Research on the Churn of Daily Necessities Users on E-commerce Platform based on Binary Logistic Regression

Jiazhen Tian[1], Weipeng Zhang[2,*]
[1]*School of Business Administration, Ningbo University of Finance and Economics, Ningbo, China*
[2]*College of Digital Technology and Engineering, Ningbo University of Finance & Economics, Ningbo, China*
**Corresponding Author*

Abstract: To analyze the influencing factors of churn of daily necessities users, questionnaires were designed and data were collected from four aspects: user behavior, after-sales service, transportation and offline comparison. The binary logistic regression model was used to provide a preliminary churn prediction, and extract the key influencing factors from nine factors: gender, age, educational background, occupation, monthly income, user behavior, after-sales service, transportation problems and offline comparison. After analyzing the model, it can be found four factors, namely educational background, user behavior, after-sales service, transportation and offline comparison have a great influence. The results show that the model is of high accuracy and interpretability in predicting the churn of daily necessities users on e-commerce platforms. This study has certain reference significance for E-commerce platforms to improve user retention rate and improve operation strategy.

Keywords: Daily Necessities; Purchase Intention; Binary Logistic Regression; Customer Churn

1. Introduction

In the highly competitive e-commerce market, user churn is an important index to measure the health of shopping websites. The rising user churn rate will both lead to the reduction of website user base and sales and weaken its market competitiveness. In recent years, with the rapid progress of Internet technology and the vigorous development of E-commerce market, various shopping websites have sprung up, providing users with richer choices [1]. Under this background, it is urgent for shopping websites to deeply analyze the root causes of user churn for the effective preparation of user retention strategies [2,3]. Therefore, in-depth research on the user churn will help shopping websites to accurately identify and improve the weak links in user experience, thus improving user satisfaction.

Based on the basic information of users, user behavior and other dimensions, this paper constructed the index system of influencing factors of purchasing daily necessities on e-commerce platforms. Logistic regression analysis method and binary logistic regression were employed to identify the key factors affecting the churn. On this basis, this paper put forward targeted policy measures and suggestions, aiming at helping enterprises to understand consumer psychology and purchasing intention more deeply. Thus, they could expand their sales channels and promote the widespread popularization on e-commerce platforms [4,5].

2. Data Collection

2.1 Questionnaire Design

In this study, the survey results were analyzed by means of questionnaire survey. A total of 482 valid questionnaires were collected, and the index system is shown in Table 1.

Table 1. Distribution Table of Questionnaire Content

Part I	The basic personal information includes the respondent’s gender, age, education level, occupation and monthly income.
Part II	Intention to buy daily necessities on E-commerce platforms, including: user behavior (7), after-sales service (4), logistics and transportation (4).
Part III	Whether they are willing to continue to buy daily necessities on e-commerce shopping platforms.

2.2. Questionnaire Test

2.2.1 Reliability test

Reliability analysis was mainly utilized to test the consistency of the survey data. On this basis, the index of Cronbach's compatibility coefficient was employed to test the internal consistency of the index, and the reliability analysis is shown in Table 2.

$$\alpha=\frac{k}{k-1}(1-\frac{\Sigma s_i^2}{s_x^2}) \quad (1)$$

Table 2. Reliability Analysis

Cronbachs' Alpha	Number of items
0.899	15

It can be seen from Table 2 that the reliability α value is 0.899 and greater than 0.8, which indicates the better data reliability.

2.2.2 Validity test

Reliability primarily examines the internal consistency among items within a scale, whereas validity focuses on the construct validity, the extent to which each item contributes meaningfully to the overall measurement [6]. The KMO test was leveraged to verify the validity of the questionnaire. In KMO test, the suitability of questionnaire data was evaluated by comparing the size of simple correlation coefficient and partial correlation coefficient between original variables. We are able to evaluate the validity of questionnaire more accurately and ensure the accuracy and reliability of research results by using this method. The validity analysis is shown in Table 3.

