Comprehensive Analysis and Mining Big Data on Smart E-commerce User Behavior

Yinggui Wang, Ben Wang, Yuxin Huang

Department of information science and engineering, Hangzhou Normal University, Hangzhou 311100, China
2276598112@qq.com, 20170056@hznu.edu.cn, 616710866@qq.com

Abstract. With the gradual development of big data technology and the rapid growth of e-commerce industry, big data comprehensive analysis technology is particularly important. Therefore, the user behavior and historical shopping data have been focused in recent progress of e-commerce industry. Based on e-commerce user data sets, this paper analyzes the user behavior based on Pareto principle (80-20 rule) and RFM data models. By statistically analyzing the behavior characteristics of users, user groups, user preferences, and purchase behavior characteristics, several models and research algorithms have been proposed to service targeted products and functions for e-commerce users, and figure out further research topics of e-commerce.

1. Introduction
With the increasing popularity of the Internet and computer information technology, a new business model for e-commerce is developing rapidly. It has attracted the attention of many domestic scholars. They have made a lot of good suggestions, new evaluation system and models by studying the network integrity, influencing factors of store performance and reputation evaluation system in the development process of e-commerce in China. Fan Wenjie explored the time interval between user review and purchase and the factors influencing the timeliness of customer feedback in e-commerce platform, so as to understand the law of the time interval of consumer review publication, explore the factors that affect the timeliness of consumer feedback, and master the motivation of influencing the release of feedback [1]. Zhou Pei [2] organized user data sets of ten universities in Nanjing as examples, based on TAM and ECM models, put forward the influencing factor model of shopping app users’ continuous usage, and conducted empirical test on the model. Zhang Chunlian analyzed the traditional customer purchase behavior prediction model, improved SMC model, built BG / NBD model, and verified its effectiveness [3]. According to the significant difference between online and offline sales of C2C e-commerce, Tian constructed a prediction model using user transaction frequency and time [4]. Zuo utilized SVM model to process RFID data generated by consumers in supermarkets [5]. As a result, SVM model significantly improved the prediction accuracy of consumer purchase behavior, compared with linear regression prediction models. HJ Chang proposed a prediction model of potential purchase behavior through cluster analysis and association rule analysis, which is used to locate the attributes of potential customers and detect their interest in goods [6].

From current literatures, most of them are based on the impact of a single factor on user behavior, but different scholar draw different conclusions [7]. Based on this situation, it is quite necessary to further mine the influencing factors of user behavior [8].
2. Data and Environment

2.1. Software and hardware environment
A computer with windows 10 OS system was used in this experiment, and the memory is 8G. Anaconda is an integrated scientific computing environment. In order to facilitate the later data visualization, the conda is executed to install numpy, scipy, and panda commands & toolkits at the same time.

The compiler is an open source web application jupyter notebook, which allows the creation and sharing of documents, including code, equations, visualization and narrative text. It is able to achieve data cleaning and conversion, numerical simulation, statistical modeling, data visualization, machine learning, etc.

2.2. Environmental data set and preprocess
Data set: https://tianchi.aliyun.com Between November 26, 2017 and December 3, 2017, all behaviors of about one million random users with behaviors (including click, collect, Additional purchase, purchase). The meanings of the data fields are listed in Fig. 1.

| Data Field  | Description |
|-------------|-------------|
| User ID     | String type, user ID |
| Commodity ID| String type, commodity ID |
| Commodity category ID| String type, commodity category ID |
| Behavior type| String, enumeration type, including ('pv browse', 'buy purchase', 'cart add purchase', 'fav collection') |
| time stamp  | Integer, the timestamp of the occurrence of the behavior |

Fig. 1 Meaning of data fields

Firstly, a series of data cleaning work is carried out, such as deletion of duplicate values, processing of missing data, data type conversion, data sorting and exception handling. The statement parameters and initialization of data are listed in Fig. 2.

| Statement | Description |
|-----------|-------------|
| df.duplicated().sum() | Delete duplicate data values |
| df.isnull().sum() | View missing values |
| data['time']=pd.to_datetime(data['time'], unit='s')+datetime.timedelta(hours=8) | Data type conversion of time column |
| data=data.sort_values(by=['date', 'hour'], ascending=False) | Sort the data in descending order according to date and hour |
| df=(data.loc[:, 'date']>'2017-11-24')&(data.loc[:, 'date']<'2017-12-04')data1=data.loc[df, : ] | Data beyond this time range will be deleted |

Fig. 2 Initialization data statement parameters

After data cleaning, the initialization of user behavior data is completed. And then a series of Taobao user behavior data analysis can be carried out with Python language. Data analysis includes flow index analysis, conversion rate analysis, sales index, and user value analysis. The specific data models and algorithms will be discussed in the next section.

