E-commerce Platform User Behavior Data Analysis

Yaxuan Wen, Haoyu Wang, Zhanyi Pan

1. Introduction:

In recent years, the customer acquisition cost in the internet e-commerce industry has been gradually increasing. Maximizing the benefits for online shopping platforms and merchants while providing convenience to customers has become the primary focus of development in the e-commerce sector. Against this backdrop, this analysis utilizes a dataset of Taobao users to analyze their shopping behaviors and *identify user behavior trends*, *operational issues within the online shopping platform*, and *the underlying causes of these issues*. Moreover, it aims to provide recommendations for optimizing and improving the e-commerce platform at a systemic level.

2. Methodology:

The subject of this analysis is Taobao, an e-commerce platform owned by Alibaba and the largest e-commerce platform in China. To ensure data reliability, we have chosen to obtain user behavior data through the official channel, Alibaba Cloud Tianchi website. We have selected one million rows of data from this source for the purpose of analysis.

Through this experiment, we conducted an in-depth analysis on the following aspects:

- **User attrition rate on the Taobao platform:** We examined the extent to which users disengage or cease their activities on Taobao.
- **Underlying causes contributing to user attrition:** We explored the factors that lead to this type of user churn, aiming to identify the root causes.
- **Strategies to mitigate and prevent similar attrition:** Based on our findings, we provide recommendations and actionable strategies to effectively minimize user attrition and foster long-term engagement on the platform.

To conduct the analysis, we employed Spark and MongoDB as analytical tools. Given the substantial volume of data for this experiment, we opted to leverage the high-speed processing capabilities of Apache Spark for handling large-scale data tasks, along with the interactive and ad hoc querying capabilities of MongoDB. We utilized Spark for data transformation and aggregation tasks, subsequently storing the processed data in MongoDB for further querying and analysis purposes.

For this experiment, we utilized funnel analysis to analyze user behavior and optimize conversion rates within the defined process. Furthermore, we employed hypothesis testing analysis to test assumptions and draw inferences about the population based on the sample data.

3. Data Processing

3.1 Field translation

user_id: user ID

item_id: item number

behavior_type: User behavior type (including four behaviors of click, wishlist, add to shopping cart, and purchase, which are represented by numbers 1, 2, 3, and 4, respectively, in the original field)

user_geohash: geographic location

item_category: product category number

time: the time when the user behavior occurred

3.2 Data cleaning

3.2.1 Select field

Most of the geographical location data in the item_category column are NULL , and the location information is encrypted, making it difficult to study. So do not select this column for analysis, that is, delete it directly.

