Analysis and mining of user behavior on e-commerce platforms based on decision trees and RFM models

Xiuyou Zhou *

School of Computer Science and Technoloy, Kashi University, Kashi, China, 844000

* Corresponding Author Email: z1694638133@163.com

Abstract. User transaction behavior has driven the development of e-commerce. This article reveals the key role of the development of the e-commerce industry and user behavior in driving it through in-depth research on e-commerce sales data. Firstly, by grouping and summarizing user IDs through Pandas, a comprehensive user consumption situation was obtained, including consumption frequency, amount, etc., and user consumption behavior was visualized through visual means. On the basis of constructing effective features, Matplotlib was used to depict the number of users who receive coupons every week and day, providing a basis for the establishment of subsequent coupon distribution prediction models. By splitting the payment dates, an RFM model was constructed to analyze the purchase time period, frequency, consumption amount, and the proportion of each user's time behavior, forming a detailed user profile. Finally, the decision tree model is applied to predict the probability of coupon usage.

Keywords: E-commerce analysis, Decision tree model, RFM model.

1. Introduction

Commerce is one of the largest trading markets for users today, and under the influence of the Internet era, online consumption behavior satisfies users of all ages. The transaction behavior of users has driven the development of e-commerce. In order to analyze and mine user behavior on e-commerce platforms. Gao Yun used clustering algorithms and BERT models to study the behavior and mining of e-commerce platform users. He proposed an improved K-means clustering algorithm combined with density Canopy algorithm to divide social e-commerce groups. However, the dataset used was a privately collected dataset with significant errors [1]. Ye liming uses big data frameworks Hadoop and Spark to analyze user behavior and data mining on e-commerce platforms, in order to guess and recommend user preferences and products to meet user needs. However, Spark's learning curve is relatively steep, requiring a certain amount of time and effort to learn and master [2]. Wang Zhibo conducted research on the improved Apriori algorithm based on Hadoop. Further exploration is needed to determine whether the operational efficiency of the algorithm can still meet the needs of production and daily life after the data collective reaches TB level [3]. The approach to solving this problem in this study is as follows: Firstly, define a cal_ The PV() function is used to calculate the monthly payment amount for each user; Draw line charts and histograms to better understand user consumption patterns. Next, extract four fields including user ID, merchant ID, usage status, and month from the source data for secondary conditional filtering, and study the behavior of users, merchants, and coupons and extract effective features. Furthermore, construct an RFM model for customer value analysis and draw user profiles to analyze user characteristics [4-6]. Finally, construct a binary decision tree model and perform probability prediction [7-10].

2. Analysis of user consumption behavior

2.1. Basic problem-solving idea

Define a cal_ The PV () function is used to calculate the monthly payment amount for each user; Draw line charts and histograms using the matplotlib library to better understand user consumption patterns. Group by receiving the value of the key and return the monthly payment amount of the user. Draw a monthly user payment volume chart.

2.2. Data preprocessing

Firstly, use the Pandas library in Python to read user data saved to the local e-commerce platform. Then, use the info method to check the type of data (Data source: Data from a certain e-commerce platform of the 2023 National College Student Data Analysis Competition organized by the China Financial Analysis Association). Due to empty values in the usage status, fill them with non-negative values and handle duplicate and outlier values. Finally, remove the day, minute, and second data from the date format and only retain the year month format.

2.3. Analysis of User Consumption Behavior

2.3.1. Monthly payment statistics of users

We define cal_ The pv() function groups by receiving the value of the key and returns the monthly payment amount of the user. Draw a monthly user payment volume chart.

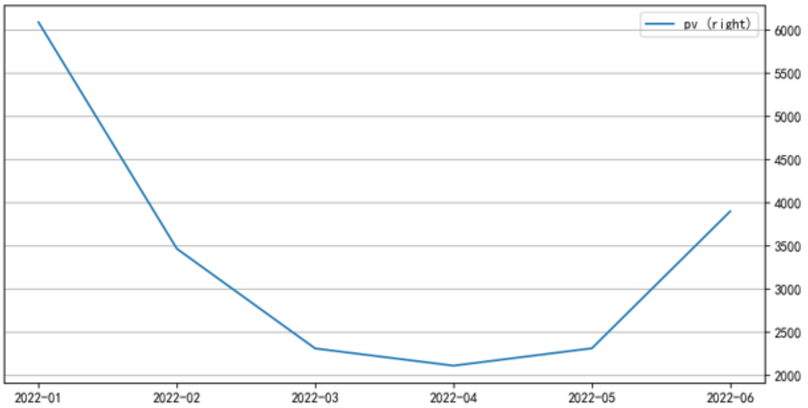

Figure 1. User payment line chart.