Table 3. Validity Analysis

KMO sampling appropriateness quantity		0.975
Bartlett's sphericity test	Approximate chi-square	2576.227
	Degree of freedom	136
	Significance	0.000

$$KMO=\frac{\Sigma_{i\neq j}r_{ij}^2}{\Sigma_{i\neq j}r_{ij}^2+\Sigma_{i\neq j}r^2_{ij\ *1,2\ldots K}} \quad (2)$$

Bartlett spherical test is a statistical method used to test whether the correlation matrix is an identity matrix, with the purpose of judging whether the variables are independent of each other. Through this test, we can know whether variables are highly correlated, so as to further analyze the construct validity of the questionnaire [7,8]. The formula is:

$$k^2=\frac{1}{c}\left[(n-r)lnMSe-\Sigma_{i=1}^{r}(n_i-1)lns_i^2\right] \quad (3)$$

Based on 482 valid questionnaires, the reliability of the questionnaires was tested by SPSS statistical analysis tool to ensure the reliability and stability of the data.

It is found through the questionnaire test of the scale that the Cronbach's α coefficient is above 0.8, which indicates good reliability. Then, the validity of the collected questionnaire data was tested and its results showed that the coefficient of KMO test was 0.975. It was qualified in the significance test of the validity of the questionnaire, and this method also fits other factor analysis, such as logistic regression.

3. Analysis of Influencing Factors of User Churn in E-commerce Platform Supplies based on Binary Logistic Regression Method

3.1 Principle of Binary Logistic Regression

Binary logistic regression is a probabilistic model, which can be used to predict the probability of events. Binary logistic regression requires that the explanatory variables should be category II, and the frequency of each category is higher than zero. The number of samples and dependent variables should have sufficient category samples for the use of binary logistic regression.

The logistic regression model is as follows:

$$P=\frac{\exp(\beta_0+\beta_1x_1+\ldots+\beta_mx_m)}{1+\beta_0+\beta_1x_1+\ldots+\beta_mx_m} \quad (4)$$

Then the probability of non-occurrence is:

$$1-P=\frac{1}{1+\exp(\beta_0+\beta_1x_1+\ldots+\beta_mx_m)} \quad (5)$$

After logical transformation, the data is then converted into $-\infty<log_{it}(P)<+\infty$. After logistic transformation, the logistic regression model can be expressed in the following linear form:

$$Logit_{(p)}=\beta_0+\beta_1x_1+\ldots+\beta_mx_m \quad (6)$$

3.2 Dummy Variables in Binary Logistic Regression Analysis

In binary logistic regression analysis, dummy variables play a key role. It is essential to construct a set of dummy variables that can effectively characterize the attribute characteristics of these variables affected by multiple categories of categorical variables, which helps us to more accurately analyze the influence of various factors on the results. In this study, gender, age and monthly income are selected as control variables, consumer behavior, after-sales service and transportation are explanatory variables. The explanatory variables are set as whether consumers will shop through E-commerce platforms in the future (0 means no, 1 means yes). Gender, age, education level, occupation and monthly income are all categorical variables, and corresponding dummy variables need to be generated for them in the analysis process. The relevant results obtained from the study are shown in Tables 4 and 5, respectively.

Table 4. Coding of Dependent Variables

Original value	Internal value
Won't	0
Will	1

Table 5. Coding of Categorical Variables

Variable	Classification	Frequency	Parameter coding		
			(1)	(2)	(3)
Monthly income	Below 1500 yuan	234	.000	.000	.000
	1500-2999 yuan	102	1.000	.000	.000
	3000-4999 yuan	90	.000	1.000	.000
	5,000 yuan and above	56	.000	.000	1.000
Age	18 years of age and under	124	.000	.000	.000
	19-29 years old	129	1.000	.000	.000
	30-39 years old	109	.000	1.000	.000
	Aged 40 and above	120	.000	.000	1.000
Level of education	Below high school	226	.000	.000	.000
	High School or College	116	1.000	.000	.000
	Undergraduate	89	.000	1.000	.000
	Graduate and above	51	.000	.000	1.000
Occupation	Student	295	.000	.000	
	Office worker	158	1.000	.000	
	Freelancer	48	.000	1.000	

In binary logistic regression analysis, categorical variables were usually coded in the way of dummy variables In this table, three dummy variables are set for monthly income, age and education level, corresponding to three categories except the baseline group. Only 2 dummy variables are set for 3 occupational categories. This coding method helps us to accurately estimate the influence of each category on the outcome variables in regression analysis.

3.3 Test of Binary Logistic Regression Equation

Test was carried out for binary logistic regression equation, the significance of the regression equation, and the goodness of fit of the regression equation. The Hosmer-Lemeshow Goodness-of-Fit test was used to confirm the goodness of fit between the fitting and the actual data, that is, the difference between the fitting and the actual situation.