3.2.2 Delete duplicate values

No duplicates were found in the data

3.2.3 Missing value processing

No missing values were found in the analysis fields in the data

3.2.4 Consistent processing

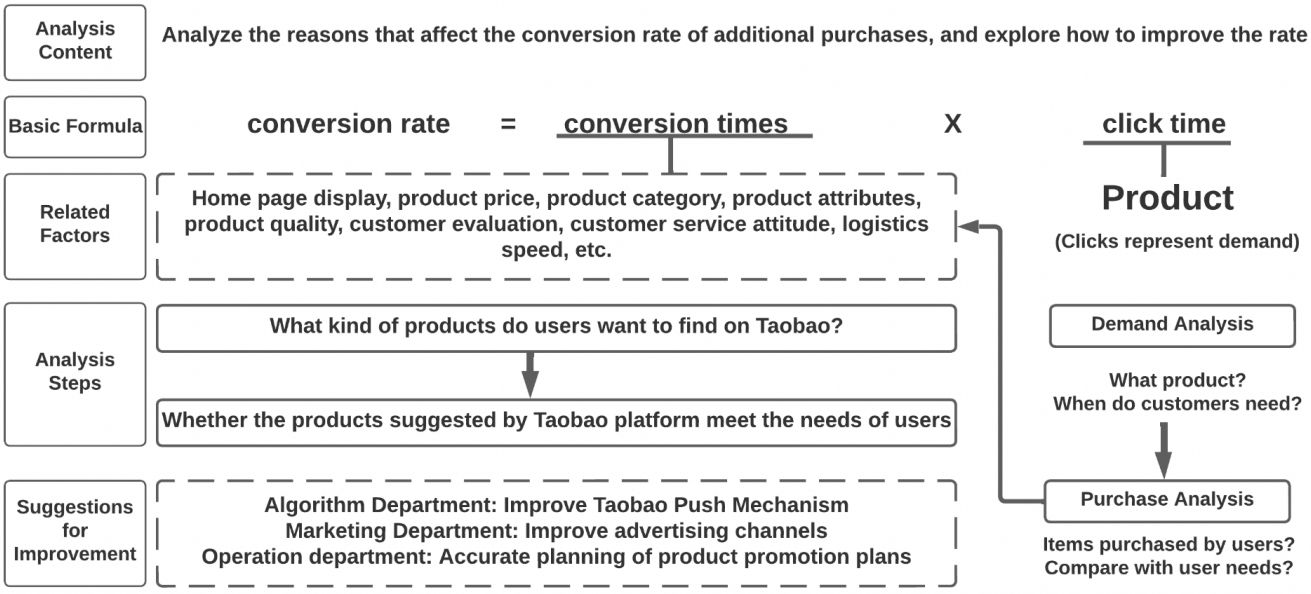
Since the time of the time field includes (year-month-day) and hour, for the convenience of analysis, this field is divided into 2 fields, a date column and an hour column.

3.2.5 Outlier processing

Check whether there are abnormal values in each field and whether it meets the specification. After checking, all data is normal, the data conforms to the specification, and there is no need to delete the data.

4. Findings and Analysis:

4.1 Overview of Analysis Ideas



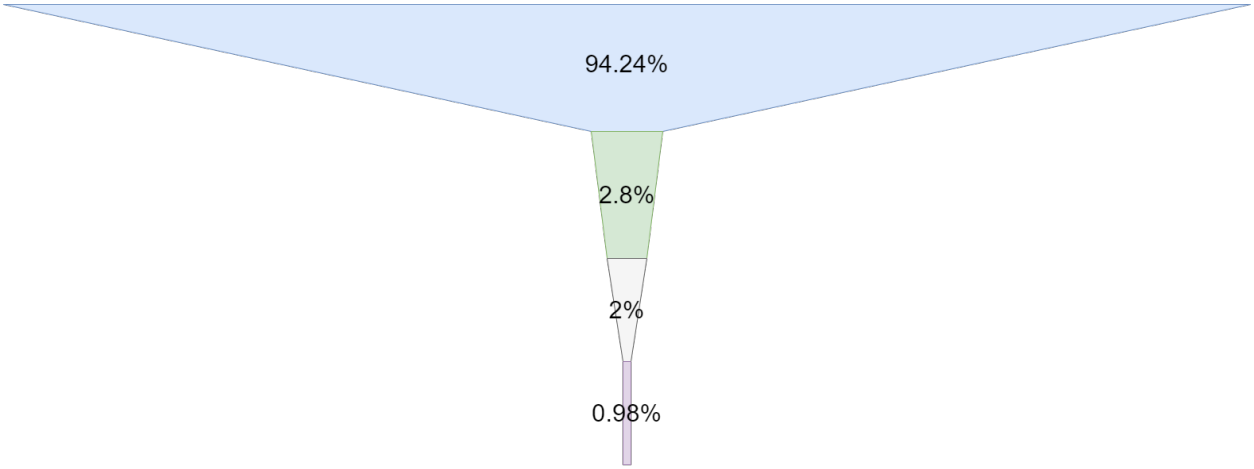
On the user level, we utilized the funnel model to analyze the conversion rates between different stages of customer engagement, including clicks, adding to cart, wishlisting, and purchasing. We examined the distribution of conversion rates across different time periods within a day. Based on these findings, we strategically increased advertising efforts during high-

conversion time periods, aiming to maximize the conversion rate and enhance both traffic utilization and store sales.

On the product level, we conducted hypothesis testing to analyze the distribution of traffic-generating products versus top-selling products. This analysis aimed to optimize new product introductions, allocate investments in paid promotions, manage product inventory levels, and determine restocking cycles.

