Advances in Engineering Research, volume 149
2nd International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018)

The Research of Online Shopping Customer Churn Prediction Based on Integrated Learning

Guoen Xia [a], Qingzhe He [b, *]
Guangxi University of Finance and Economics, Nanning, 530001, China
[a]1343257009@qq.com,[b]2235692517@qq.com

Keywords: customer churn, artificial neural network,support vector machine.

Abstract. The prediction of customer churn is an important research direction of customer churn management. In this paper, take the non-contract scenario of online shopping customers as an example, select transaction data of a domestic e-commerce website for empirical research. On the basis of the single model-BP neural network and support vector machine, apply the integrated learning theory to the online shopping customer classification. The empirical results show that the combined forecasting model has a significant improvement in the hit rate, coverage rate, accuracy rate and lift degree, and so on. In order to effectively identify different types of lost customers, use the RFM theory to classify the different value of the lost customers, thus implementation the strategy of customer churn retention.

1. Introduction

In recent years, with the development and improvement of network technology and the process of social information, Internet technology has an important impact on people's lives. Especially, the arrival of the “Internet +” e-commerce, traditional enterprises have to use the Internet platform to open up the market, looking for new profit growth point, the online business promote the rapid development of the network retail market. The development of the online retail market intensifies competition among the industry. For the e-commerce industry, the customer churn rate is high, business operators need to consider how to minimize the customer churn rate of online shopping. Because of the customer's behavior is predictable, through the relevant data collected to carry out the relevant analysis can find the customer's future trading tendencies. For business operators, reduce the number of lost customers, an effective way is to find the customer who has the lose tendency and do the relevant pre-control work. In recent years, online shopping customer churn prediction has become an important direction of e-commerce business research.

2. Customer Churn

2.1 The Concept of Customer Churn

Customer churn is the state that customer contribution to profit of the enterprise is declining . [1] According to whether the enterprise and the customer entered into a contract to divide into two categories: customer churn in the context of contractual relationship and non-contractual relationship.

(1)Customer churn in the context of contractual relationship. Customer relationship in the contract scenario means that the firm and the customer in order to reduce the loss of both parties in the transaction process and to bind the contract by binding the contract. The contract has a clear stipulation on the rights and obligations of both parties. After the customer signs a contract with the enterprise, the customer needs to fulfill the relevant obligation in accordance with the contract to enjoy the corresponding right. In the contract relationship, the transaction between the customer and the enterprise is bound by the agreement, the customer needs to pay a higher cost of transfer, such customer's trading behavior is relatively stable. Telephone communication business, insurance business, etc. are typical enterprise that based on the contractual relationship.

(2) Customer churn in the context of non-contractual relationship. In the case of non-contractual circumstances, the relationship between firm and customer is started with the customer's initial transaction.It is not necessary to sign a contract between customers and enterprises, customers can

freely enter and leave the enterprise, the customer's trading behavior and the loss behavior is uncertainty. Customers may purchase products for a long time after their initial transaction or never generate trading behavior. As the enterprise is less binding on the customer, coupled with the low cost of customer transfer, so in the context of non-contractual relationship, the customer churn rate is higher.

The difficulty of judging customer churn in non-contractual relationships is that the criteria for lost customer is more vague and there is no clear distinction between lost customers and non-lost customers. E-commerce enterprises need to be based on the characteristics of enterprise products, set the standard for customer churn, in order to be able to timely discover the tendency of the lost customers, and then take effective retention measures.

Table 1. The comparison of contractual and non-contractual customer

	Contractual	Non - contractual
Trading behavior	stable	random
Customer churn	low	high
Predictiondifficulty	easy	difficult

2.2 Online Shopping Customer Churn

Online shopping customers are the typical of customer churn in the context of non-contractual relationship, take the duration of the transaction as consideration, the loss of online shopping customer is divided into the following two types: interruptions and permanent lost.

