Research on User Classification of e-commerce Live Broadcast Platform Based on Bullet Screen——Take Beauty Live Streaming on Taobao as an Example

Wen Lei*

School of Economics and Management, Northeast Petroleum University, Daqing, Heilongjiang, 163318, China

*lw366499@163.com

Abstract. In the live broadcast of e-commerce, users and anchors interact in real time through the bullet screen, so a large amount of user information and data are stored in the live broadcast bullet screen. Effectively utilizing this information and accurately identifying effective users can help e-commerce platforms and broadcast rooms better understand the needs of users. In this paper, an improved RFM based user classification method for live streaming of e-commerce is proposed. By constructing an improved RFM model, the index and attribute of user classification are added. SOM neural network algorithm combined with K-means clustering algorithm is used to classify barrage users. The experimental results show that the SOM model combined with K-means method has a good effect in user classification. Through the classification of Taobao live streaming platform beauty pop-screen users, four user groups are obtained: deep loyal users, active active users, active cultivation users and potential users. Finally, the corresponding marketing strategy is given according to the characteristics of the four groups.

Keywords: e-commerce live broadcast; bullet screen users; user classification; RFM model; cluster analysis.

1 Introduction

Electricity live this emerging way of shopping widely welcomed by consumers, for electricity enterprises, this also means a huge opportunity and more fierce competition, therefore, a deeper understanding of the user and targeted marketing is electricity enterprise in the fierce market competition is an important way. Categorizing users is the first step to understanding them.

The RFM model is a classic model for distinguishing users, which can distinguish different users from a large amount of transaction data, and is often used to classify users. In recent years, many scholars have studied the user classification and improved RFM models from different perspectives. Wei Ling et al. (2020) established an RFLP index system for MOC user learning behavior and attrition prediction by

A. K. Draman Mud et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Social Sciences (ICBDSS 2024)*, Advances in Computer Science Research 116,
https://doi.org/10.2991/978-94-6463-562-1_3

improving the RFM model, and built a MOC user loss prediction model combined with the data grouping processing network[2]. Yan Chun et al. (2020) improved the RFM model from both macro and micro perspectives, built the RFMC model, and finally divided non-life insurance users into seven categories, and put forward marketing strategies for different categories of users respectively[3]. Le Chengyi et al. (2020) took the reading information and behavior data of a university library as an example, and proposed the construction method of user portrait of university library based on improved RFM clustering[4]. You Qijun et al. (2022) took users of a typical creative crowdsourcing community as an example for empirical research, selected user indicators to improve the RFM model, and divided users into star users, genius users, loyal users, ordinary users and sunken users through K-means clustering[5].

Bullet(Danmaku) is the important embodiment of electricity live real-time, specifically refers to in the process of live, when the audience gathered in the same scene to watch video or live, will send some text or symbols to express their thoughts and emotions, these characters in the form of rolling in the live interface, when a lot of barrage, as dense bullets flew, so the comments are named "barrage"[6]. Since the function of bullet screen greatly improves the real-time interactivity of e-commerce live broadcast, many scholars have studied how bullet screen affects users' purchase. Han Yutong et al. (2022) paid attention to the impact of bullet screen on users' consumption behavior, and confirmed that bullet screen has a positive impact on the sales of goods in e-commerce live broadcast[7];Li Yiping and others (2023) proved through empirical research that real-time comments play a complete intermediary role in the relationship between live streaming e-commerce and consumers' purchase intention[8]; Liu Chenglin et al. (2023) constructed a path model of the influence of barrage information quality on consumers 'purchase intention based on the detailed possibility model, and found that the quality of barrage information can then influence consumers' purchase intention through their perception of kilometer and perceived attribution[9]; Jiao Yuanyuan et al. (2024) constructed a research model of "the characteristics of instant comment information to reflect customer stickiness" based on the theory of emotional cognition, and found that the characteristics of instant comment information affected the stickiness of customer access and purchase stickiness to varying degrees[10]. Accordingly, it can be found that bullet screen has a great impact on users and their consumption behavior, and the research on bullet screen users is essential.