Through Figure 1, we can analyze that the payment volume of users reached its peak in early January 2022, as it happened to be on New Year's Eve, with over 6000 users shopping and making payments on e-commerce platforms. From mid to late January until April, user shopping behavior significantly decreased compared to before, and reached a low point in April, as most users may be working and have no time to shop during this period. From the end of April, user payment volume began to increase and continued to rise sharply in May, possibly due to the increase in holidays leading to a rebound in user payment volume.

2.3.2. User consumption frequency and amount

Group the user ID field and use the purchase quantity field as the query criterion to draw a histogram of user consumption frequency and user amount frequency.

As shown in Figure 2, the majority of users spend less than 2 times, with a small portion of the data causing interference.

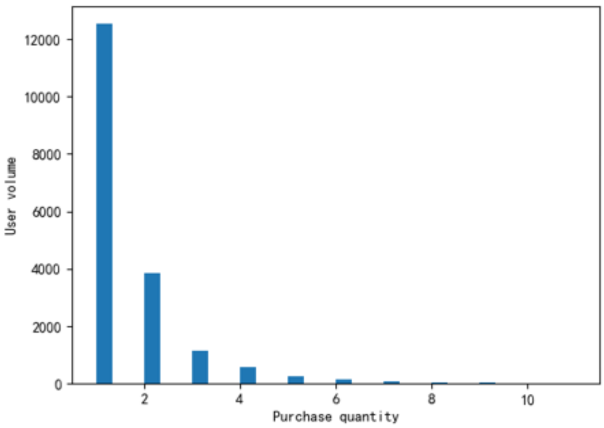

Figure 2. User consumption frequency histogram

2.3.3. Proportion of cumulative user consumption

Sort by consumption amount and use the cumulative sum function to calculate the proportion of user consumption.

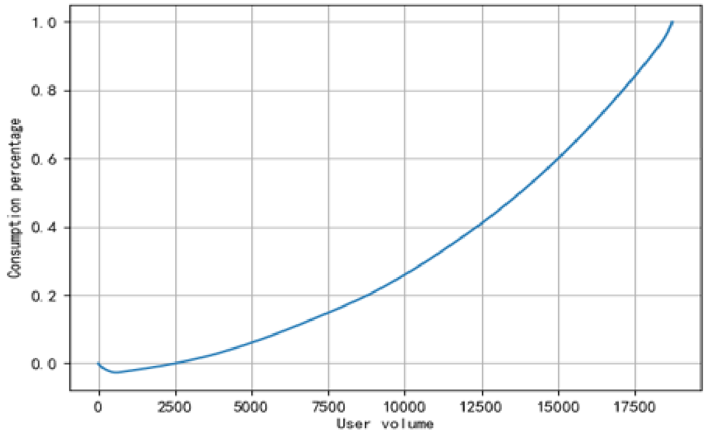

Figure 3. Line chart of the proportion of user consumption.

From Figure 3, we can see that there are a total of 20183 users, and 50% of the users, including those before 10091st, only account for nearly 28% of their consumption. The top 7500 users contribute 60% of their total consumption, which means maintaining these top 7500 users can achieve 60% of the KPI.

3. Feature screening

3.1. Basic problem-solving ideas

Filter out four fields from the source data: user ID, merchant ID, usage status, and month, and process the data again through secondary criteria.

First, deduplicate and only retain the data that appears for the first time. We studied the behavior of users and extracted the following six effective characteristics: the number of times users visited a specific merchant, the total number of times users made purchases, the total number of times users made purchases and used coupons, the number of times users received coupons, the proportion of times users wrote off coupon purchases to the total number of times users spent, the proportion of times users wrote off coupon purchases to the total number of times users received coupons, and finally, Use pd.merge() to merge the effective features of the appeal and save it.