(1) Intermittent lost. Intermittent lost means that the customers did not buy the enterprise's goods or services during the specific time threshold, the main feature is the decrease in trading frequency. Such customers may not purchase the enterprise's product within the time threshold, however it does not mean that the customer has been completely lost, and the customer may continue to purchase the product or service of the enterprise beyond the time threshold.

(2)Permanent lost. Permanent lost means that the customer will not buy enterprise's products or services in the future. E-commerce enterprises will not write off the customer's account, even if the customer does not use the account for a long time, the customer can still use the registered account to log in, the enterprise can not distinguish whether the customer is permanent lost by the registered account. Permanent lost means that the customer is exhaustive loss, leading to the emergence of such customers for many reasons, such as the change of customer's spending habits, the change of the growth stage that the customer is no longer need the product or customer natural death.

2.3 Overview of Online Shopping Customer Churn

After reading the literature on customer churn, it has been found that there is little research on customer churn prediction for non-contractual relationships. The research on non-contractual relationship customer behavior, mainly focused on the purchase behavior. Schmittlien Morrison and Colombo proposed the SMC model for predicting customer transactions in 1987. The model obtained the customer activity through mathematical calculation and judged the customer's activity in the future based on the customer activity. Liu Xuewei (2006)[2]in order to improve the prediction accuracy of individual activity of e-commerce website customers,use the combination of naive Bayesian and SMC models to predict the performance. The new model is superior to the SMC model, Naive Bayesian algorithm. Dai Yisheng (2010) [3] calculated the customer activity level of the e-commerce website with the non-contractual relationship with the SMC model. The empirical results show that the higher the degree of customer's activity, the smaller the probability of customer churn. Wu Hong (2015)[4] uses the SMC model to calculate the potential value of customers and introduce them into the characteristics of customer attributes. Empirical analysis shows that the artificial neural network model with potential value is better.Zhu Bangzhu (2010) [5] selected active degree 0.5 for the threshold to judge the existing customers, and attribute reduction sample with support vector machine to establish forecast model.The results show that the SMC model-rough set-support vector machine is more accurate.

The use of SMC model to predict customer's future activity, provides a new method to solve non-contract customer churn prediction. However, the SMC model has the limitation of the accuracy of the activity level at the individual level when the activity is predicted. At the same time, the SMC

model is based on strict assumptions, it does not allow customers to be inactive and then back to active state. In fact, online shopping customers are more casual, and the loss of customers' time threshold does not mean that customers are permanently lost. Therefore, from the strict sense, SMC model theory in the field of online shopping customer churn applications also have some limitations. In the prediction of customer churn, building a single classification model, the generalization ability is weak, and it is difficult to solve similar problems. Using the combined model can take advantage of the various prediction methods to increase the reliability and stability of the prediction results. The establishment of the combined forecasting model and the improvement of the prediction accuracy have become an important research direction in the field of prediction.[6]

3. Empirical Study

3.1 Data Processing

The data collected by online crawling technology has the characteristics of incomplete and non-linearity, and the collected data can not be directly applied to the research. It needs to be standardized to deal with the original data to meet the actual research needs.

(1) In the process of online transactions, the customer has a system assigned ID, and it is the only constant. For the effective distinction different customers, an ID corresponding to the unique number.

(2) The transaction time is time point, it is necessary to convert the transaction time of the customer into a continuous variable, and it need to be standardized to [0, 1].

(3) The price of the goods, the score in the numerical show great differences, it needs to standardized to reduce the data gap on the impact of the result, and data normalization to [0,1].

3.2 Data Understanding

Based on the characteristics of electronic products, this paper selects the customer transaction data from January 2014 to December 2014 as the observation period, and the customer transaction data of January and February 2015 as the forecast period. If the customer does not purchase the product or service on the website during the forecast period, such customers are considered to be lost customers, marked "1"; if the customer purchases the product or service on the website during the forecast period, such customers are considered to be non-lost customers, marked as "0". The customer's attributes include the customer's number, the price of the product, the first time of trading behavior(FirstTimeDiff), the last time of behavior (TimeDiff), the frequency of trading behavior, the score of the customer, customer's category (Label). The customer's attribute table is shown in table 2.