Table 6. Model Summary

Chi-square	Cox & Snell R-Square	Nagelkerke R square	Significance
43.52	0.158	0.323	0.000

From the perspective of model fitness, the relevant results are presented in Table 6. Specifically, the chi-square test statistic was 43.52, corresponding to a P-value equal to 0.000. The significance test was successfully passed under the test condition that the significance level was set to 5%. This shows that the significant difference between the model and the observation data is not caused by random errors, and the model is reasonable to a certain degree.

Based on the current results, Nagelkerke R-square value is 0.323, which is relatively high, indicating that the logistic regression model can be used to explain the variation of dependent variables and can fit the data to a certain extent.

Table 7. Hosmer-Lemeshow Test

Chi-square	Degree of freedom	Significance
8.991	7	0.335

As can be seen from Table 7, the chi-square value is 8.991, and the obtained p-value is 0335 greater than 0.05. The original assumption should be acknowledged, that is, there is no significant difference between the predicted result and the actual value. In other words, the prediction of this model is very good.

3.4 Prediction Table

Test was carried out for the binary logistic regression equation, the significance of the regression equation, and the goodness-of-fit of the regression equation [9]. The

Hosmer-Lemeshow Goodness-of-Fit test was used to confirm the goodness of fit between the fitting and the actual data, that is, the difference between the fitting and the actual situation.

Table 8. Before Prediction

Measured		Predicted		
		Intention to buy		Correct percentage
		Yes	No	
Intention to buy	Yes	359	0	100%
	No	123	0	0%
Overall percentage				74.48%

Table 9. After Prediction

Measured		Predicted		
		Intention to buy		Correct percentage
		Yes	No	
Intention to buy	Yes	326	22	94.67%
	No	87	42	32.56%
Overall percentage				76.35%

Based on the data in Table 8 and Table 9, the overall prediction accuracy rates under the two modes are 74.48% and 76.35% respectively, which are relatively higher than before. This shows that the model is of certain practicality and reference value in predicting whether consumers will shop in E-commerce, and the prediction effect is ideal. However, the fluctuation of prediction accuracy in various samples cannot be ignored. In the future, the model needs to be optimized to improve the prediction accuracy of various consumers and provide stronger support for the formulation of e-commerce marketing strategies.

3.5 Binary Logistic Regression Model Output

Table 10. Variables in Equation

	B	Standard error	Wald	Degree of freedom	Significance	Exp(B)	95% confidence interval of EXP (B)	
							Lower limit	Upper limit
Age			0.500	3	0.796			
X2(1)	-0.112	0.344	0.101	1	0.708	0.889	0.324	1.767
X2(2)	-0.074	0.263	0.036	1	0.741	0.929	0.445	1.775
X2(3)	0.197	0.482	0.167	1	0.683	1.218	0.454	3.145
Education background			4.678	3	0.173			
X3(1)	-0.945	0.561	2.477	1	0.100	0.389	0.122	1.564
X3(2)	-0.578	0.379	2.398	1	0.016	0.561	0.341	1.565
X3(3)	-1.476	0.684	4.568	1	0.027	0.229	0.049	0.879
Occupation			1.221	2	0.531			
X4(1)	-0.390	0.792	0.149	1	0.670	0.677	0.126	4.013
X4(2)	0.349	0.555	0.656	1	0.346	1.418	0.565	3.789
Monthly income			2.639	3	0.451			
X5(1)	0.415	0.777	0.285	1	0.593	1.514	0.330	6.939
X5(2)	-0.843	0.596	2.000	1	0.157	0.430	0.134	1.384
X5(3)	-0.251	0.450	0.311	1	0.577	0.778	0.322	1.879
User Behavior X6	-1.496	0.582	7.687	1	0.005	0.224	0.137	0.646
After-sales service X7	1.899	0.908	4.528	1	0.025	6.679	1.175	41.902
Transportation X8	0.232	0.124	3.903	1	0.046	1.261	1.012	1.332
Constant	-1.438	2.591	0.416	1	0.535	0.237		

