eCommerce behavior data from multi category store

Mzoxolo Mbini

MIT805 Big data
Department of Computer Science,
University of Pretoria,
Hatfield,
Pretoria,
South Africa

email: u16350244@tuks.co.za

14 November 2021

1 Introduction

The trend of shopping online with e-shops goes hand in hand with the digital age. Shopping online seems to be a success story.

However, several vendors support developing retail shops to support face-to-face contact with customers as a unique selling point. Each way of shopping has its perks but also has downfallen the customer who chooses to buy.

Using commercial data to drive eCommerce business has become the gold standard. An eCommerce strategy must be based on data just like any other business. Without reliable data, there's no way to know the traffic on the website or even how the business is doing.

Understanding what drives customers' behaviour is the holy grail of the eCommerce industry. Knowledge can be used to improve the shopping process and ultimately result in higher sales and customer satisfaction.

REES46 Technologies built an eCommerce platform Open Online CDP for the United States-based retail stores. The Open Online CDP listens to the user activities (business processes) such requests, product view, move product(s) cart and update user details. All these processes are actioned and recorded in real-time.

The paper focuses on the technical aspects of the output of the recorded business processes. The paper continues to describe why and how the dataset is generated, its metadata, and insights that can be drawn from it.

2 Data exploration

The code base used to produces the data exploration is stored in the below repository
url: https://github.com/mzoxolombini/MIT805-Part2

2.1 Customer activity categories

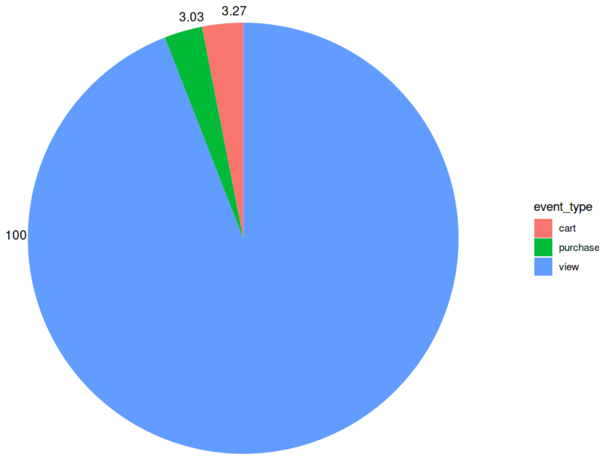

Figure 1: Customer activity categories

2.2 The date which customers shop the most on

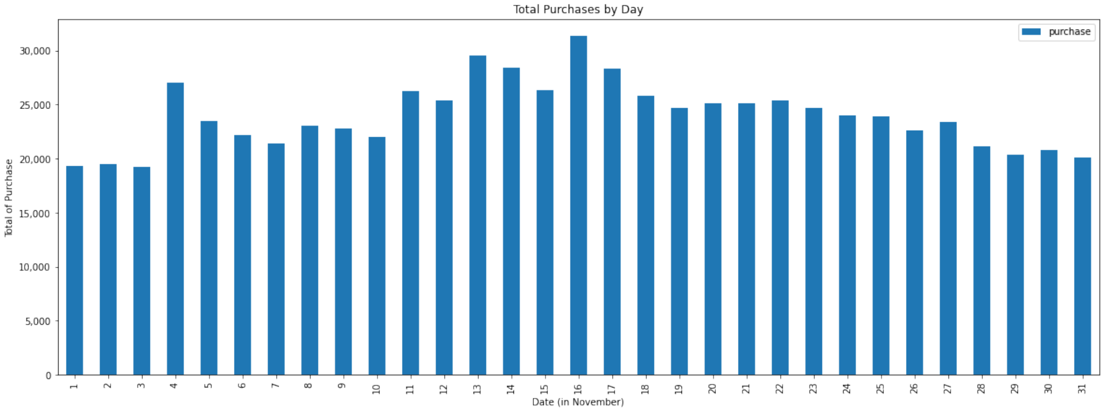

Figure 2: Date the customers shop the most in

The purchasing interest of users is gradually increasing from day 11 to 16 in the middle of the month, mid-month sale/discount offers could potentially start from day 11 to 16.

