Research on E-commerce User Churn Prediction Based on Logistic Regression

Qiu Yanfang
School of Information Management
Beijing Information Science and Technology University
Beijing, China
366043744@qq.com

Li Chen
School of Information Management
Beijing Information Science and Technology University
Beijing, China
lichen@bistu.edu.com

Abstract—**With the development and popularization of Internet technology, e-commerce platform has provided satisfying products for customers and cultivated customer loyalty. Nevertheless, the loss of user is still a popular issue in business field and academic field. Based on logistic regression model, this paper established an e-commerce user churn prediction model through preliminary research on e-commerce customer churn behavior. By using the factor analysis method, the user's online duration, number of logins, attentions, and other user behavior factors were analyzed which concludes the factor affecting the loss of users. Finally, the empirical study proved that the proposed EBURM model can predict user churn behavior in a high confidence level.**

Keywords- e-commerce; Logistic regression model; User behavior; User retention rates

I. INTRODUCTION

According to the research data released by CNNIC, as of December 2016, the extent of China's e-commerce users reached 467 million. With the rapid growth of e-commerce users, how to predict the possibility of user churn in advance has become an urgent problem for e-commerce platform. The factors affecting user retention of e-commerce includes user's attention to shop, browsing rate of recommended information, demand for the sharing function, number of online times per day, and length of the daily online, those are important factors affecting e-commerce platform user churn, which can influence the prediction accuracy to a large extent.

The problem of e-commerce customer churn prediction has its own particularity. The e-commerce platform can't accurately judge whether the user is really lost, which increases the level of difficulty and complexity of prediction greatly. Currently, the algorithms applied to user churn prediction include decision tree, artificial neural network, Logistic regression model, K-Means algorithm, naive Bayes and so on. In 2011, Yucheng Zhang proposed Markov model to predict user churn whose disadvantages were of low accuracy, low prediction coverage and high storage complexity, etc. [1] In 2016, Yang Tao Liu chose embedded vector and recurrent neural network method to conduct the research. However, they failed to make the model in a stable level [2]. Sun et al chose the SVM model when establishing a bank credit card user churn prediction model. [3]

The prediction of user churn in e-commerce platform is a classical dichotomy problem [4]. The prediction results are the possibility of being retained or lost, rather than classifying user behavior directly. As a common statistical analysis method used for classification, logistic regression can obtain probabilistic prediction results that is applicable to predict the user churn behavior of e-commerce platform.

II. ESTABLISHMENT OF EBURM MODEL BASED ON LOGISTIC REGRESSION

A. Logistic regression

The prediction of user churn in e-commerce is an obvious two classification problems. Logistic regression is a commonly used statistical analysis method which can be used for classification, prediction results can be obtained and probability, which belongs to a kind of probability type nonlinear regression. [5] Let the conditional probability $P(z) = p$ be based on the probability of an observation relative to an event, then the logistic regression model can be expressed as:

$$p(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} \tag{1}$$

Since the result of the prediction is yes or no between the two possibilities, the range is [0,1], so we can estimate the probability that the variable P = 1 is based on its value.

Maximum likelihood estimation, also called maximum likelihood estimation, the basic idea of this method is: when the model group n were randomly selected from the total sample observations, the most reasonable parameter estimates from the model should make the probability of extracting the n group sample observation value maximum. This is an iterative algorithm, which takes an estimate value as the initial values of the parameters, according to the algorithm to determine the direction and change of parameters can increase the log likelihood value, estimation of the initial function, test the residuals and re estimated by the update function improved, until the log likelihood value is no longer significant change so far [6]. Because the solution is more complex, this is no longer the case, and the application of the SPSS software is usually calculated using the SPSS software. Finally, the prediction model of logistic regression can be established by substituting the obtained parameters (1).

B. EBURM model building

Here we need to build a model, because there are two kinds of things that are active and churn users of e-commerce, define y_n as the category of e-commerce users in sample data. When $y_n = 1$, it represents the user as active user, and $y_n = 0$ represents the user as the churn user. The retention rate R (retention rate) is defined as a real number between the range $[0,1]$ and is used to indicate the possibility of loss of the user y_n of the e-commerce platform. The greater the value, the greater the likelihood that the user will remain on the e-commerce platform [6]. Set $x = (x_1, x_2, x_3, \cdots, x_n)$ as the dependent variable of the user's y_n behavior index, the logistic regression model can be used to calculate the retention rate of e-commerce users. The formula of R is:

$$R = P(y_n = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \quad (2)$$

Based on the above model is trained using the sample data, using the maximum likelihood method or the use of SPSS software can obtain the estimation value of each parameter of the model, thus getting the final e-commerce user's retention situation EBURM (Electronic Business User Retention Model).