4.2 Process of Analysis

4.2.1 User Attrition



User behaviors encompass clicking, adding items to the shopping cart, wishlisting, and making purchases. According to Figure 1, clicks account for a substantial 94.24% of all user actions, whereas adding items to the cart represents a mere 2.8%. Finally, actual purchases comprise less than 1% (0.98%) of the total. Notably, user attrition primarily transpires at the stage of adding items to the cart.

Consequently, we posit a hypothesis: it is plausible that users spend a significant amount of time on Taobao but struggle to find the desired products, prompting them to abandon Taobao as a purchase platform and seek alternatives elsewhere.

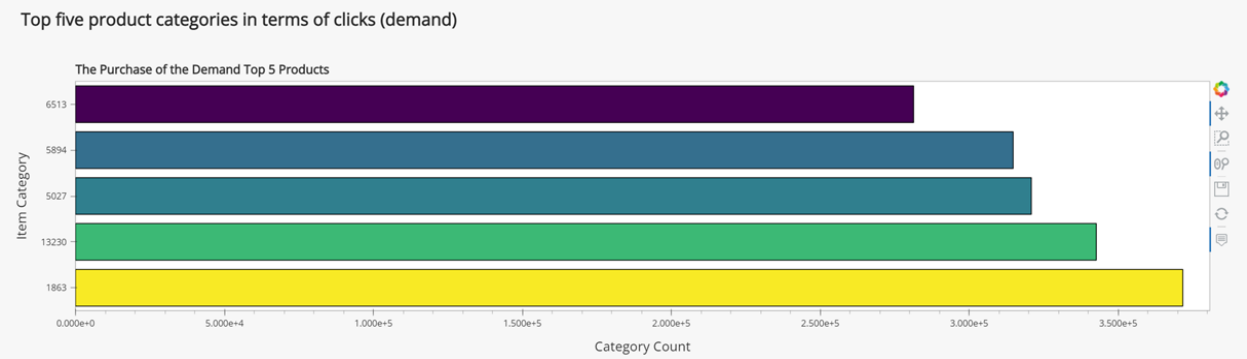
To validate this conjecture, we will analyze the hypothesis through the following three dimensions:

1. *What types of products are users seeking on Taobao?*
2. *When (during which time periods) do users typically make purchases?*
3. *To what extent do the platform's product recommendations align with users' needs and preferences?*

4.2.2 Analysis of the causes behind user attrition

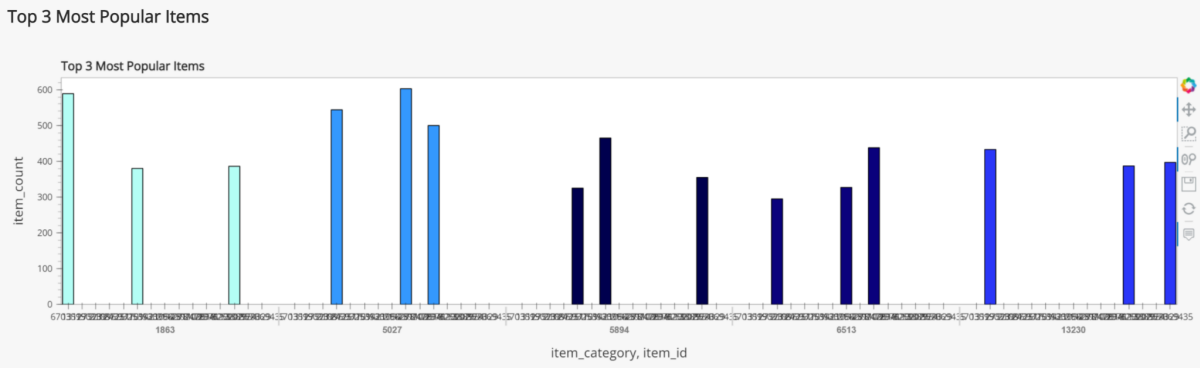
4.2.2.1 What types of products are users seeking on Taobao?