Next, study the behavior of merchants. The following five valid features were extracted: total sales of merchants, sales of coupons cancelled by merchants, issuance of coupons by merchants, usage rate of coupons issued by merchants, and proportion of coupons cancelled by merchants. Finally, the appeal valid features were merged using pd.merge() and saved.

Finally, study the distribution and use of coupons and extract relevant effective features. What day of the week do users receive coupons, the number of users who receive coupons on each day of the week, and whether users receive coupons on weekends. Merge effective features extracted from users, merchants, and coupons and save them.

3.2. User feature extraction

Extract effective user features: the number of times the user has visited a specific merchant, the total number of times the user has made purchases, the total number of times the user has made purchases and used coupons, the number of times the user has received coupons, the proportion of times the user has verified coupon purchases to the total number of times the user has received coupons, and the proportion of times the user has approved coupon purchases to the total number of times the user has received coupons. Finally, use pd.merge() to merge the effective appeal features and save them. Some effective feature data of users can be found in Table.1 below.

Table 1. User Effective Characteristics.

User id	Number of uses	Number of times to claim	Use ratio	Claim ratio
Xiao Xi Jia	3	3	1.0	1.0
Bud Zhang	1	1	1.0	1.0
2day day up	1	1	1.0	1.0
9chs05	4	4	1.0	1.0
ejaneinmecn	1	1	1.0	1.0

3.3. Merchant feature extraction

Extract effective features of merchants: Filter out three fields from the source data: merchant ID, usage status, and monthly payment date, and then filter the data again through secondary conditions for processing. And extract the following five characteristics: total sales times of merchants, sales times of coupons cancelled by merchants, times of coupons issued by merchants, usage rate of coupons issued by merchants, and proportion of coupons cancelled by merchants. Finally, merge the effective features of the appeal using pd.merge() and save it. Partial data can be found in Table.2.

Table 2. Effective characteristics of merchants.

Merchant ID	Write-off sales frequency	Usage rate	The proportion of write offs
100110	33	1.0	1.00
103840	1	1.0	1.00
100017	133	1.0	0.98
100180	24	1.0	1.00
102464	1	1.0	1.00

3.4. Coupon feature extraction

Firstly, use the sequence method dt. dayof week to convert the date to the day of the week. Then use get in pandas_ the dummies () method heats up the days of the week. Using the list derivation formula again, add the corresponding column names for each week to obtain the effective features of the user's coupon collection on the day of the week. Finally, use the number of weeks (Monday to Sunday) as the field to filter and calculate the number of users who receive coupons for each day of the week, and use pyecharts to draw a weekly coupon bar chart.

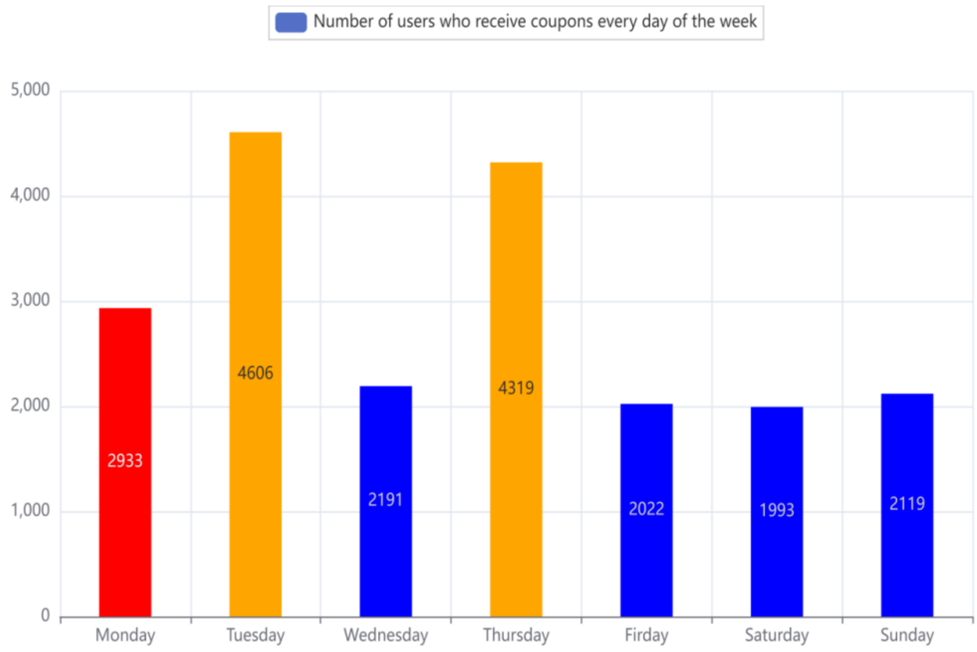

Figure 4. Number of users who receive coupons in four weeks.