Therefore, this paper tries to construct the user classification model of e-commerce live broadcast bullet screen, and classify the bullet screen users through empirical analysis, so as to deeply explore the characteristics of e-commerce live broadcast bullet screen users, and put forward targeted management suggestions for e-commerce live broadcast platforms and live broadcast rooms.

2 Model Building

2.1 Traditional RFM Model

The traditional RFM model is based on three indicators: The interval (R) of the latest purchase shows the time distance of the user's latest consumption, reflecting the user activity and demand for the platform. Purchase frequency (F) indicates the number of purchases in a specific period, which is related to the user's desire to buy and loyalty to the enterprise. The consumption amount (M) is the total consumption amount of users in the statistical cycle, reflecting the consumption ability and contribution degree of customers.

Although RFM has been widely used in the field of user segmentation, the RFM model still has some limitations in the research of bullet screen user classification on e-commerce live streaming platforms. The first is the difference in user behavior. The traditional RFM model segmentation is mainly based on the user's purchase behavior, while in the barrage user segmentation, the barrage behavior of the user is more discussed. Second RFM model only consider the user's direct economic value, for live platform, user value is not limited to its consumption behavior, the user's attention, user interest, and the user itself is more important, more basic wealth, the traditional RFM model ignored the represented by traffic user potential value, has great limitations.

2.2 Improved RFM Model

Since the traditional RFM model can only classify users according to the recent transaction data, and the ability to distinguish the user behavior of the platform is weak, many scholars have modified the traditional RFM model and expanded the application field of RFM. Among them, Chen Danhong et al(2022) Based on the RFM model and the crowdsourcing platform user features, will User credit is incorporated into the user value model to build the crowdsourcing user value Measure the model RFMC[11]; Xi Yunjiang et al. (2023) will average users based on the importance of user engagement and long-term value in the live broadcasting platform.Watch time is integrated into the existing RFM model as an important dimension, and the RFMT model is constructed to realize user classification[12]. At the same time, many scholars have carried out a lot of research on online users. For example, Chen Yang et al. (2023) borrowed the existing user activity research method and proposed the user activity measurement algorithm of the scientific research information service platform[13]; Based on the dimensions of contribution quality and contribution quantity, Chen Jian et al. (2023) subdivided users into four categories: users in core concern area, users in priority improvement area, users in auxiliary concern area and users in advantage maintenance area[15]. Therefore, this paper draws on the research results[14] of previous scholars, considering the limitations of traditional RFM and the characteristics of barrage users, and combines with the literature[1]Build an improved RFM model for barrage user classification. The model mainly includes three indicators: user activity, user contribution and user engagement. Among them,

the activity index mainly includes average barrage time interval and barrage end time; contribution index mainly includes user emotion degree and user identity number; engagement index mainly includes average barrage length and sending frequency. As shown in Table 1.

Table 1. The improved RFM model

target system	Subdivision attribute
User Activity (A)	A_1: Average barrage time interval
	A_2: Bullet screen end time
User Contribution Degree (C)	C_1: User emotion
	C_2: Number of user identifiers
User engagement (P)	P_1: Average barrage length
	P_2: Frequency of sending a barrage screen

3 Research Technique

3.1 The Emotion Value Calculation Based on the Emotion Dictionary

Emotion analysis is an analysis method used to judge the polarity and tendency of text emotion. Its processing steps usually include data acquisition and number. According to pre-processing, extracting emotional information, and calculate emotional values. this research. Using the method based on the emotional dictionary, referring to Liu Siyuan[16]To build a special beauty makeup bullet screen emotional dictionary. By analyzing the vocabulary, phrase and sentence structure of the text, we can calculate the emotion value of the barrage, judge the emotion of the barrage, and construct the user engagement index.

3.2 SOM-K-Means Algorithm

This paper uses the reference literature The SOM-K-means algorithm is selected, which combines the self-organization of SOM neural network and the efficient characteristics of K-means algorithm, and effectively compensates for the long convergence time of SOM neural network and the improper random selection of K-means algorithm, which is easy to cause poor clustering results. The clustering of the algorithm is divided into two stages: in the first stage, the data is input into SOM neural network model, and the SOM model is used to find the clustering center and its initial range; in the second stage, the output results of SOM neural network are input into K-means algorithm, and the second stage clustering is conducted to obtain the final clustering results.

