Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **7**, Number **4**, November **2022**

Consumption Behavior Analysis of E-commerce Users Based on K-means Algorithm

Junli Zhang*

Major in Data Science and Big Data Technology
Xi'an Eurasia University
No.8, Dongyi Road, Yanta District, Xi'an, China
zll319@qq.com

Jingyang Wu

Xi'an Eurasia University
No.8, Dongyi Road, Yanta District, Xi'an, China
156588351@qq.com

Chenyan Gao

New Image International
Auckland, New Zealand
24200926@qq.com

*Corresponding author: Junli Zhang
Received July 6, 2022, revised August 21, 2022, accepted September 29, 2022.

Abstract. *At present, the pattern of China's e-commerce retail market has been basically formed, Ali already occupies the first place in the market with more than half of the market share, and Jingdong accounts for less than 50 percent, these two companies together account for more than three-quarters of the overall market share in China. However, our country's national online shopping market is about to become saturated. Therefore, under the premise of the rapid development of the era of big data, the competition of e-commerce platforms has gradually changed from the competition of the number of users to the competition of refined user management. This paper constructs consumer behavior indicators through feature engineering, and then uses the K-means algorithm to cluster customers into four categories, and name these customers as "iron powder customers" ,"general customers", "develop customers" and "zombie customers" according to their importance to the enterprise from high to low, and analyze the characteristics of different customer groups at the same time, and give corresponding marketing suggestions according to different customer groups..*
Keywords: consumer behavior analysis; feature engineering; K-means Algorithm

1. **Introduction.** E-commerce refers to the realization of electronic and grid operation of all aspects of traditional commerce by carrying out commodity trading activities and related business activities in the form of e-commerce transactions on the network, intranet and value-added networks. According to the National Bureau of Statistics, in 2021, the national online retail sales reached 13.1 trillion yuan, an increase of 14.1% year-on-year, and the growth rate was 3.2 percentage points faster than that of the previous year. Among them, the online retail sales of physical goods reached 10.8 trillion yuan, exceeding 10 trillion yuan for the first time, an increase of 12.0% year-on-year, accounting for 24.5% of the total retail sales of social consumer goods, and the contribution rate to the growth

of total retail sales of social consumer goods was 23.6% [1]. This series of phenomena fully supports the continuous improvement of China's e-commerce penetration rate, and online consumption has also played a pivotal role in China's consumption upgrade.

At present, the layout of China's e-commerce retail market has been basically formed, of which Ali occupies the first place in the market with more than half of the share, JD.com accounts for only 50% , and the two companies together occupy more than three-quarters of China's overall market share. In the second echelon, the fastest increase in the number of users is Pinduoduo, and now it has become the leader of the second echelon, followed by e-commerce platforms such as Suning and Vipshop [2]. In addition, in recent years, new e-commerce platforms have been pouring into the market, which can be described as quite fierce competition [3]. Among them, leading e-commerce companies such as Ali and JD.com have also thought of breaking through, and they have successively launched the new concept of community group buying. However, it was revealed that the dishes were not fresh, and the operation ability of the regiment leader was insufficient [4]. And with the release of relevant policies on community group buying in the first half of 2020, this new online shopping method has also been strictly regulated. At this point, the wave of community group buying has come to an end for the time being, and the development has also tended to be flat.

In the case of the rapid development of the big data era and the development of new fields, the competition of e-commerce platforms has gradually changed from the competition in the number of users to the competition for user quality, that is, the implementation of refined management of users, which needs to be classified according to the big data characteristics of users such as markets, channels, and usage behaviors, and targeted marketing activities for different categories of customers to provide them with more personalized and differentiated operational strategies, so as to finally stand out from other angles in this fierce competition [5]. The realization of this method requires a large amount of data as a support, Taobao with its leading position as an e-commerce platform for many years, should have a large amount of consumer-related behavior data [6]. Related researches on ecommerce platforms mainly adopt classification models based on supervised learning [7-13]. However, for new e-commerce products, it is difficult to obtain target variables due to data accumulation problems in the initial stage of operation, so user behavior classification based on unsupervised learning is of more commercial value.According to this point, this paper will use the K-means algorithm to classify customers by using the relevant consumer behavior data of consumers on Taobao to choose different marketing methods for different categories of customers, so as to achieve precision marketing.