Table 2. The customer attributes

The customer attributes
Number
Price
FirstTimeDiff
TimeDiff
Frequency
Score
Label

3.3 Data Sampling

In order to extract the training data set, the difference in the quantity of customers and non-lost customers is taken into account. At the same time, in order to extract the characteristics of the lost customers fully, in this paper,by stratified sampling method the same number of lost customers and non-loss customers are extracted. The number of lost customers and non-lost customers, such as table 3. According to the existing empirical study, the training sample and the test sample are used in the proportion of 2: 1 in the quantity of the training sample and the test sample.

Table 3. The training sample data analysis

	Frequency	Percentage	Effective percentage	Percentage accumulation
0	1890	50.2	50.2	50.2
1	1873	49.8	49.8	100.0
Total	3763	100.0	100.0	

3.4 Data Analysis

The relationship between customer attributes and customer churn is analyzed before the model is established, and the impact of customer attributes on customer churn is understood.

(1)The relationship between commodity prices and customer churn

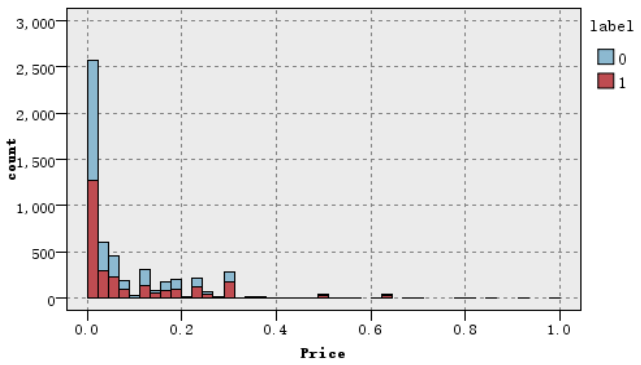

Figure 1. The relationship between price and customer churn

The horizontal axis indicates the price of the product that is purchased by the customer. As can be seen from the figure, the price of the purchased goods is mainly concentrated in the range of 0-0.4. Cheaper electronic products tend to have a lot of consumers, while higher-priced electronics have fewer buyers. In the case of the same price,As commodity prices rise, the proportion of lost customers accounted for a downward trend.

(2)The relationship between customer's first trading behavior and customer churn

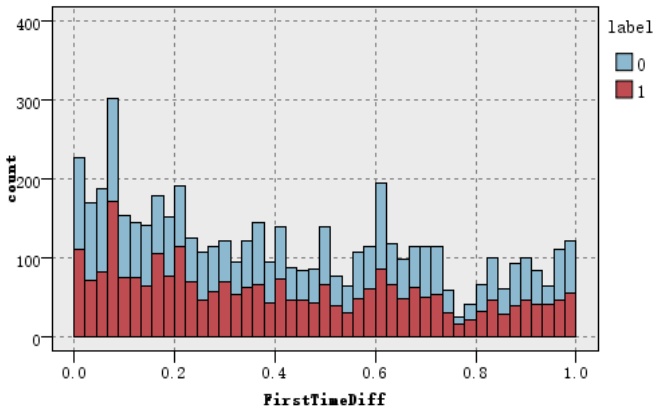

Figure 2. The relationship between customer's first trading behavior and customer churn

The horizontal axis represents the time of the first trading behavior. Customer churn and the first trading behavior did not show a clear linear law. There is no specific explanation for the relationship between the two.

(3)The relationship between customer's last trading behavior and customer churn

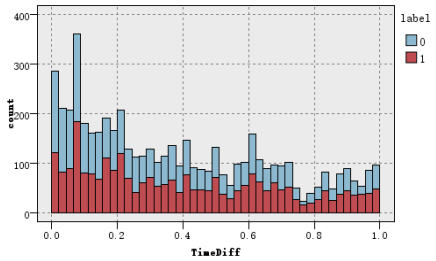

Figure 3. The relationship between the trading behavior and customer churn

The horizontal axis indicates the time the customer has recently trading behavior the item. The relationship between the time of the last trading behavior and the customer churn is not particularly noticeable, and there is no significant difference in the number of customers entering or the last trading behavior threshold.