4 Empirical Analysis

4.1 Data Source

Considering that the bullet screen of Taobao Live Studio contains a large number of user needs and tendencies, this paper, according to the historical situation of the total list of Taobao Live Platform, selects the first Estee Lauder Live Studio to capture the bullet screen. The live time of capture is 17:00-24:00 every day from October 12,2023 to October 22,2023. Select Air test IDE software for data collection, and finally get 52,084 barrage data sent by 27,654 users, which was saved as local files in CSV format to facilitate subsequent data processing.

4.2 Data Processing

First, the barrage data is cleaned. The data cleaning of this paper mainly includes two parts: cleaning the irrelevant symbols in the data and cleaning the advertising and merchant guide information. Irrelated symbols mainly refer to the symbols without substantive content, such as dashes, ellipses, and brackets, etc. Such symbols do not carry any specific information, but will cause interference to the effective identification and analysis of the bullet screen content, so this part of the content is cleared. There are also some contents in the bullet screen, such as "Photo 3,20 tasting box; 200 available; Photo 4,100 tasting box available at 900". This kind of information belongs to the publicity and marketing information of merchants and is not related to the actual needs and wishes of consumers, so this kind of information is cleared.

Secondly, the data were statistically analyzed according to the modified RFM model. The frequency of sending barrage refers to the number of barrage sent by the user during the data capture period, and the number of user identity refers to the number owned by the user that cannot dial the member identity. The data of the above attributes is calculated using the function of EXCEL software.

Average barrage length refers to the average of the length of all the barrage sent by a user. Let the length of each barrage sent by the user be L1, L2, L3... Ln, then the average barrage length Laverage is calculated by formula (1):

$$L_{average} = \frac{1}{n}\sum_{i=1}^{n} L_i \tag{1}$$

The end time of the bullet screen is the time difference between the time when the user sends the barrage in the studio and the deadline time of data capture on the same day. Let the time when the user last sends the bullet screen on the day be Tuser, and the deadline time of data capture on the day is Tend, then the end time of the bullet screen T can be calculated by formula (2), the unit is hours.

$$T = T_{end} - T_{user} \tag{2}$$

Individual bullet screen time interval refers to the time interval between two adjacent barrage screens sent by a user in the live broadcast of the same day, and the average barrage time interval refers to the average value of all individual barrage time interval of the user. Let the time sequence of a user be (t1, t2, t3... Ttn), then the individual barrage time interval $T_{individuality}$ And the average barrage time interval of $T_{average}$It can be calculated by formula (3) and formula (4) respectively:

$$T_{\text{individuality}_i} = t_{i+1} - t_i, 1 \le i \le n-1 \tag{3}$$

$$T_{average} = \frac{1}{n-1}\sum_{i=1}^{n-1} T_{\text{individuality}i} \tag{4}$$

User emotion refers to the average emotional value of all the bullets sent by the user on that day. Let each bullet screen emotion sent by a user be E1, E2, E3... En, then the user emotion degree Eaverage can be calculated by formula. (5):

$$E_{average} = \frac{1}{n}\sum_{i=1}^{n} E_i \tag{5}$$

The processed data is summarized to obtain the user vector, which is used for the subsequent cluster analysis. Some user vectors are shown in Table 2.

Table 2. Example of the user vector

user	Contribution de-		degree of partic-		vitality	
	C1	C2	P1	P2	A1	A2
User 1	4.5643	0	1	12	0	1.8490
User 2	0.8288	1	1	2	0	4.1385
User 3	1.0216	1	1	7	0	0.9511
User 4	3.7393	1	6	2.5	0.0231	4.0082
User 5	1.5534	3	2	9	0.0598	4.1993

4.3 Bullet Screen User Classification Based on the Improved RFM

1). Empowerment of indicators

Since there are two subdivision attributes in each index in the model constructed in this study, considering the common influence on its respective indicators, the method of each attribute is used to calculate the weight of each attribute, as shown in Equation (6).

$$\mathrm{C} = \omega_1 C_1 + \omega_2 C_2 \tag{6}$$

Among them, C refers to the user contribution index in the improved RFM model, C1 and C2 represent the subdivision attributes of the user contribution C respectively, and represent the weights of C1 and C2. If user engagement and user activity are similar in expression and contribution, they will not be repeated. ω_1, ω_2.