2.3 Prime time for customer shopping activities

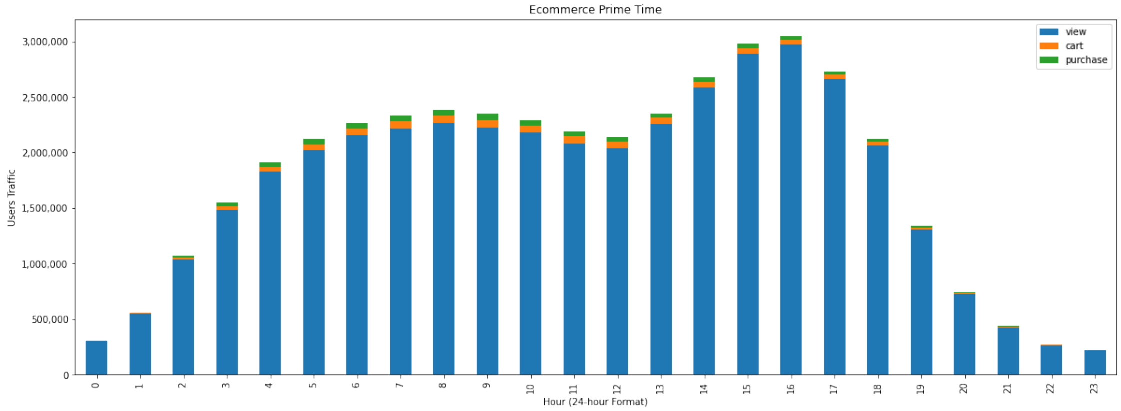

Figure 3: Customer's prime time shopping activities

eCommerce platform had already been accessed by 1.5 million users by 3:00 p.m. The rate is lower in the morning and rises significantly in the afternoon, peaked at 16:00. The use of flash sales from 13:00 to 16:00 can increase the user's impulsiveness to engage on shopping activities.