C. Extraction of Characteristic Factors

In this paper, we analyze the user behavior of e-commerce, combine and transform the original features of e-commerce users through reasonable logical induction, and extract the following characteristics as the factors of the model and take the variables the value gives a specific formula.

- the user's interest rate for e-shops

The user's interest rate for the store is expressed in this paper by β_1 which refers to the degree of attention that the user pays attention to the e-commerce store, which can be measured by the number of times the user clicks into the shops of interest, from the user's payment to the particular shop concerned Analysis, that is, the user clicks into the store and successfully ordered the more orders to pay more, the more the number of comments that the user concerned about the e-commerce concerns the higher the rate of shops.

$$\beta_1 = \frac{\text{number of times a user has successfully paid A shop}}{\text{number of users paid}} \quad (3)$$
$$\times \frac{\text{number of shops purchased by user}}{\text{number of users concerned}} \times \alpha$$

The loss of users of the following characteristics, the concern of fewer shops, orders are also less; or basically no orders. However, the more users retain, the more orders will be placed. That retained user is loyalty to some shops in e-commerce, so users focus on shops can be replaced by $\alpha = \frac{1}{1+e^{-\sigma}}$, $[0,1]$ and σ in the range, is the dispersion between the shops and the number of orders, so according to the standard deviation formula, where $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ represents x_i shops under the singular, i for users the number of shops to pay. Therefore, the meaning of α can be defined as: when users pay more attention to the shops, the greater the dispersion of orders, the user is more likely to be retained users, α closer to 1.

- Recommended CTR

In this article, the user's attention to the recommendation information by β_2 that the e-commerce platform is now a user to recommend a variety of information, and recommended information is generally the user's personalized needs, if a user clicks the recommended number of times the more information E-commerce platform to understand the user, you will get the user's degree of love, so the higher the recommended rate of attention, the more the more likely to retain the user retention.

$$\beta_2 = \frac{\text{times people views recommendations}}{\text{user views}} \quad (4)$$

- Share rate

Where β_3 said the sharing rate, users share an e-commerce platform each time, indicating that the e-commerce platform products or activities have been the user's favorite, share to the third-party platform, indicating that users of our e-commerce platform promotion, the possibility of retention is higher, the formula is as follows:

$$\beta_3 = \frac{\text{click to share btn times}}{\text{users per click btn times}} \quad (5)$$

- Number of daily

Here, the daily number is expressed by beta _4, one of the factors that users may be wasting is the number of days on which the platform is used, and if the number of times used is low, the long-term retention rate is low. Therefore, it is also an important characteristic factor to analyze the user's daily login. The formula is as follows:

$$\beta_4 = \frac{\text{log times}}{\text{all users times}} \quad (6)$$

- User churn time

β_5 is used to indicate the length of the drain. The standard of determining whether a user is really missing is the login frequency of the user in this e-commerce platform within 3 months, if a user is registered for use a week after the frequency of re use of a linear downward trend, and in three months after basically no landing, the platform shows that this user drain. According to data from an e-commerce channel shown in Figure 1 below, it will take at least three months for a user to have a significant loss in one channel. And from the data shown on the map, a week time you can see the user's loss situation, the user churn is 7 days a week cycle.

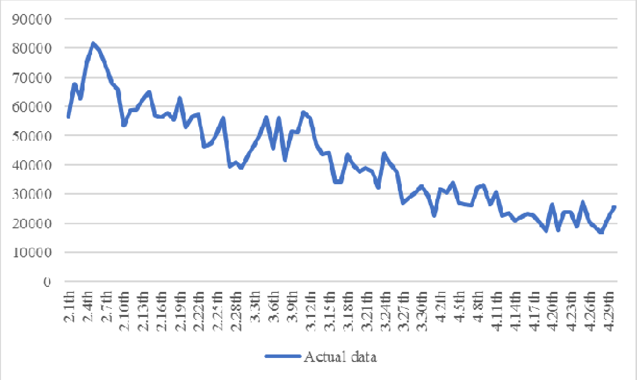

Figure 1. Electricity Business Platform 3 Months User Data Volume

- Constant

The data selected in this paper is extracted from an e-commerce platform of 6000 data, which retained the amount of data 3211, and the amount of data lost 2789. The formula for the constant β_0 is thus given as follows:

$$\beta_0 = log(p) = ln\frac{2789/6000}{1-2789/6000} = -0.1409 \tag{7}$$

III. Empirical Research

A. data collection

Through the sample data, the model can be trained and the maximum likelihood method can be used to obtain the estimated value of each parameter of the model, so as to get the EBURM of the final e-commerce user. The concrete process is as follows:

- The Classification Process

Step 1 quantifies the behavior of the test user to form the model's independent variable value.