In order to eliminate the subjective factors, the entropy method was selected to empower each subdivision attribute of the improved RFM model, and the results are shown in the Table 3.

Table 3. The improved subdivided index weights of the RFM model

metric	ω_1, ω_2	ω_1, ω_2
User Contribution Degree (C)	0.8281	0.1719
User engagement (P)	0.8074	0.1926
User Activity (A)	0.5855	0.4145

Take the calculated weight into formula (6), and the contribution index is shown in formula (7).

$$C = 0.8281C_1 + 0.1719C_2 \tag{7}$$

Similarly, the activity and participation indicators are shown in Equation (8) -(9).

$$P = 0.8074P_1 + 0.1926P_2 \tag{8}$$

$$A = 0.5855A_1 + 0.4145A_2 \tag{9}$$

2). Data normalization

Because there are great differences in the values of the three key indicators. In order to eliminate the differences in measures and dimensions, further normalization of the three indicators is needed. In this paper, the data were normalized by the maximum and minimum normalization method (Max-Min Normalization). This normalization method works by scaling the data so that it falls within a specified small range, specified to [0,1].

3). The K-value is determined by the contour coefficient

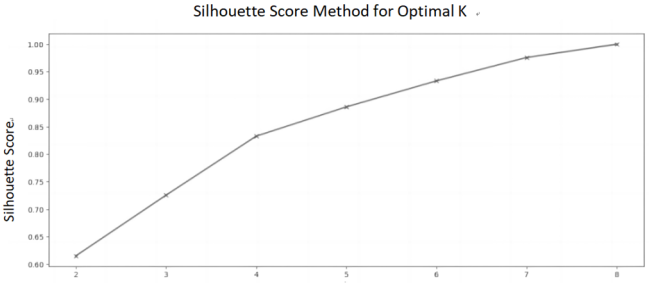

Fig. 1. contour coefficient calculation results

The optimal number of clusters was determined by calculating the contour coefficient, as shown in Figure 1, where K was 8, the contour coefficient was maximum, so the number of clusters was set to 8.

4). Bullet screen user clustering results

Clustering using the SOM-K-means algorithm yielded eight classes of users, and the number of each class and its clustering center are shown in Table 4. [14,15]

Table 4. Cluster results and each cluster center

class name	Number of users	Cluster center		
		C	A	P
Class 0	7829	0.2066802	0.0306603	0.067093
Class 1	2941	0.0223587	0.0285274	0.0817865
Class 2	1470	0.9569324	0.0214368	0.091004
Class 3	3771	0.194743	0.1852433	0.0900575
Class 4	1856	0.0908925	0.1625822	0.1084365
Class 5	2822	0.3770969	0.1082488	0.0939125
Class 6	1509	0.8421032	0.0160523	0.090926
Class 7	2433	0.1279814	0.049269	0.074673

Considering the classification balance and in the actual marketing process, assigning different marketing strategies for 8 types of users will lead to too much cost, so the method of index segmentation is adopted to further classify the clustering results. The specific classification process is as follows: by calculating the average index mean of the sample population, when the central value of the group index is higher than the overall mean of the sample, the index is evaluated as, and when the central value of the group index is lower than the mean, the index is evaluated as (where A2 has been converged). Eventually more than three indicators above the mean of users are divided into the first type of users, the user contribution and user participation is higher than the average user is divided into the second type of user, the user engagement and user activity is higher than the average users divided into the third type of users, for all indicators below the average users, classified as the fourth type of users. The final user classification results are shown in Table 5.