2.4 Goods brand distribution by activity

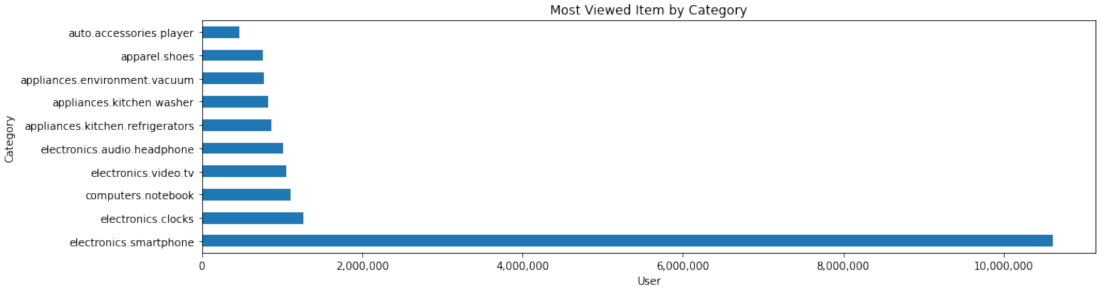

Figure 4: Most viewed item by category

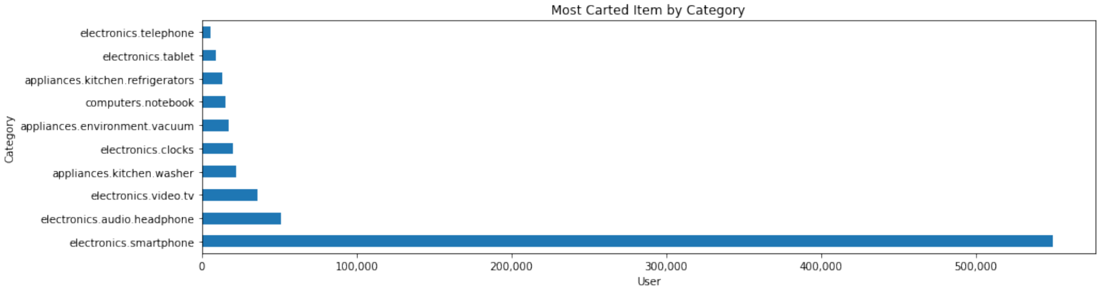

Figure 5: Most carted item by category

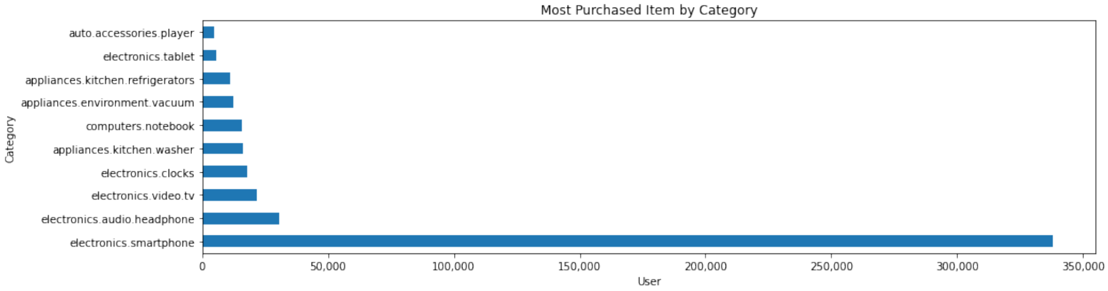

Figure 6: Most purchased item by category

Electronic smartphones are the most popular purchase in this eCommerce site. The most viewed, carted and purchased item in this eCommerce is an electronic smartphone, with a large lead over other items. As well as the electronic smartphone, users seem to be interested in other items such as audio headphones, video TV, clocks, computer notebooks, and so on.

As an eCommerce specializing in electronics, it can be difficult to establish a strong brand positioning for the business. Sales can be increased by identifying what users are interested in; for example, besides electronic smartphones, a lot of customers add audio headphones and video television to their carts; by doing so, a promo codes or hold a sale on that particular category to attract more customers.

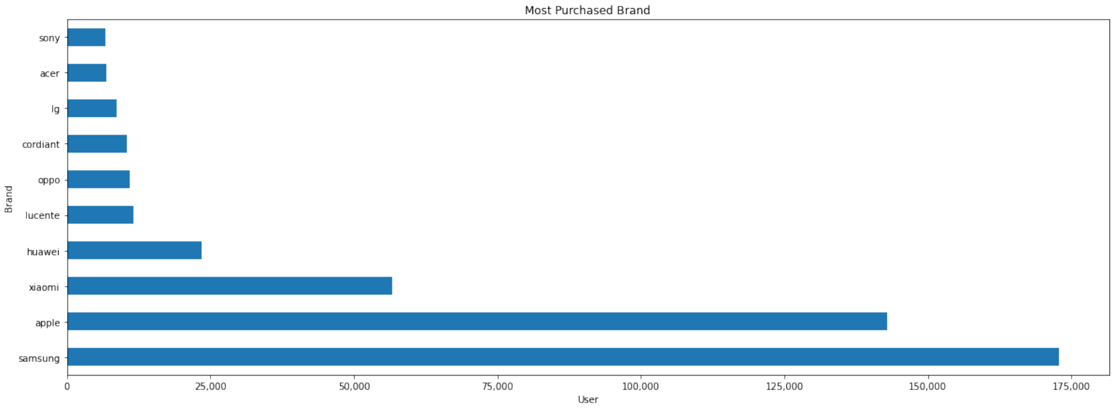

Figure 7: Most purchased brand

Samsung is the most popular brand among users, followed by Apple and Xiaomi, so we can collaborate with Samsung to increase sales.