2. K-means algorithm Several definitions and theorems.

The K-means algorithm is also one of the most commonly used clustering algorithms, and is often widely used due to its simplicity and speed. In the business scenario of e-commerce, the information generated by consumers when purchasing can divide the customer groups through cluster analysis and carry out different business activities to each customer in a more accurate manner.

2.1. Optimal K-value determination method.

The optimal K-value determination method in K-means clusters is the elbow method [14]. The core indicator of the elbow method is the calculation of the error sum of squares (SSE), which is calculated as follows:

$$SSE = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2 \tag{1}$$

As the number of clusters k increases, the sample division becomes progressively finer, each cluster becomes more aggregated, and the error-squared sum SSE becomes progressively smaller. When k is less than the actual number of clusters in the cluster [15], the cohesion of the entire population is greatly improved because of the increase in k , so the reduction of SSE is more dramatic [16]. When k reaches the true number of clusters of clusters, if k continues to increase, the return on the degree of aggregation will rapidly shrink, but at the same time the decline in SSE rises sharply and tends to smooth as k continues to rise. At this time, there will be an inflection point similar to the human elbow, which we call the elbow, and the k value corresponding to this elbow is the optimal k value of this set of data.

2.2. **Criteria for evaluating K-means clustering results.** Mutual Information (MI): Refers to the degree of proximity between two labels of the same data information content, that is, it is used to measure the similarity of the distribution of two data [17]. To judge the clustering effect through mutual information, you must first grasp the actual category information. Assuming that U and V are the distribution of N sample labels, the entropy of the two distributions is as follows:

$$H(U) = \sum_{i=1}^{U} P(i) \log(P(i)) H(V) = \sum_{i=1}^{V} P'(j) \log\left(P'(j)\right) \tag{2}$$

Mutual Information (MI) between U and V is defined as:

$$MI(U,V) = \sum_{I=1}^{U} \sum_{j=1}^{V} P(I,J) \log\left(\frac{P(i,j)}{P(i)P'(j)}\right) \tag{3}$$

Normalized Mutual Information:

$$NMI(U,V) = \frac{MI(U,V)}{\sqrt{H(U)H(V)}} \tag{4}$$

Adjusted Mutual Information:

$$AMI = \frac{MI - E|MI|}{MAX(H(U), H(V)) - E|MI|} \tag{5}$$

MI and NMI range $[0,1]$, AMI range $[-1,1]$, generally the larger the value, the better the clustering effect.

3. **Feature engineering.** In order to understand the behavior characteristics of different consumers more accurately. Firstly, the data is characterized, the consumption behavior index is constructed, the date variable of the original data table is derived from the year, month and day variable, and then the different consumption behaviors of different consumers are aggregated by month, and the new variables of consumption behavior are derived, as shown in Table 1.

The statistics for the description of the new variables after the completion of the derivation are shown in Table 2.

By observing the situation of the derived variables, it can be found that the consumer operation in this data is basically click-based. Only less than half of consumers make a purchase. At the same time, consumers also make more frequent operations to add to the shopping cart than to add to the collection. However, the difference between the largest collection and the smallest collection is the largest except for clicks, which indicates that there is a consumer preference for collection operations.

TABLE 1. Feature engineering

Behavior	New Variables
Clicks	Monthly maximum clicks, monthly minimum clicks, monthly average clicks
Collection	Monthly largest collection, monthly minimum collection, monthly average collection
Add to cart	Monthly maximum added to cart, monthly minimum added to cart, average collection
Purchase	Monthly maximum purchase, monthly minimum purchase, monthly average purchase

TABLE 2. The new variable describes the statistic

Variable name	Maximum	Median	Average Value
Clicks	575	30	45.95
Maximum number of clicks	267	16	23.46
Minimum number of hits	106	2	4.15
Average number of hits	128	8.67	12.25
Number of favorites	145	0	1.48
Largest collection	128	0	0.93
Minimum favorite	17	0	0.07
Average collection	38.67	0	0.39
The number of times you added to your cart	87	1	2.79
Maximum add to cart	75	1	1.75
Minimum add to cart	16	0	0.12
Average number of times you add	17.4	0.33	0.76
Number of purchases	84	0	1.03
Maximum number of purchases	30	0	0.76
Minimum purchase	13	0	0.02
Average number of purchases	17	0	0.27

4. K-means clustering based on the consumption behavior of e-commerce users.

4.1. **Determine the optimal K value.** Based on the above two different determination methods, I chose the more intuitive elbow method to determine the optimal k-value of this set of data [17] .