(4)The relationship between trading frequency and customer churn

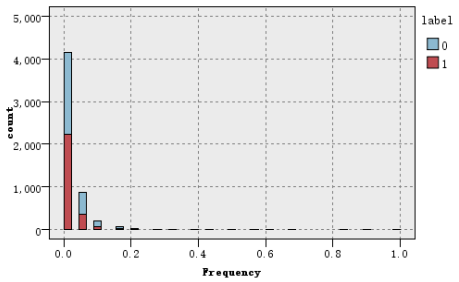

Figure 4. The relationship between trading frequency and customer churn

The figure shows the horizontal axis of the customer's trading frequency, lost customers and non-lost customers are mostly concentrated in the 0 to 0.2 range. As can be seen from Figure 4, with the increase in customer purchases, the proportion of non-lost customers is increasing in the same purchase frequency.

(5)The relationship between product's score and customer churn

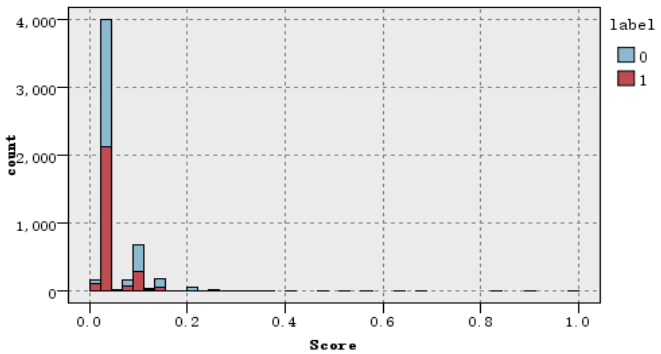

Figure 5. The relationship between product's score and customer churn

The horizontal axis represents the customer's score on the product, and the lost customer and the non-lost customer are mostly concentrated in the lower score range. However, with the increase of customers' score on commodities, the proportion of lost customers shows a decreasing trend

3.5 Model Establishment

From the exploratory analysis of customer's attributes and customer churn, it is known that there is no simple linear relationship between customer's attributes and customer churn. It can consider the ability of artificial neural network to deal with non-linear data and effectively classify customer churn.In a certain time the number of the E-commerce enterprise's customers is basically in a relatively stable range, it can take advantage of the support vector machine model in the classification of small sample data. In recent years, the use of composite forecast is more extensive, many researchers have combined neural neural networks, genetic algorithms and other methods together.

According to the characteristics of online shopping customer transaction data, use artificial neural network and support vector machine for integrated learning to play their respective advantages to predict the situation of customer churn.

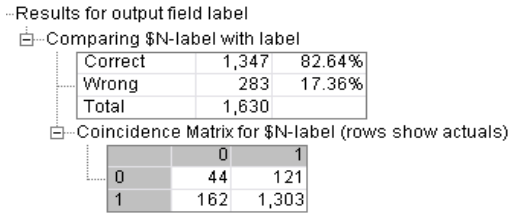

Figure 6. The results of the BP artificial neural network

The trained BP artificial neural network model is used to predict the customer churn from January 2015 to February, and the correct rate is 82.64% and the error rate is 17.36%.

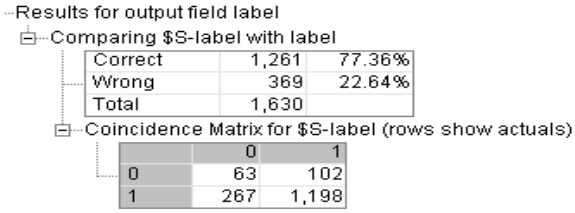

Figure 7. The results of the support vector machine

By training the mature support vector machine model, the customer churn from January to February 2015 is predicted, and the correct rate is 77.36%.