Based on the above analysis, the Wald statistic of the age variable was not significant(p>0.05), indicating that the effect of age on the dependent variable was not statistically significant. Among the educational variables, the p-values of X3(2) and X3(3) were smaller than 0.05, indicating that the effects of these two educational levels on the dependent variables were statistically significant. Specifically, the Exp(B)values of X3(2) and X3(3) were smaller than 1, indicating that the advantage of the occurrence of the dependent variable was reduced in these two education levels compared to the reference group; Neither level of the occupational variable was significant(p>0.05), indicating that the effect of occupation on the dependent variable was statistically insignificant. None of the three levels of the monthly income variable were significant(p>0.05), indicating that the effect of monthly income on the dependent variable was statistically insignificant; The p-value of the user behavior variable was smaller than 0.05, indicating that its effect on the dependent variable was statistically significant. The Exp(B)value was smaller than 1, indicating that the user behavior reduced the advantage of the occurrence of the dependent variable; The p-value of the after-sales service variable was smaller than 0.05, indicating that its influence on the dependent variable was statistically significant. The Exp(B)value was greater than 1, indicating that after-sales service improved the advantage of the occurrence of the dependent variable; The p-value of the transportation variable was smaller than 0.05, indicating that its effect on the dependent variable was statistically significant. The Exp(B)value was greater than 1, indicating that the transportation improves the advantage of the occurrence of the dependent variable.

The model constructed according to Table 10 is as follows:

Logit(p)=-1.4388-0.578*X3(2)-1.476*X3(3)-1.496*X6+1.899*X7+0.232*X8

According to Wald's value, after-sales service X7 greatly affected the problem whether customers would buy daily necessities on E-commerce platforms, followed by user behavior X6, education level X3(3), X3(2), and finally transportation X8. The constructed Logit model became a quantitative tool for predicting consumers' purchasing behavior on e-commerce platforms. Enterprises can input the values of variables such as different education levels, user behaviors, after-sales services and transportation, and calculate the probability of consumers buying daily necessities on E-commerce platforms with this model, so as to predict market trends and consumer demand in advance, and provide a strong basis for formulating accurate marketing strategies, product improvement strategies and customer service strategies.

4. Discussion

The variables of age, education, occupation, monthly income, user behavior, after-sales service and transportation were only considered in this study. Maybe other potentially important influencing factors were not included in the model, such as consumers' personal preferences, social media influence, promotional activities, etc [10]. The range of variables can be further expanded in future studies to more comprehensively reveal the influencing factors of consumers' purchase of daily necessities on e-commerce platforms. The research results may be limited by sample regions and groups, and the purchasing behavior of different consumer groups in different regions may vary. The sample scope can be expanded in future researches to cover consumers from different regions, with different ages, genders, and consumption levels to test the universality of the model and explore the differences between different groups in depth.

References

[1] Xu Lei, Wang Fang. Research on Consumer Purchase Decision-Making Mechanism under Live E-Commerce Mode. Journal of Commercial Economics, 2024, 30(05): 135-139.

[2] Li Xiang, Bai Yijie, Shang Meng. Analysis of Factors Affecting the Purchase Intention of New Energy Vehicles. Technology and Market, 2024, 31(08): 189-191.

[3] Zhu Lin, Huang Jian. Research on the Influencing Factors of User Stickiness of Social E-Commerce Platforms-Based on Multi-Case Comparative Analysis. E-commerce Letters, 2023, 28(11): 67-72.

[4] He Wei, Ma Jing, et al. Early Disease Screening Method Combining Binary Logistic Regression and Feature Engineering. Chinese Journal of Biomedical Engineering, 2024, 35(04): 345-350.

[5] Zhang Chen, Liu Xin. Construction and Empirical Analysis of Cross-Border E-Commerce Logistics Efficiency Optimization Model. Logistics Technology and Management, 203, 34(09): 56-60.

[6] Wang Ming, Zhao Qiang, Liu Ting. Research on Credit card Default Risk Prediction Model based on Binary Logistic Regression. Financial Technology Research, 2024, 32(06): 78-82.

[7] Chen, Yu, Zhang Hua, Li Na. Application Analysis of Binary Logistic Regression in Benign and Malignant Tumor Diagnosis. Journal of Medical Data Analysis, 2023, 29(03): 245-249.

[8] Yang Fan, Sun Li. Optimization and Practice of Binary Logistic Regression Model in Credit Risk Assessment. Review of Economy and Management, 2023, 38(12): 189-194.

[9] Zhou Tao, Wu Xiao. An Empirical Study on Binary Logistic Regression Algorithm in Users' Purchase Intention Prediction. Marketing Science, 2024, 31(01): 112-116.

[10]Luo Yun, Chen Hui. Research on Precision Marketing Strategy of E-Commerce Platform Based on Big Data. Modern Business, 2024, 33(02): 88-92.