The specific results are shown in Figure 1:

According to the principle of the elbow method, it can be found that the curvature of the curve is the highest when $k = 3$ or 4, and here, $k = 4$ is taken for clustering.

4.2. **Analysis of differences between categories.** As shown in Table 3, after selecting some of these variables and comparing different categories of customers, it is found that the average number of clicks is much higher than that of other customers. In the variables related to collections, the data of category 4 customers is better than that of other customers, where the smallest gap is the smallest collection. Although the gap between technical customers is not very obvious in terms of joining the shopping cart, there is still a certain lead for category 2 customers. Finally, in terms of purchases that affect

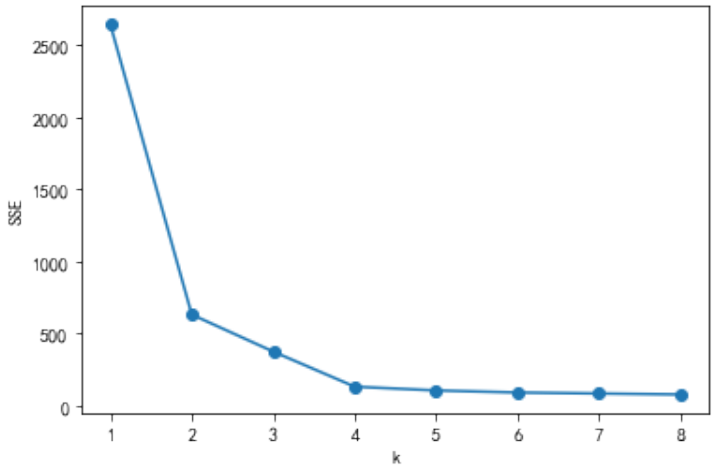

FIGURE 1. Elbow test

TABLE 3. Operational statistics for different categories of customers

Variable	1	2	3	4
Average number of hits	13.96	10.56	11.96	10.54
Average number of purchases	0.23	1.56	0.52	0.45
Add to cart on average	1.02	1.24	1.14	1.04
Average collection	1.07	1.21	1.22	1.34
Maximum purchase	1.11	1.32	1.27	1.20
Minimum purchase	0.16	0.27	0.21	0.17
Largest collection	2.31	2.32	2.35	2.88
Minimum favorite	0.40	0.44	0.52	0.53
Maximum add to cart	2.39	2.62	2.56	2.44
Minimum add to cart	0.44	0.54	0.52	0.48
Sample size ratio	33.23 %	16.92 %	39.99 %	9.85 %

corporate profits, the number of purchases by Category 2 customers is much higher than that of other 3 categories, followed by Category 3 customers, and the least purchased is Category 1 customers.

After selecting some variables to visualize the results, as shown in Figure 2, it can be found that the customer of the first type has much higher average clicks than the other customers. Category 4 customers are very inclined to collect and buy lower. Category 2 customers are arguably the most distinctive customers and not only have a much higher advantage in addition to their shopping cart but also in terms of purchasing. Finally, the third type of customer can be said not to be related to any aspect of the characteristics are not very obvious, belongs to the more decent customers, but this type of customers because there is a high purchase volume, so it is very easy to induce it to increase the company's revenue.

4.3. **Feature analysis of different categories.** According to the clustering results, customers are divided into four categories: 1, 2, 3 and 4, and their different characteristics are defined as "iron powder customers", "general customers", "development customers" and "zombie customers".

The specific reasons are as follows.