3. Data Models and Algorithm Optimization

3.1. Pareto principle (80-20 rule)
In any group of things, the most important part is about 20%, and the remaining 80% is less important. It is also called Pareto principle (80-20 rule).
For a long time, under the influence of production concept and product concept, enterprise marketing managers have been commonly concerned about the sales of products or services, and new customers. However, comparing with new customers, old customers often bring more benefits to enterprises. Smart enterprises try to create new customers, and at the same time, they are good at transforming customer satisfaction into lasting loyalty. They add extra interests for old customers, just as new customers, and establish long-term relationship with customers. More important, the user behavior data are analyzed by Pareto principle. The relevant algorithm is shown in Fig. 3 and results are shown in Table. 1.

| Function: Analyze user behavior |
|---------------------------------|
| 1. Cumulative sales purchase classification |
| 2. \( \text{value}_8 = \text{data}_{-}	ext{Category}[\text{purchase quantity}]. \sum() \times 0.8 \) |
| 3. \( \text{value}_{10} = \text{data}_{-}	ext{Category}[\text{purchase amount}]. \sum() \) |
| 4. \( \text{[cumulative purchase]} = \text{Data}_{-}	ext{Category}[\text{purchase quantity}]. \text{Cumsum()} \) |
| Map (X: 'Top 80%' if \( x \leq \text{value}_8 \) else 'last 20%') |

Fig. 3 Classification of cumulative sales volume purchase

## Tab. 1 Classification of cumulative purchase volume

| Behavior type | Hits | Add cart volume | Purchase volume | Conversion rate | Interest rate | Cumulative purchase | Classification |
|---------------|------|-----------------|-----------------|-----------------|---------------|---------------------|----------------|
| 6344          | 7230 | 322             | 174             | 0.024066        | 0.044537      | 174                 | Top 80%        |
| 1863          | 3207 | 804             | 140             | 0.004365        | 0.025070      | 314                 | Top 80%        |
| 5232          | 1132 | 374             | 138             | 0.012183        | 0.033018      | 452                 | Top 80%        |
| 3424          | 4298 | 128             | 110             | 0.025593        | 0.029781      | 562                 | Top 80%        |
| 3472          | 4345 | 127             | 104             | 0.023936        | 0.029229      | 666                 | Top 80%        |

3.2. Analysis of RFM model for user value

RFM model is a key tool to measure customer value and customer profitability. The model describes the user value status through three indicators: the recent transaction behavior (recency), transaction frequency (frequency), and transaction amount (monetary). Normally, three indicators are divided into several intervals for scoring, and the valuable users are ordered and classified by the scores.

(1) Statistics F value: in the users who have purchase behavior, count the number of users to buy in Figure 4.

| Function: Analyze user behavior |
|---------------------------------|
| 1. The number of purchases per user determines the F value |
| 2. \( \text{userBuyNum} = \text{buyDF}'\text{user ID'} . \text{value_counts()}. \text{sort_values()} \) |
| 3. Descriptive statistics of user purchases |
| 4. \( \text{userbuynum. describe()} \) |
| 5. mean 3.008202 std 2.974372 min 1.000000 25% 1.000000 75% 4.000000 max 84.000000 |

Fig. 4 Customer purchase times

Here, the maximum number of user purchases is 84 times, but about 50% of the users only purchase 2 times or less, and 75% of the users only purchase 4 times or less.

(2) Statistics F value: Calculate the time between the last consumption and the last consumption. Firstly, the R values are grouped, and the corresponding R values are set in different time periods. As shown in Table 2.

| Tab. 2 Classification of F values |
|---------------------------------|
| Purchase times | R value | Last purchase time | F value |
| (40, 84]  | a      | 2017/12/3          | A       |
Table 1: RFM model description

| (4,40) | b | 2017/12/1-2017/12/2 | B |
| (2,4) | c | 2017/11/29-2017/11/30 | C |
| 2 | d | 2017/11/27-2017/11/28 | D |
| 1 | e | 2017/11/25-2017/11/26 | E |

The last purchase time of each user determines the R value, we assume that each corresponding time period corresponds to an R value. The corresponding r value and F value are analyzed by matrix transformation. The code description is shown in the figure 5.