Table 5. Each user categories and their user clusters

Bullet screen user category	User Cluster number	subscriber number	C	P	A
Deep loyalty to users	5	2822	↑	↑	↑
Activate users actively	2,6	2979	↑	↓	↑
Active cultivation of users	3,4	5627	↓	↑	↑
potential user	0,1,7	13203	↓	↓	↓

5 Conclusion

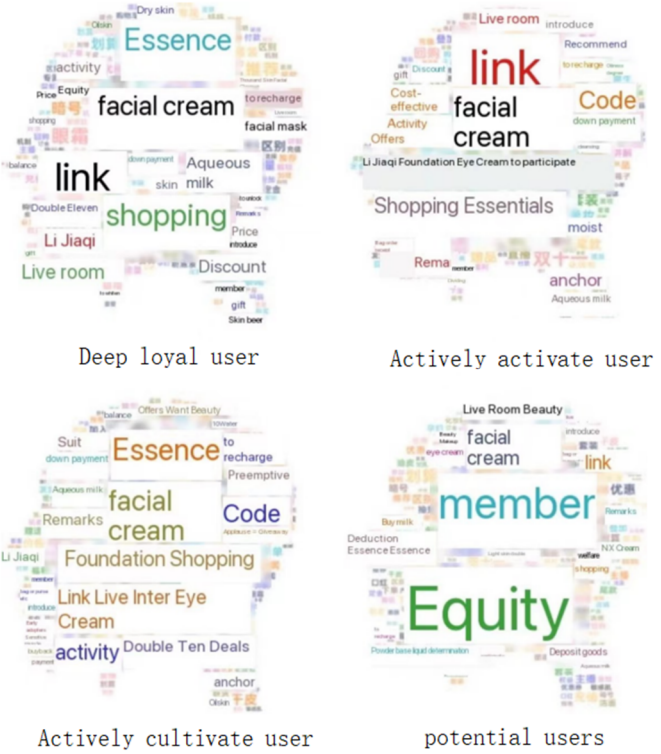

Fig. 2. Cloud map of user group words

5.1 Deep and Loyal to Users

As shown in Figure 2, Deep loyal users are the mainstay of live broadcasting platform, they not only perform well in all aspects, with high contribution, active participation and strong activity, but also loyal supporters of live broadcasting room. Their activity in the live broadcast process injects vitality and vitality into the entire live broadcast room, while showing high loyalty to the live broadcast platform or brand. Such users often return to the platform, watch live programs for a long time, and actively participate, not only as consumers, but also as the backbone of the live community. They are well aware of live products, highly sensitive to new product launches and event information, and show strong purchase intention and loyalty. Therefore, by providing personalized live recommendations and customized services, setting up targeted exclusive reward programs, creating exclusive community areas, etc., to continuously enhance their loyalty and satisfaction. In addition to increasing satisfaction and loyalty, they should also leverage their influence to attract a broader user

base, creating a virtuous cycle that drives the platform's continued growth and success.

5.2 Actively Activate Users

Active cultivation of users' emotional connection to live content and brands is weak, but high engagement and activity show their potential loyalty. They need the right incentives and guidance to become deeply loyal users. In e-commerce live streaming, it is crucial to cultivate users' shopping habits. It is recommended to set daily live time, such as 5pm to 12pm every night, welfare promotion on Saturday, interactive or new product launch on Sunday. This regular live streaming helps build user habits, increases user engagement, and stimulates purchase intentions through limited-time events, increasing sales and conversions. Through these strategies, the studio can create a positive, healthy and interactive live environment.

5.3 Actively Cultivate Users

Active cultivate users have weak emotional connection to live content and brands, but high participation and activity show their potential loyalty. They need the right incentives and guidance to turn them into deeply loyal users. In e-commerce live broadcasting, it is very important to cultivate users' shopping habits. It is recommended to set daily broadcast time, such as 5 to 12 every night, welfare promotion on Saturday, interaction or new product release on Sunday. This regular live streaming helps to establish user habits, improve user engagement, stimulate purchase intention through time-limited activities, and improve sales and conversion rates. In addition, when there are many users in the broadcast room, it is difficult for anchors to answer questions one by one. Therefore, it is very necessary to set up a special person to manage the barrage screen. Anchors should formulate clear bullet screen rules, arrange special personnel to monitor in real time, and timely respond to and guide improper content. Anchors should also actively participate in the interaction of the barrage screen, reply to users' questions, guide the topic direction, and avoid the emergence of negative topics. Through these strategies, the live broadcast room can create a positive, healthy and highly interactive live broadcast environment.