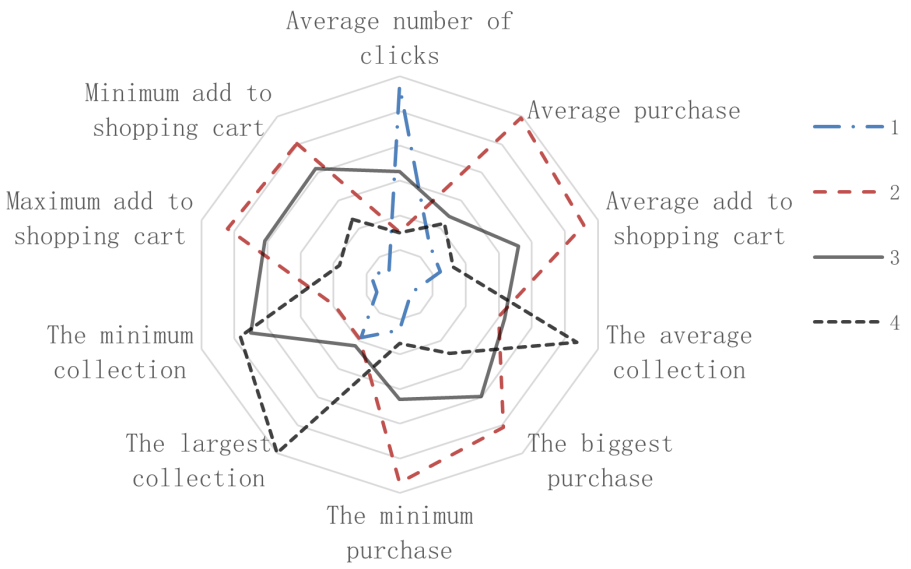

FIGURE 2. Radar chart of different types of customers

First class customers are doing clicks much more frequently than other types of customers but their other actions are very low, so I infer that these customers just clicked on the product detail page to view it, but this type of customer is not useless, as long as he clicks into the detail page of this product, it means that the customer is interested in this kind of product. As for why the follow-up operation is not carried out, then there will be many reasons, it may be because of dissatisfaction with the price, dissatisfaction with the benefits of the purchase, and of course, it may also be dissatisfaction with the quality of the product. It's often difficult for these types of customers to have much impact on a company's revenue, so we define them as "zombie customers".

Although Category 4 customers do not have as many clicks as Category 3 customers, their favorites are much higher than those of other types. It shows that this type of customer may collect several items that they feel good about, and then compare the characteristics of these products. However, in the end, such customers did not make a purchase much more than categories 2 and 3, indicating that a large part of them did not choose the product in the end. Nevertheless, this type of customer has added the brand's products to the alternative list and most likely only needs some benefits to attract such customers to carry out subsequent operations, so I call this kind of customer who only needs a little effort on the brand side to lead to customer purchases: "developing customers".

The third type of customer, though, does not have obvious characteristics, they are the second most purchasing type of customer among the four types of customers. This shows that the purchase process of this kind of customer is largely from click-to-click directly to purchase, and no other superfluous operations are carried out. On the one hand, they have a very high conversion rate for shopping, but on the other hand, because they do not have a lot of collection or add to the shopping cart type operations, most of them are probably disposable customers. Although this kind of customer creates a lot of profits

for the enterprise, the customer who needs the enterprise to enhance its impression of the brand through some marketing means is called the "general customer".

The last category 2 customers did not have many clicks, but they did much more operations to add cart types and purchase types than other customers. This type of customer is often the most important thing for the company because this type of person often does not need the company to invest much publicity, they are either impulsive customers attracted by the benefits or old customers with a high degree of trust in the brand, such customers are generally the focus of the company's profit source. Therefore, we call this kind of customer who will create most of the profits for the enterprise "iron powder customer".

5. **Conclusion.** In this paper, k-means clustering model is used to analyze the recommended data of consumption behavior of e-commerce users. First of all, we proved through research that no matter what consumers are, they are more or less connected internally, so it is feasible to increase corporate profits by subdividing customers. Secondly, we further excavated the data of this experiment and found that both the clustering and classification models divided the data into four categories with different classification standards. Finally, according to the "80/20 rule", 80 % of the profits of enterprises are created by 20 % of the consumers. So it's very important for companies to maintain that 20 % of customers. Therefore, we named these four types as important value customers, general customers, development customers and potential customers in descending order of importance to the enterprise, and analyzed the characteristics of these four types of customers, explained the characteristics of different customer groups, and finally put forward corresponding suggestions according to different customer groups.