Function: Analyze user behavior

```python
1 Begin
2 The last purchase time of each user determines the r value
3 for u in userBuyNum.index:q=data_buy.loc[:, 'user_id']==u
4 if d=='2017-12-03': R='A'
 elif d in ('2017-12-02', '2017-12-01'): R='B'
 elif d in ('2017-11-30', '2017-11-29'): R='C'
 elif d in ('2017-11-28', '2017-11-27'): R='D'
 else: R='E'
5 Comprehensive evaluation value of RFM model
6 scoreDf['RF']=scoreDf['Rscore']+scoreDf['Fscore']
7 Statistics by RF value
8 scoreDf['RF'].value_counts().sort_index()
9 End
```

Fig. 5 RFM model description

4. Results Analysis

In Tab. 1, the first 80% of sales are contracted by about 30% of commodity categories, which is close to the 80-20 rule. In fact, it is obvious that nearly 20% of the sales are provided by 80% of the commodity categories. In traditional retail industry, because of the high cost, only top 20% products get profits. But in e-commerce, the space cost is reduced or even zero, so that the last 80% of the goods can also be sold out. Therefore, the optimization and recommendation of the long tail part of the goods can bring greater benefits to enterprises.

In Fig. 6, there are four types of user behavior: Click, collect, add cart, and purchase. In which click behavior accounts for the highest proportion, the other three types account for a very low proportion, and the conversion rate of purchase is also very low. The possible reason is that most of the products that users browse on the platform are not purchased. The e-commerce platform can push the products that users may need according to the user profile, and optimize the filtering function of the platform, so as to facilitate users to find suitable products and improve the conversion ratio.

In Tab. 3, users with RF value of "Aa" have recently spent a lot of time and bought many times, so they are VIP customers. If the RF value is "Ca or Da", it is an important customer to keep in touch. If the recent consumption time is long, and the consumption frequency is very high, it indicates a loyal customer, who has not come for a period of time, and needs to actively keep in touch with them. The customers with RF value of "Ab, Ac, Ad, Ae" are important development customers. They recently spent a short time, but the consumption frequency is not high (the closer to e, the lower the frequency), and the loyalty is not high. Users with great potential should focus on development customers. The customers with RF value of "Ee" are the most likely to leave. They only buy once and spend the longest time, and we should take the initiative to contact them, otherwise these customers may be lost.

In Fig. 7, PV and UV are quite similar during 0-5 am. The number of visits is relatively small. However, at about 18:00 p.m., PV fluctuated violently, but UV was not obvious. Therefore, we should pay attention to the recommendation of the key time period when making user recommendation.
5. Conclusion

In this paper, we analyzed user behavior data, the factors affecting the user purchase. We also mined user behavior characteristics, preferences, potential needs, and behavior habits. The research can help enterprises truly understand users, accurately locate marketing target users, improve user retention rate, and bring higher profits for enterprises. Through the definition of target users or groups, it helps us to achieve accurate product information push and efficient marketing. In conclusion, the models and algorithms have a great impact on platform managers and potential consumers. Next step, multiple e-commerce user behaviors will be combined and organized together to find more interesting useful
results. In addition, clustering algorithm would improve the efficiency and accuracy of user behavior analysis. These will be our further research directions in the near future.

Acknowledgments
This paper is funded by projects: Basic public welfare research program of Zhejiang Province (lgf19f020011); Independent application project of agricultural and social development scientific research in Hangzhou (20191203b14).

Reference
[1] Fan Wenjie, Deng Weihua. Characteristics and influencing factors of timeliness feedback behavior of e-commerce users [J]. Agricultural Library and information, 2019, 31 (11):64-71.
[2] Zhou Pei, Fu Suyun, Zhao Yuechun. An Empirical Study on the influencing factors of shopping app users' continuous use [J]. Journal of Nanjing Normal University (NATURAL SCIENCE EDITION), 2020, 43 (02):140-148.
[3] Zhang Chunlian. BG / NBD prediction model of customer purchase behavior and its application research [D]. Harbin Institute of technology, 2006.
[4] Ye Tian, Zuoliang Ye, Yufei Yan, Miao Sun. A practical model to predict the repeat purchasing pattern of consumers in the C2C e-commerce[J]. Electronic Commerce Research, 2015, 15(4):571-583.
[5] Zuo, Yi, Ali, A.B.M. Shawkat Yada, Katsutoshi. Consumer Purchasing Behavior Extraction Using Statistical Learning Theory[J]. Procedia Computer Science, 2014, 35:1464-1473.
[6] Chang HJ. Hung LP., Ho CL. An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis. [J]. Expert Systems with Applications, 2007, 32(3):753-764.
[7] Wang Ben, Zhou Wenli. Comprehensive Integrated Platform for Garbage Classification in Reduction, Innocuity, and Resource[J]. Computer, Communication and Network Technology, 2018, 26-31.
[8] Wang Ben, Zhou Wenli, Li Zhihua. Research on Distributed Intelligent Mattress on the Internet of Things[J]. Computational Modeling[J], Simulation and Applied Mathematics, 2018, 323-327.