From the bar chart in Figure 4, we can analyze that the number of users who receive coupons on Monday, Tuesday, and Thursday is relatively high, with 2933, 4606, and 4319 users respectively. The remaining days are all below 2500 users.

4. RFM model construction

4.1. Basic problem-solving ideas

Build an RFM model for customer value analysis and draw user profiles to analyze user characteristics. Filter out users with purchasing behavior, construct panel data, divide time periods into early morning, morning, noon, afternoon, and evening, and calculate the number of purchases made by users in each time period. Calculate the user's recent purchase time, purchase frequency, and consumption amount to construct an RFM model for customer value analysis, and use pie charts in pyecharts to draw a user profile.

4.2. Analyze user purchasing behavior

4.2.1. Statistical user time behavior

Statistics on the usage status of coupons, filtering out customer data orders with purchasing behavior, calculating the time period during which users made more purchases, and using pyecharts to draw a user time pie chart.

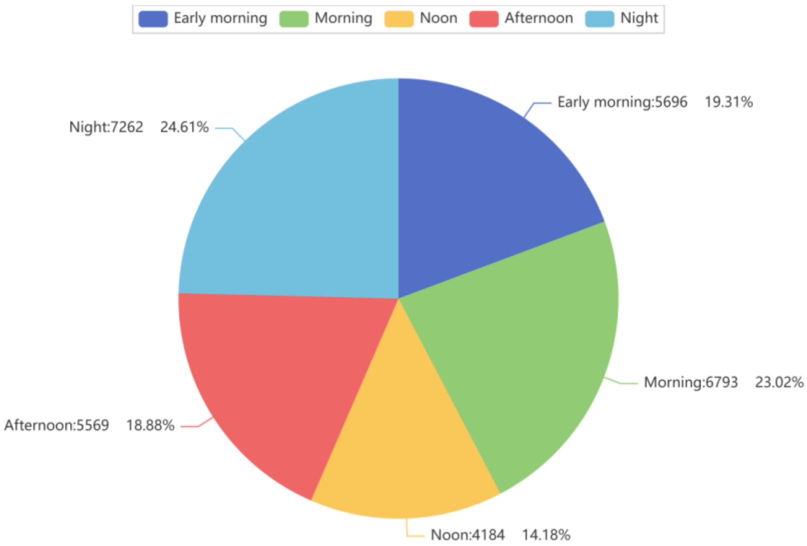

Figure 5. Proportion of user time behavior.

From Figure 5, we can see that users are most active during the evening hours and least active during the noon hours.

4.3. Building an RFM model

4.3.1. Calculate the most recent purchase time

Calculate the final purchase time for each user, merge the earliest and latest purchase times into one table (through the user ID field), obtain the most recent purchase time and name its field as days, and finally merge the data.

4.3.2. Building an RFM model

Using the most recent purchase time (R) calculated in 4.3.1, and this step calculates the purchase frequency (F) and consumption amount (M) through grouping and summing, merge the data and fill in the blank value as 0. Finally, use the average value to define and extract RFM indicators and draw an RFM user profile.