The most crucial indicator for gauging the products that users most desire to find on the Taobao platform is the number of product clicks. By utilizing this metric, we can gain valuable insights into the product categories that exhibit high demand among users, as well as those with comparatively lower demand. (Note: Click count reflects user demand.)



From the graph, it can be observed that the product categories with the identifiers 1863, 13230, 5027, 5894, and 6513 have the highest number of clicks on the Taobao app. This indicates that users are most interested in finding products from these five categories.

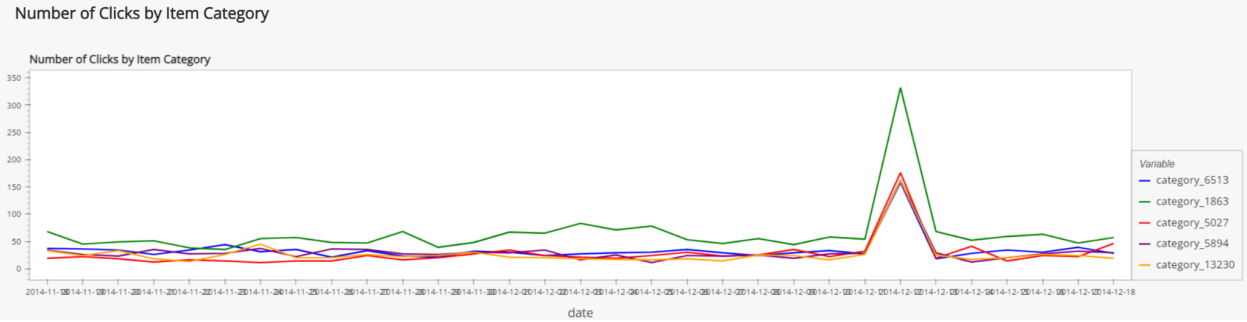
Within each of these five categories, the top three products with the highest number of clicks are as follows:



According to Figure 3, it can be observed that among the various product categories, there is a high demand for products in category 5027, with all the top three products in this category having a click count of over 500. On the other hand, category 6513 exhibits relatively lower demand for its products.

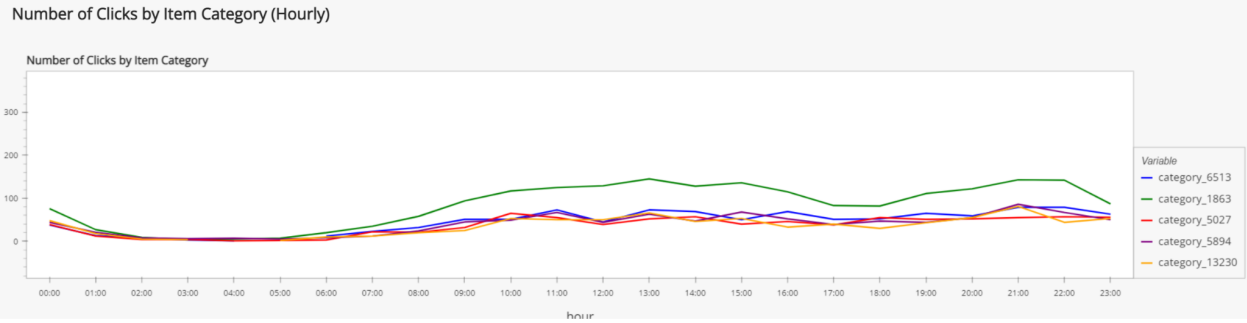
4.2.2.2 When (during which time periods) do users typically make purchases?

Based on the above results, we analyzed the primary time periods during which users search for these five product categories, both within a month and within a day. The findings are as follows:



According to Figure 4, it can be observed that the click counts for these five product categories experienced a significant surge on the eve of and during the Double 12 (December 12th) event on Taobao (Note: Double12 is a promotion launched by Taobao). After Double 12, the click counts returned to normal levels, with no significant fluctuations observed during other

time periods.



In terms of the primary time periods within a day when users search for these five product categories, it can be seen that the click counts for these categories rapidly decline from 10 PM to 5 AM, reaching the lowest point around 4 AM. From 6 AM to 11 AM, the click counts increase rapidly, followed by a relatively stable period from 11 AM to 4 PM. The click counts then rise rapidly again from 6 PM to 10 PM, reaching the highest point around 10 PM.

