X. W., & Lee, L. (2018). Research on customer value segmentation of online shop based on RFM. *Computer Knowledge and Technology, 14*(18), 275-276, 284.

Qian, S. L., He, J. M., & Wang, C. L. (2007). Telecom customer churn prediction model based on improved support vector machine. *Management Science,* (01), 54-58.

Ren, H. J., & Xia, G. E. (2018). Summary of customer churn research. *Chinese Business Theory*, (32), 166-167.

Shao, D. (2016). *Analysis and prediction of insurance company's customer loss based on BP neural network*. Lanzhou University.

Su, M. (2016). *Customer relationship management* (p.15). Higher Education Press.

Tian, L., Qiu, H. Z., & Zheng, L. H. (2007). Modeling and implementation of telecom customer churn prediction based on neural network. *Computer Application*, (09), 2294-2297.

Weng, J. (2018). *Implementation of telecom customer churn prediction system based on big data mining*. Nanchang University.

Wang, R., Zhou, Y., Liu, Y. L., & Jue, S. P. (2018). Research on MOOC learner segmentation based on improved RFM model. *Computer Products and Circulation,* (07), 239-240+268.

Wang, T., Cong, Q., Shang, Y. L., Chen, H., Zhang, Z. F., & Fan, B. B. (2018). Customer churn prediction model for product service system based on improved support vector machine. *Combined Machine Tools and Automated Machining Technology,* (05), 181-184.

Weng, D. (2016). *Research on corporate customer churn and customer segmentation.* Nanjing University.

Wu, X. J., & Meng, S. S. (2017). Research on e-commerce customer churn prediction based on customer segmentation and Ada-Boost. *Industrial Engineering, 20*(02), 99-107.

Yao, D. (2017). *Research on customer churn prediction model and its application.* Northwest University.

Zhai, X. K., & Liao., M. H. (2018). Research on WeChat social e-commerce information communication behavior based on SIR model. *Journal of Changsha University, 32*(05), 70-73.

Zhang, D. (2015). *Establishment and application of customer churn prediction model.* Beijing Institute of Technology.

Zhang, J. (2015). A customer churn prediction model based on C5.0 decision tree. *Statistics and Information Forum, 30*(01), 89-94

Zhou., D. (2014). *Research on telecom customer churn prediction based on feature selection of rough sets.* Jiangsu University of Science and Technology