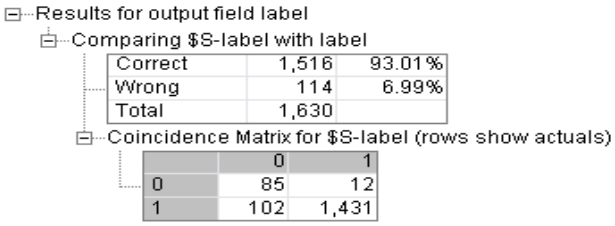

Figure 8. The result of the combined model

The forecasting result shows that the correct rate of the combined forecast is 93.01% and the error rate is 6.99%. From the results of customer churn prediction, it can be seen that increasing the accuracy of customer churn prediction is necessary to reduce the actual lost customers forecast for the non-lost customers, the actual non-lost customers forecast for the lost customers.

3.6 Model Evaluation

Table 4. The classification table of customer churn

Actual / Prediction	Non-lost customers	Lost customers
Non-lost customers	f_{11}	f_{12}
Lost customers	f_{21}	f_{22}

The customer churn classification table is a form that can visually describe the true classification of lost customers and non-lost customers in actual forecasts. In assessing the customer churn model, the model is evaluated by selecting the hit rate, coverage rate, accuracy rate, and lift degree.

The hit rate is expressed as: $f_{22} / (f_{12} + f_{22})$

The coverage rate is expressed as: $f_{22} / (f_{21} + f_{22})$

The model accuracy rate is expressed as: $(f_{11} + f_{22}) / (f_{21} + f_{22} + f_{11} + f_{12})$

Lift degree: The ratio of hit rate to accuracy.

Table 5. The evaluation results of the three model

	BP neural network	SVM	Combination
Hit rate	91.50%	92.15%	99.17%
Coverage	88.94%	81.77%	93.54%
Model accuracy	82.64%	77.36%	93.01%

The lifting degree is an important index to evaluate the merits of the model. In order to see the pros and cons of the model more intuitively, this paper uses the lift curve to show the effect of the model.

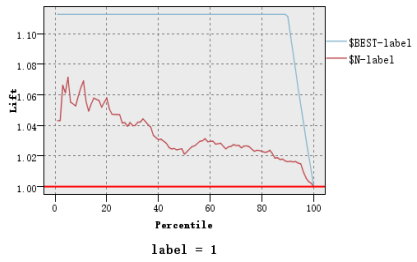

Figure 9. The lifting curve of the BP neural network model

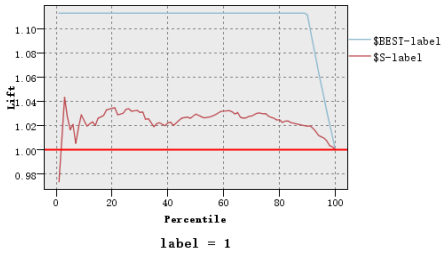

Figure 10. The lifting curve of the Support Vector Machine model

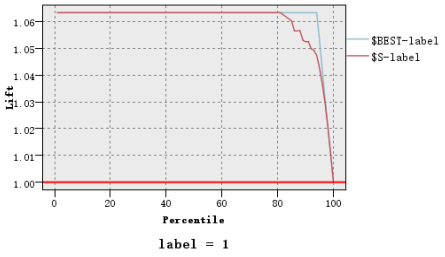

Figure 11. The lifting curve of the combined model

There are three lines in the lift curve, the bottom horizontal line is the baseline, the top is the optimal curve, the middle line is the actual lift curve. When the lift curve is closer to the optimal curve, the model's performance is better. Compared with the improvement curve of the models, and it is found that the lifting curve of the combined model is basically close to the optimal curve . That is, the performance of the combined forecasting model is better than that of the single model.