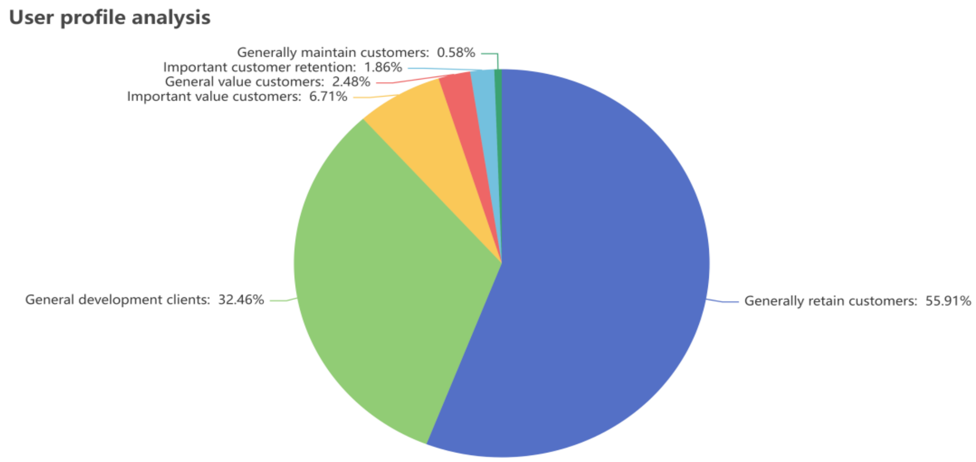

Figure 6. User Profile.

From Figure 6, we can clearly see that the proportion of general development users and general retention users on this e-commerce platform is 32.46% and 55.91%, respectively, while the proportion of important value users is relatively low at 6.71%. Corresponding strategies should be formulated to improve user loyalty.

5. Decision tree model construction and result analysis

5.1. Comparison between binary and ternary classification models

Evaluate the performance of the two models based on the overall prediction accuracy, as well as the accuracy and recall of each classification. Model 1 represents a three-class decision tree model, while Model 2 represents a two-class decision tree model.

Table 3. Accuracy Comparison.

model	Accuracy 1	Accuracy 2	Accuracy 3
1	0.0%	99.71%	99.15%
2	100%	99.7%	none

From Table.3, it is easy to analyze that Model 2, the binary classification model, achieved better performance than the three-classification model in terms of various indicators. Specifically, both the status of receiving coupons but not using them and the status of not receiving coupons were classified as negative, while the decision tree model of receiving and using coupons was classified as positive and achieved better results. Based on the above, a binary classification model was ultimately chosen to predict the probability of coupon issuance.

5.2. Building a binary decision tree model

Establish a decision tree model using DecisionTreeClassifier, divide feature variables and target variables, and divide the training and testing sets for fitting training. Draw a line graph of their true and predicted values, and use this model to predict whether to issue merchant vouchers to the user.

5.2.1. Implementation of Decision Tree Model

Convert the three-classification problem into a two-classification problem. Divide the non use and non use of coupons into negative classes (0, 2), and the non use and non use of coupons into positive classes (1). Then, divide the training and testing sets into feature variables and target variables. Finally, perform fitting and training (70% for the training set and 30% for the testing set).

5.2.2. Decision Tree Model Prediction Results

Predict the probability of users using coupons. Compare the predicted values with the actual values and draw a line graph using pyecharts.

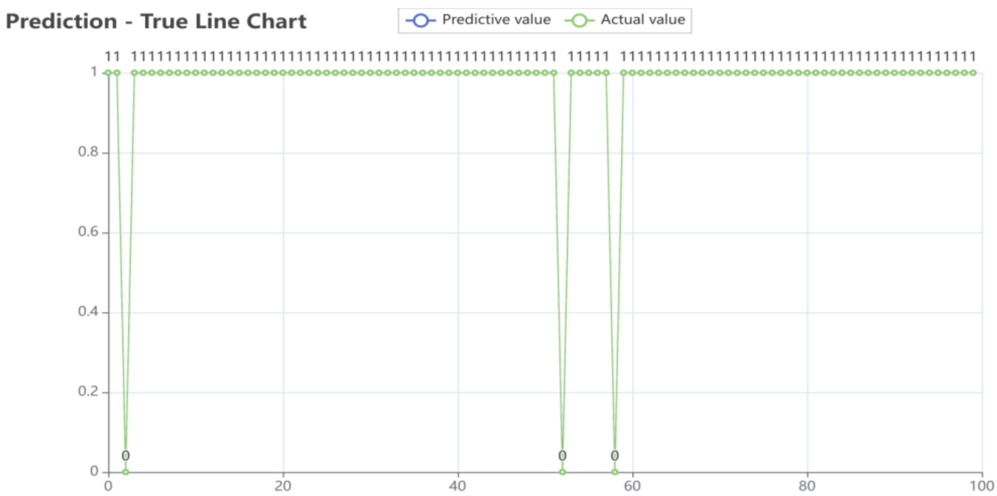

Figure 7. Comparison between predicted and true values.