Step 2 uses the EBURM model to calculate the probability that the user is a real user, that is, the user's retention rate R.

Step 3 determines the e-commerce user category according to the set classification standard value (threshold value).

- Data Processing

The data used in this study is the real user data of an e-commerce, which collects 3 month users' data of one channel of the platform, and extracts the information of 6000 user data. And in 3 months of data statistics, identified 3211 active retained users and 2789 lost users.

B. Parameter estimation and explicit test

In this paper, the binary logistic regression analysis module in SPSS software is used to train the model, and the parameters of the model are estimated and tested. Set the dependent variable and covariate, the classification standard value is set to 0.5, the other settings are the default value, the calculation results as shown in Table Ⅰ below.

TABLE I. THE FACTOR COEFFICIENTS IN THE EQUATION

	B	S.E.	Wald	df	Sig.	Exp(B)
β_1	40.095	4.892	40.262	1	0.000	70.971
β_2	15.125	2.337	12.236	1	0.002	20.112
β_3	3.142	0.969	26.107	1	0.006	8.861
β_4	3.425	1.847	27.136	1	0.000	8.326
β_5	6.21	0.137	4.326	1	0.001	0.129
β_0	-0.141	0.347	9.763	1	0.001	0.310

From Table Ⅰ, we can see the user's attention to the shop, the recommended information attention rate, the number of hits, and the length of the online sig. Values are less than the critical value of 0.05, and users share the sig value of more than 0.05, Has nothing to do with the model, so the following models no longer use the user to share the rate of click on the independent variables. On the whole, the model is feasible and four indicators are available, so the model changes to 4 variables for the final variable. The B value is the coefficient of each factor, and the parameter estimate of the model can be obtained by Table Ⅰ. The following is the formula (8):

$$R = \frac{1}{1+e^{-(-0.141+42.095x_1+7.125x_2+11.425x_4+6.21x_5)}} \tag{8}$$

From the estimation of the parameters of the factors are positive, indicating the degree of attention, recommendation information, as well as the number of logins and duration and retention rate is positively related, the greater the value, the higher the retention rate, the less likely the user is lost. Because of its sig are within the critical value, indicating that the significant significance of the variable, so the use of the model is very high.

C. Result analysis

The prediction of the retention rate of e-business users is that a classifier is created to classify the categories of users belonging to an e-commerce user. The performance of a classifier is evaluated, the performance evaluation metrics usually have the following:

- Accuracy: The ratio of correct sample size to total sample size
- Precision: Predict the ratio of the correct number of samples to the total number of samples in the state
- the full rate: Predict the correct sample number to the actual sample ratio

- rate of omission: the ratio of the sample size of the prediction error to the total sample number

In the test set to extract 10 user data, including five retained users and five lost users, according to the formula forecast, the result is lost users have two predictions are wrong, respectively, have to buy, but the online length And the number of online are relatively small, that some of the occasional users is also a loss of the other one is no purchase behavior, but online length, browse the recommended information values are relatively low, indicating that this may be accidental click , But the platform does not really need the user, this is the loss of the user.

TABLE II. CALCULATION RESULTS OF RETENTION RATE R

user	category	x1	x2	x4	x5	R	prediction
1	Retained	0.0085	0.0107	0.9279	0.0037	0.97242	1
2	Retained	0.0124	0.4851	0.5258	0.0026	0.999926	1
3	Retained	0.023	0.2419	0.3333	0.0063	0.996391	1
4	Retained	0.0024	0.0671	0.2965	0.0012	0.880069	1
5	Retained	0.0012	0.0067	0.1308	0.0019	0.614967	1
6	chum	0.0018	0.0051	0.0012	0.0002	0.503415	0
7	chum	0	0.0029	0	0.0001	0.47589	1
8	chum	0	0.0017	0.2975	0.0015	0.713607	0
9	chum	0	0.0007	0	0.0022	0.470845	1
10	chum	0	0.0005	0	0.0049	0.474271	1

The data of the user data of a channel of an e-commerce platform was selected, and the behavior information of 6000 users was selected after the data processing. The standard value of the classification evaluation was set to 0.5, that is, the retention rate R was 0.5 or more Indicating that the user to retain the user, on the contrary, when R is less than 0.5 for the loss of users. According to this setting, according to the model statistics out of Table Ⅲ, Table Ⅳ