4. The Lost Customer Value Analysis

In order to analyse the value of the lost customers, this paper uses the RFM theory, the recent purchase date R(Recency), the purchase frequency F(Frequency) and the purchase amount M (Monetary) on the loss of the value of the customer's contribution to the enterprise The value of the customer is analyzed, and the weight between the three is equalized. At the same time, regardless of the impact of the time node division on the customer's recent purchase date, then the two variables of F and M are divided into two states. F_1 is lower than the average trading frequency , F_2 is higher than

the average trading frequency, M_1 is lower than the average price of product, M_2 is higher than the average price of product.

Table 6. The classification of lost customer's value

Customer category	Quantity	Rate(%)
F_1M_1	834	57.5
F_1M_2	363	25
F_2M_1	142	9.7
F_2M_2	113	7.8

(1) F_1M_1 customers. This type of customer is the largest proportion in the lost customers, more than 50%. Such customers are less valuable to the business, in the traditional management concept, the lower value of the customer can be ignored, but for e-commerce enterprises, such customers should not be ignored. Reducing the loss of low-value customers is an issue that e-commerce companies must be solved.

(2) F_1M_2 customers. This type of customer's trading frequency is lower than the average level, the amount is higher than the average level, the proportion is about 25%. Compared with the F_1M_1 type of customers, in the trading frequency is basically the same circumstances, with the increase in the amount of purchase,the customer churn significantly reduced.

(3) F_2M_1 customers. The proportion of lost customers in this category is 9.7%. Compared with F_1M_1 customers, the increase in customer purchase frequency can significantly reduce the churn rate when the purchase amount is at the same level. A certain period of time, the customer's income is basically stable, increase the amount of customer purchase is more difficult. Therefore, the most effective way is to increase the frequency of customers .

(4) F_2M_2 customers. Such customers, whether the purchase frequency or the purchase amount are above average, the percentage of such customers in the lost customers is the lowest proportion. In the implementation of the lost customer retention measures, on the one hand, increase the trading frequency of online shopping customers and increase the amount of the product can effectively reduce the customer churn rate. On the other hand, such customers because of their high-value customers, and the probability of loss is relatively low, such customers should be as much as possible to restore.

The existing research shows that the old customers are more likely to be higher than the new customers in frequency and monetary. By analyzing the value of the lost customers, it is shown that the likelihood of customer churn is significantly reduced as the amount and the frequency of purchase increases. For e-commerce enterprises, operators should start from the purchase amount and the purchase frequency of two aspects, implement customer churn retention strategy, effectively reduce the customer churn rate, achieve sustainable development of the enterprises.

Acknowledgments

Author: Guo-en Xia(1977-), male, Sichuan Neijiang, Professor, Doctor of management. Research direction: business intelligence, customer relationship management.

Corresponding Author : Qingzhe He(1991-),male, Master graduate.

Funds: Guangxi Province Universities and Colleges Excellence Scholar and Innovation Team Funded Scheme; The Foundation of Guangxi Key Laboratory Cultivation Base of Cross-border E-commerce Intelligent Information Processing, Guangxi University of Finance and Economics; Innovative Governance and Intellectual Property Discipline of Guangxi University of Finance and Economics.

References

[1]. Yu Xiaobing, Cao Jie, Gong Zaiwu.The review of customer churn research [J]. Computer integrated manufacturing system, 2012(10):2253-2263.

[2]. Liu Xuewei.The prediction of e-commerce customer churn based on Pareto/NBD+ naive bayesian combined model [D]. Sichuan university, 2006.

[3]. Dai Yisheng, Shen Peilan, Sun Hongxia. The research of customer churn prediction based on Pareto/NBD model [J]. Science and technology and engineering, 2010.

[4]. Wu Hong. The Comparison and empirical study on the model of e-commerce customer churn [D]. Capital economics and trade university, 2015.

[5]. Zhu Bangzhu.The prediction of e-business customer churn based on smc-rs-lssvm model [J].Systems engineering theory and practice, 2010(11):1960-1967.

[6]. Ren Jianfeng, zhang Xinxiang. The modeling and prediction of the e-commerce customer churn[J]. Computer simulation, 2012(05):363-366.