Looking specifically at the click distribution within a day for the top three products in the highest-demand category, 1863, users predominantly search for these products on the Taobao platform from 6 PM to 11 PM. Although there may be slight variations in search patterns for specific products within each category, the overall trend includes a focus on the time period 12 PM to 3 PM and time period 9 PM to 10 PM.

It is evident that users are particularly active in the time slot above. For businesses engaged in internal paid promotions (such as direct advertising and display ads), it is advisable to concentrate efforts during these time periods to maximize traffic conversion.

In conclusion, the analysis indicates that users on the Taobao platform are primarily interested in finding products from the categories with identifiers 1863, 13230, 5027, 5894, and 6513. Among these categories, there is the highest demand for products in the category 1863. Users tend to concentrate their searches for these categories of products in time period 12 PM to 3 PM and time period 9 PM to 10 PM on the Taobao platform, with slight variations depending on the specific products within each category.

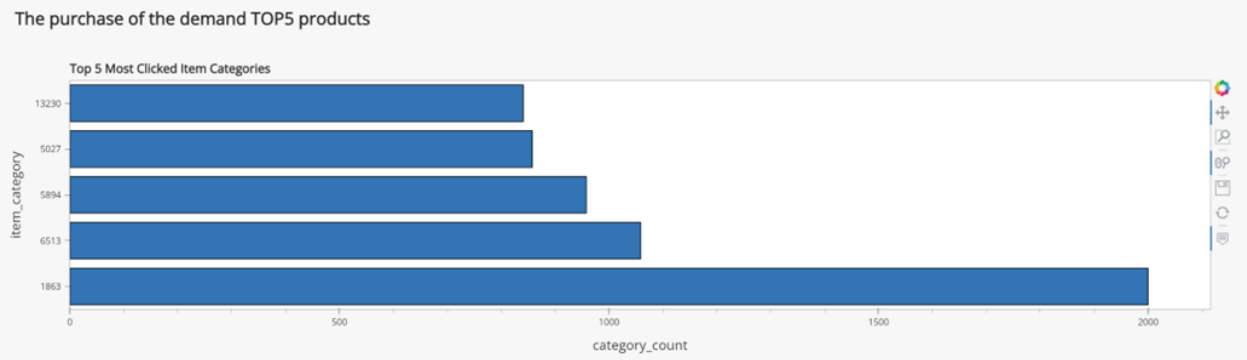
4.2.2.3 To what extent do the platform's product recommendations align with users' needs and preferences?

After analyzing what products users want to find on the Taobao platform, the next step is to determine whether the products recommended and presented on the platform meet user demands.

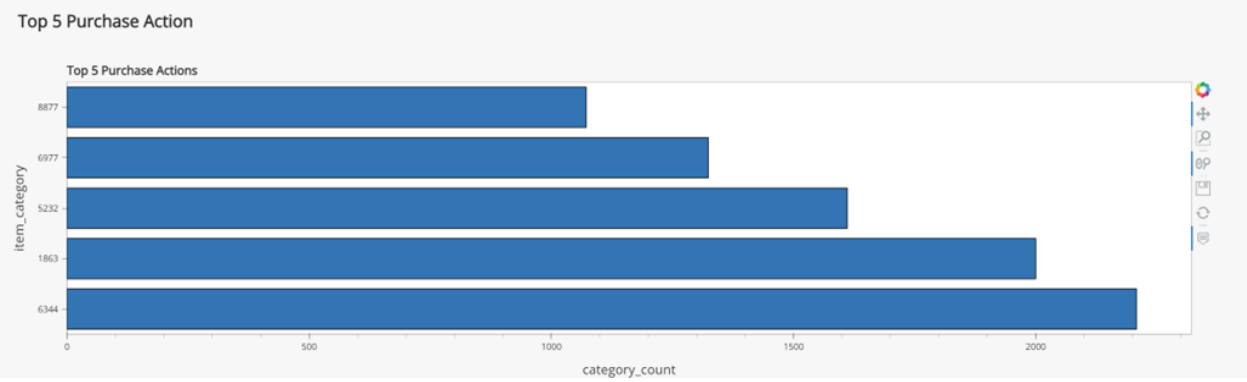
Firstly, we will analyze the proportion of these five product categories in terms of the total number of products available on Taobao, to assess whether there is an adequate selection for users.