TABLE III. EBURM MODEL PREDICTION RESULTS

	Predict retention users	**Predict churn users**	**Predictive accuracy**
Actual retention users	3084	127	96.03%
Actual churn users	257	2532	90.78%
Total Accuracy	-	-	93.6%

TABLE IV. EVALUATION INDICATORS

（T）	（P）	（R）	（M）
93.6%	96.03%	95.22%	6.4%

a. （Accuracy -T； Precision -P； Check the rate -R； missed rate-M）

From the results predicted in Table Ⅲ, the test set has a total of 3211 retained users, 3084 judged to retain the user, the predicted accuracy rate of 93.6%, a total of 2789 users were lost, was found to have lost 2532 users, the prediction accuracy rate is 95.22%, which indicates that the accuracy of the model is very high in the prediction of the retention and loss of the user's behavior. The accuracy of the whole model is 93.6%, and the accuracy of the model's prediction of user churn Is the highest. Indicating that the model is reasonably available.

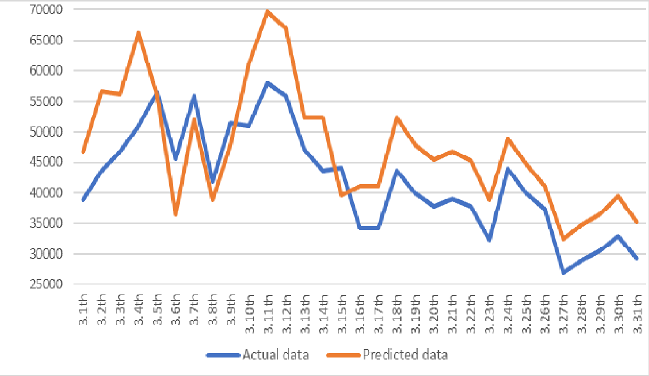

Figure 2. Comparison of the actual data and the predicted data obtained from the model

After comparing the predicted model data with the original data, figure 2 is obtained, after obtaining the predicted data, the predicted data is basically close to the practice data, and the error is about 7%. And from the different AUC value can be seen, EBURM model logistic regression to establish the correct rate of the value is relatively high based on the data, so we can conclude that the accuracy of predicting the loss behavior of EBURM user model in electronic commerce is relatively high. So that the EBURM model can predict the availability of relatively high.

IV. SUMMARY

The EBURM model is evaluated by the AUC test method. The results show that the EBURM model is consistent with actual expectations for active and churn users. Based on the different influencing factors of the user retention rate, the EBURM model provide a personalized operational recommendation strategy. Comparing to the method of user type predication, this model can predict user behavior more accurately to reduces user churn. Through the construction of the EBURM model to predict e-commerce user churn behavior, it helps e-commerce platform to formulate operational strategy more precisely, provide users with personalized recommendations, increase user activity, retain users, and improve the economic effects of e-commerce platform.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China (No. 61272513), based on Multi-Agent, the study of collaborative supply chain data integration under emergency status.

V. Reference

[1] Zhang Yucheng, xu big grain, Wang Xiaojuan. Active user behavior based on weighted markov chain prediction model [J]. Computer engineering and design, 2011, (10): 3334-3337 + 3418.

[2] Liu Yangtao, south slope, Yang Xinfeng. Based on embedded vector and circulation of the neural network user behavior prediction method [J]. Journal of modern electronic technology, 2016 (23): 165-169.

[3] Li Shi bo, Sun Bao hong, Wilcox R T. Cross-selling sequentially ordered products: An application to consumer banking [J]. Journal of Marketing Research, 2005,42(2):233-239.

[4] Tang Xing Quan Yi ning, Song Jianfeng, Michael Dunn e, Zhu Hai, MiaoQi widely. Weibo forward personalized pre diction of the new algorithm [J]. Journal of xi 'an university of electronic science and technology, 2016, (4): 62-56 + 51.

[5] Musa A B. Comparative study on classification performance between support vector machine and logistic regression [J]. International Journal of Machine Learning and Cybernetics,2013,4(1);13-24.

[6] Chang Zhenhai, Liu Wei. Logistic regression model and its application [J]. Journal of Yanbian University (NATURAL SCIENCE EDITION), 2012, (01): 28-32.

[7] Gupta A, Kumar guru P. Credibility ranking of tweets during high impact events [c] //Proceedings of the 1st Workshop on Privacy and Security in Online Social Media. New York: ACM,2012.