5.4 Potential User

Although potential users are average in terms of activity, their attention to live streaming indicates its potential value. Such users may not know enough about live e-commerce platforms and need more information and guidance. The user word cloud shows that they pay attention to "members", "benefits" and product information. Livestreaming platforms can use eye movement recognition technology to optimize personalized recommendations and match users' interests. In order to stimulate the participation of such users, it is suggested to increase the diversity of content, such as new product display and use skills; Enhance interactivity, such as increasing the frequency of questions and answers, and issuing barrage coupons; Use the celebrity effect

to invite experts and stars to participate; Optimize the live broadcast screen to ensure user retention and enhance the purchase intention and engagement of potential users.

6 Research Limitations and Prospects

Classifying e-commerce live broadcast users helps e-commerce platforms and live broadcast merchants better understand users, and make personalized recommendations for different user groups to achieve precision marketing. In this paper, the traditional RFM model is improved, considering the characteristics and attributes of the bullet screen users, and the improved RFM model is established. SOM neural network and K-Means clustering are used to classify users and name user groups. Finally, corresponding marketing strategies are given according to the characteristics of various users, providing early support for the precision marketing of live streaming beauty products. However, this paper only studies the user data of bullet screen in the live broadcast room of beauty brand, which has limitations. In the future, it will be considered to combine the data of multiple types of live broadcast room to improve the accuracy of the research.

References

1. Ling Wei, Xinyue Guo. MOOC user loss prediction based on improved RFM and GMDH algorithms[J]. Distance education in China, 2020(09):39-43.
2. Chun Yan, Lu Liu. Research on non-life insurance customer segmentation based on improved SOM neural network model and RFM model[J]. Data analysis and knowledge discovery, 2020,4(04):83-90.
3. Chengyi Le, Xi Wang. Research on user portrait of university library based on improved RFM clustering[J]. Library theory and practice, 2020(02):75-79.
4. Tsao Y, Raj P V R P, Yu V. Product substitution in different weights and brands considering customer segmentation and panic buying behavior[J]. Industrial Marketing Management, 2019,77:209-220.
5. Li J. The Interface Affect of a Contact Zone: Danmaku on Video-Streaming Platforms[J]. Asiascape Digital Asia, 2017,4(3):233-256.
6. Yutong Han, Jilei Zhou, Fei Ren. The impact of real-time comments on the sales of live e-commerce products from a dynamic perspective[J]. Management science, 2022, 35 (01):17-28.
7. Yiping Li, Yanqiang Lu, Yimiao Wang. Analysis of the relationship between real-time comment content and consumers' purchase intention in the case of live e-commerce[J]. Commercial economic research, 2023(23):75-78.
8. Yuanyuan Jiao, Xue Gao, Jun Du. A study on the influence of instant comment information characteristics on customer stickiness in e-commerce live streaming: based on emotional cognition theory[J]. Management review, 2024,36(03):119-131.
9. Danhong Chen, Zhanglin Peng, Dequan Wan. User value identification and segmentation in crowdsourcing platforms: Based on an improved RFM model[J]. Computer science, 2022,49(04):37-42.

10. Yunjiang Xi, Dailing Guo, Xiao Liao. Research on user segmentation and personalized recommendation method of live broadcast platform based on improved RFM model[J]. Competitive intelligence, 2022,18(03):36-47.
11. Yang Chen, Haiyan Bai. Research on user activity measurement of scientific research information service platform -- A case study of National Science and Technology Books and Documentation Center[J]. Digital library Forum, 2023,19(03):28-35.
12. Jian Chen, Yue Ding. Type segmentation and differentiated incentive of archive crowdsourcing users in the dimension of participation contribution[J]. Archives science bulletin, 2023(06):86-94.
13. Siyuan Liu. Research on emotion classification of online review text in beauty field based on emotion dictionary[D]. Central China Normal University, 2022.
14. Limin Hou, Wenli Wang. Improved K-Means clustering algorithm based on SOM[J]. Journal of Inner Mongolia University (Natural Science Edition), 2011,42(05):586-590.
15. Faming Zhang. A dynamic credit evaluation method combining SOM and K-means algorithm and its application[J]. Operation research and management, 2014,23(06):186-192.
16. Liu Siyuan. Research on Emotion classification of online review texts in the field of beauty and makeup based on Emotion dictionary [D]. Central China normal university, 2022. DOI: 10.27159