Taobao User Shopping Behavior Analysis

By Zhengrong Lu (Ethan Lu)

A. Introduction

To refine precision marketing and operations and enhance Taobao's Gross Merchandise Volume (GMV), this project utilizes MySQL to preprocess the dataset provided by Alibaba Group, followed by multi-dimensional user and product analyses through data visualization in Tableau. Finally, actionable insights are presented in the conclusion.

B. Dataset

- Source: https://tianchi.aliyun.com/dataset/649

The dataset covers user behavior from November 25, 2017, to December 3, 2017, capturing interactions from approximately one million random users. The recorded behaviors include clicking (pv), purchasing (buy), adding to the cart (cart), and favoriting (fav). The dataset structure is similar to MovieLens-20M, where each row represents a single user action, consisting of user ID, item ID, category ID, behavior type, and timestamp, separated by tabs.

Table 1. Explanation of the columns of the datasets

Column names	**Description**
user_id	Integer type, serialized User ID
product_id	Integer type, serialized Product ID
category_id	Integer type, serialized Product Category ID
behavior_type	String type, categorical values including **'pv', 'buy', 'cart', 'fav'**
time_stamp	The time when the behavior occurred

Table 1.1: Behavior Types

Behavior Types	**Description**
pv	Product page view (clicks)
buy	Product purchase
cart	Adding a product to the shopping cart
fav	Favoriting a product

Table 2. Dataset sizes

Dimension	**Quantity**
Number of Users	987,994
Number of Items	4,162,024
Number of Categories	9,439
Total Number of interactions (i.e., Total rows of the dataset)	100,150,807

Due to time constraints and laptop capacity limitations, this project currently analyzes only the first 10 million interaction records out of the total 100 million rows.

C. Analysis Framework

The general formula of GMV:

GMV = Sales Price of Goods × Number of Goods Sold

The approximation formula of GMV:

GMV ≅ Sales Price of Product × (Product Page Views × Conversion Rate)

- The *Conversion Rate* measures the percentage of users who view a product page, engage with platform features such as adding to the cart and favoriting, and ultimately make a purchase, relative to the total number of users who viewed the page.

For enhancement of GMV, I used the approximation formula based on page views and conversion rate rather than direct order numbers, as the *Conversion Rate* captures user engagement across multiple interactions—viewing a product, adding to the cart, favoriting, and purchasing—enabling a more comprehensive analysis of the buying journey. This approach offers deeper insights into user behavior, helps identify bottlenecks in the sales funnel, and enhances marketing strategies, recommendation systems, and GMV prediction.

Without price data in the dataset, the analysis focuses solely on Product Page Views and Conversion Rates. For Product Views, metrics such as page views (PV), unique visitors (UV), and user retention rates are examined. Regarding *Conversion Rates*, the *Conversion Rate* of each buying path is analyzed using a funnel model, while a time-series table is created to track consumption trends across different time intervals.

Regarding precision marketing and operations, my approach is to analyze from both the user and product perspectives. On the user side, the RFM model is used for segmentation, while on the product side, similarity-based feature analysis is conducted.

Considering the two aspects mentioned above, the overall analysis framework is depicted in the diagram below.

Diagram 1. Overall analysis framework of the projet