From Figure 7, it can be seen that the predicted values are basically consistent with the true values. The predicted values are generally consistent with the true values in the direction of the graph, indicating that the model has good predictive performance.

Due to the large amount of data, only the probability of 10 users using coupons is predicted here.

Table 4. Prediction Results.

Merchant ID	User id	Probability of not receiving vouchers	Use rate
104594	Hualiu 3	0	1
100036	Leisure ethnic style	0	1
100742	8bear81	1	0
100105	4li616878	0	1
100027	0 Small Business 80	0	1
100143	asaralianln	0	1
100966	iheya4	0	1
100600	L Niu Niu_ Ane	0	1
100012	0 Red faced Eggs 20	0	1
100001	Make the Angelo	0	1

From Table 4, we can see that the probability of users receiving and using coupons has increased significantly compared to before model optimization, indicating that our model has overall achieved the expected predictive effect.

6. Conclusions

Under the influence of the internet era, user transaction behavior has driven the development of e-commerce. This article reveals the key role of the development of the e-commerce industry and user behavior in driving it through in-depth research on e-commerce sales data. Firstly, by grouping and summarizing user IDs through Pandas, a comprehensive user consumption situation was obtained, including consumption frequency, amount, etc., and user consumption behavior was visualized through visual means. On the basis of constructing effective features, Matplotlib was used to depict the number of users who receive coupons every week and day, providing a basis for the establishment of subsequent coupon distribution prediction models. By labeling users, an RFM model was constructed to analyze their purchasing time period, frequency, consumption amount, and the proportion of their time behavior, resulting in a detailed user profile. Finally, the decision tree model is applied to predict the probability of using user coupons.

References

[1] Gao Yun. User Behavior Analysis and Research Based on Social E-commerce Platforms [D]. Beijing University of Posts and Telecommunications, 2021. DOI: 10.26969/dcnki.gbydu.2021.001850

[2] Ye Liming. Analysis and Research Based on Spark E-commerce User Behavior Data [D]. Shenyang Normal University, 2020. DOI: 10.27328/d.cnki.gshsc.2020.00208

[3] Wang Zhibo. Research on User Data Mining of E-commerce Platforms Based on Hadoop [D]. North China Electric Power University, 2021. DOI: 10.27139/dcnki.ghbdu.2021.000366

[4] [Liu Ruiqi, Song Zikun. Analysis of Tmall User Value Based on RFM Model [J]. Economist, 2022, (12): 243-244

[5] Shi Aoxiang, Zhang Jie. Research on E-commerce User Value Classification Based on Improved RFM Model [J]. Computer Technology and Development, 2022, 32 (12): 123-128

[6] Chen Danhong, Peng Zhanglin, Wan Dequan, et al. User value identification and segmentation of crowdsourcing platforms: based on an improved RFM model [J]. Computer Science, 2022, 49 (04): 37-42

[7] Xu Xiaoping, Ma Wenjie. Quantitative Analysis of Loan Default Rate for Non listed Small and Medium sized Enterprises: Analysis Based on Discriminant Analysis and Decision Tree Model [J]. Financial Research, 2011, (03): 111-120

[8] Yan Ruiping, Wang Xiliang, Yao Fanxia, et al. Comparison of the effectiveness of decision tree model and logistic regression analysis model in identifying risk factors for hypertension [J]. Chinese Journal of Disease Control, 2022, 26 (02): 218-222. DOI: 10.16462/j.cnki.zhjbkz.2022.02-016

[9] Cui Jingjing, Hu Zewen, Ren Ping. Research on identifying potential "boutique" papers in the field of artificial intelligence based on decision trees and logistic regression models [J]. Intelligence Science, 2022, 40 (05): 90-96. DOI: 10.13833/j.issn.1007-7634.2022.05.012

[10] He Jian. Research on Decision Tree Optimization Algorithm [D]. Hefei University of Technology, 2004