Category	Item count	As a percentage
1863	20842	3.10%
5894	18693	2.78%
13230	18209	2.71%
6513	15854	2.36%
5027	15572	2.31%
In Sum	89170	13.25%

From the results shown in Figure 7, it can be observed that there are a total of 8,916 product categories on the Taobao platform, indicating a diverse range of offerings (with 2,876,947 unique items). Among these, the top five categories with the highest demand account for approximately 13.25% of the total number of products available, suggesting that Taobao provides an ample selection of options for these product categories.



With an adequate range of choices available, the next step is to analyze the mechanism through which Taobao recommends products in these five categories based on the number of purchases. This analysis aims to determine whether the products presented to users after searching for a specific category meet their demands.



From the above graph, it can be observed that apart from the category 1863, which has a relatively higher number of purchases, the remaining four categories with high demand have relatively low purchase frequencies, and none of them even rank among the top five in terms of purchase frequency.

Clearly, the purchase numbers for these top five categories with high demand are significantly lower compared to the top five categories with the highest purchase frequencies. This indicates that the recommendation mechanism employed by Taobao for these five

categories of high-demand products is not effective. The products recommended by Taobao in these categories fail to meet the users' expectations. As a result, users abandon the products after viewing them and do not add them to their shopping carts, leading to a decrease in conversion rates.

5. Conclusions:

Based on the above analysis, it confirms the previous hypothesis that the main reason for the significant user churn at the "add to cart" stage in the Taobao app is that users spend a considerable amount of time searching for products they desire but are unable to find them on the platform. Consequently, they abandon the idea of purchasing products on Taobao and turn to other platforms.

6. Recommendations:

Based on these conclusions, the following improvement suggestions are proposed:

Recommendation for the algorithm department: It is recommended that the algorithm department improve the product recommendation mechanism on Taobao, especially for the top five categories with high demand: 1863, 13230, 5027, 5894, and 6513. Monitor the purchase frequency of products in these categories and prioritize recommending products with higher purchase frequencies to users. After a user searches for products in these categories, they should be able to see products with high purchase frequencies among the top three or top five results. This will reduce the time users spend searching for products and improve their conversion rates.

Recommendation for the marketing department: Taobao users have a preference for the five categories: 1863, 13230, 5027, 5894, and 6513. It is recommended that the marketing department invest more in advertising these categories on Taobao's major marketing channels, especially focusing on advertising products with high sales within these categories. This will attract more users to the Taobao platform.

Recommendation for the operations department: Taobao users tend to search for products mainly between 12 PM to 3 PM and the time period 9 PM to 10 PM, which is the time when most people finish work and have leisure time. It is recommended that the operations department plan more marketing activities for these five categories during this time period, such as "group buy" or "discount" events, to stimulate user purchases and increase conversion rates. For products with high demand within these categories, the operations department should identify the specific time periods when users are most likely to search for them and plan promotional activities during those time slots. For example, as mentioned in the previous analysis, the highest demand within the top five categories is for products in the 1863 category. The analysis also provided the primary search time periods for the top three products with the highest click rates in the 5027 category. The operations department can plan targeted marketing campaigns during those precise time periods to activate users and improve conversion rates.

7. Limitation and further improvement:

Due to the limitations of the dataset, we can only provide a rough analysis of the types of products Taobao users seek. To conduct a more detailed analysis of the specific products users desire, additional information on users' frequent search terms is needed. By leveraging this

data, user search profiles can be established, and by combining it with product click count data, a search click-through rate metric can be developed. This will enable a more accurate identification of high-CTR (click-through rate) and low-CTR frequent search terms, thereby providing precise insights into the products users most desire to find on the Taobao platform.