In recent years, e-commerce platforms have continuously impacted the offline shopping market with their wide range of goods sold and unrestricted purchase time. And with the acceleration of express logistics, more and more people choose online shopping. However, in the context of the spread of the COVID-19, e-commerce platforms have also been hit hard by the security of their packages, especially for some overseas purchasing goods. At the same time, with the implementation of some national policies and the gradual saturation of the market scale, the development speed of e-commerce platforms has slowed down significantly, and even some small e-commerce platforms have a relationship. In this case, if the e-commerce platform wants to continue to develop healthily, it needs to reduce internal friction to expand revenue. The most direct and simplest is to reduce advertising and other marketing expenses, and accurately launch advertising, so that people who do not know the product can buy it through advertising, and those who know the product can buy it through the search interface. In this way, we can find a dynamic balance between the two to achieve the purpose of reducing enterprise expenditure and increasing enterprise profits. Through research, this paper obtains what kind of marketing methods can be adopted for different types of users to achieve precision marketing as much as possible. This provides a certain reference for e-commerce platforms in pushing advertisements and related activities.

REFERENCES

[1] L. Zhang, "E-commerce platform consumer behavior analysis and prediction model," M.S. dissertation, Chongqing University, 2018.

[2] L.-D. Han, "Cluster analysis based on consumer behavior data of e-commerce users," M.S. dissertation, Lanzhou University, 2017.

[3] J.-Z. Zhai, "Analysis of consumption behavior of Internet users based on big data," *Business Economics Research*, vol. 24, no. 12, pp. 46–49, 2020.

[4] W. Dai, "Research on accurate classification of social e-commerce based on transaction data," M.S. dissertation, Beijing University of Posts and Telecommunications, 2021.
[5] X.-X. Wu, "Research on user segmentation of Internet financial platform based on improved RFM model," M.S. dissertation, Beijing Jiaotong University, 2016.
[6] H.-L. Xing, L.-L. Zhai, S.-C. Zhang, " Research on user value identification and segmentation of big data service platform–Based on RFM revision model," *Information Theory and Practice*, vol. 42, no. 10, pp. 131–13, 2019.
[7] J.-T. Jiang, " Research on Customer Value Classification oriented to Consumer Behavior," M.S. dissertation, Donghua University, 2021.
[8] Y. Wang, "Research on user value mining method of shopping guide platform based on RFM and BP neural network," M.S. dissertation, Guangxi University, 2021.
[9] W. Dai, "Research on precise classification of social e-commerce based on transaction data," M.S. dissertation, Beijing University of Posts and Telecommunications, 2021.
[10] J.-L. Chen, L.-L. WU, "Research on Customer Relationship of E-commerce Based on RFME Model and AdaBoost Classifier," *Software*, vol. 42, no. 3, pp. 1–7, 2021.
[11] L. Li, "Research on customer relationship Management of E-commerce enterprises in big data environment," *Chinese Market*, vol. 1, no. 10, pp. 195–196, 2021.
[12] E.-K. Wang, X. Zhang, F. Wang, T.-Y. Wu, C.-M. Chen, "Multilayer dense attention model for image caption," *IEEE Access*, vol. 7, pp. 66358-66268, 2019.
[13] Q.-Y. Yang, S.-C. Chu, J.-S. Pan, C.-M. Chen, "Sine Cosine Algorithm with Multigroup and Multistrategy for Solving CVRP," *Mathematical Problems in Engineering*, vol. 2020, 8184254, 2020.
[14] Q.-H. Huang, J. Li, C. Yang, "Internal threat detection method based on elbow clustering algorithm," *Jouranal of Hubei University for Nationalities*, vol. 37, no. 3, pp. 331–335, 2019.
[15] Y.-K. Jiang, "Research on the application of data mining technology in film recommendation," *Computer Knowledge and Technology*, vol. 18, no. 15, pp. 254–256, 2019.
[16] S.-J. Shen, "Influencing factors and prediction model of alloy yield," *Electronic Quality* , vol. 12, no. 1, pp. 1–7, 2019.
[17] J.-Z. Zhai, "Analysis of consumption behavior of Internet users based on big data," *Business Economic Research*, vol. 24, no. 14, pp. 46–49, 2020.