2.5 Price distribution

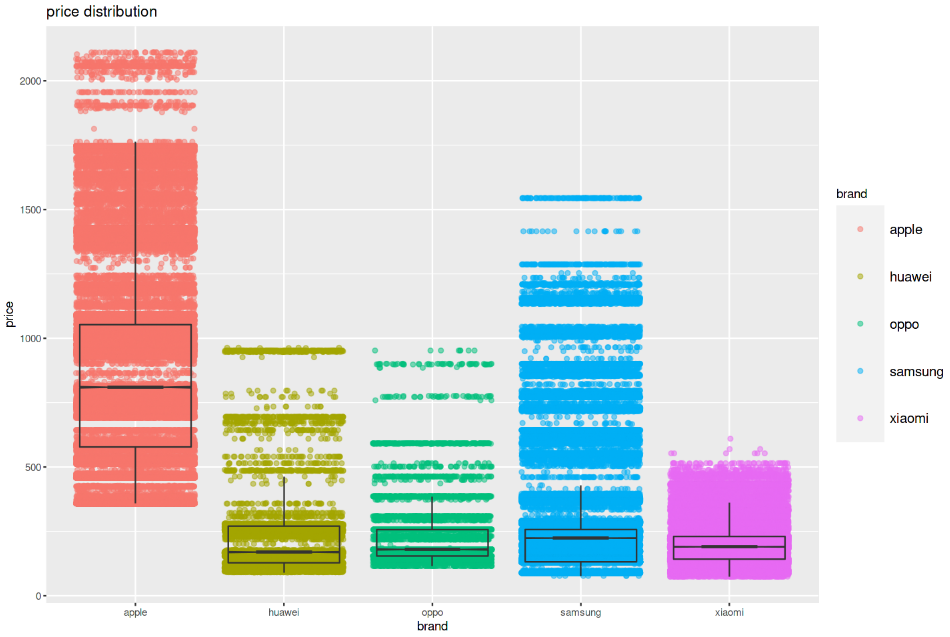

Figure 8: Goods price distribution

Huawei and Oppo have an identical distribution of prices, and this is caused by the fact that they are the new players in the market.

3 Predictive modeling

Data needs to be restructured so that machine learning models can be applied to it. In this use case, I only target customers who have placed a product in their cart. The training data set includes the following new features as well:

- category_code_level1: category
- category_code_level2: sub-category
- event_weekday: the day of the event
- activity_count: Number of activities during the session
- is_purchased: is item in the cart purchased?

So, the training dataset contains all non-duplicated cart transactions in the same session; with the above-mentioned new feature, I keep only one record per specific product. Those characteristics, along with the original price and brand, will help me predict whether customers will eventually buy the item.

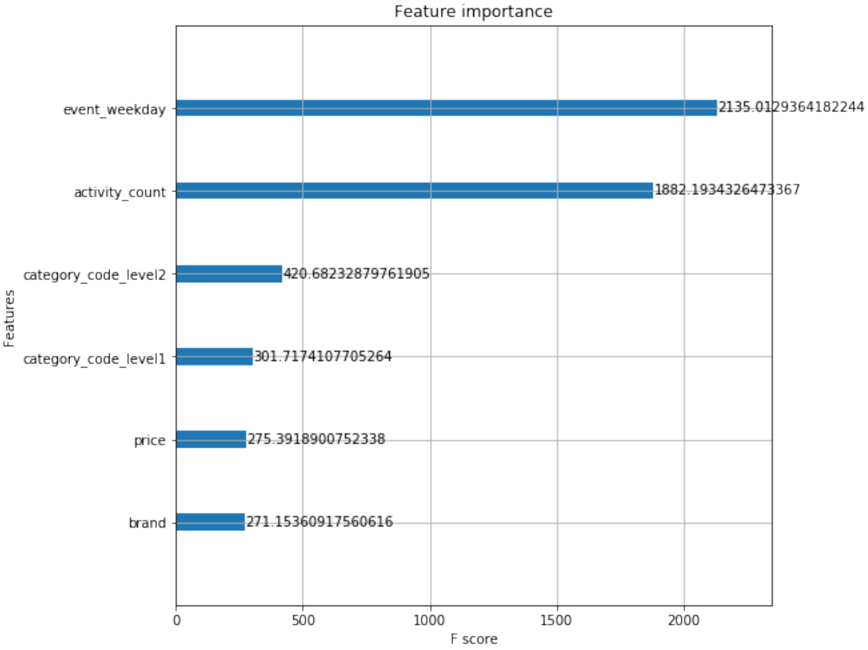

Figure 9: Feature importance