CASE STUDY: eCommerce Clickstream data for Recommender System

Problem Statement:

The objective here to measure the effectiveness of a recommendation engine on an eCommerce website by creating appropriate segments and discovering relevant KPI's to measure the same.

Overview of the Dataset:

There are two datasets provided -

1. Events data - represent the clickstream data created to emulate data collected from a client's website / app
2. Product metadata - contains all the data associated with the products on the client's catalog

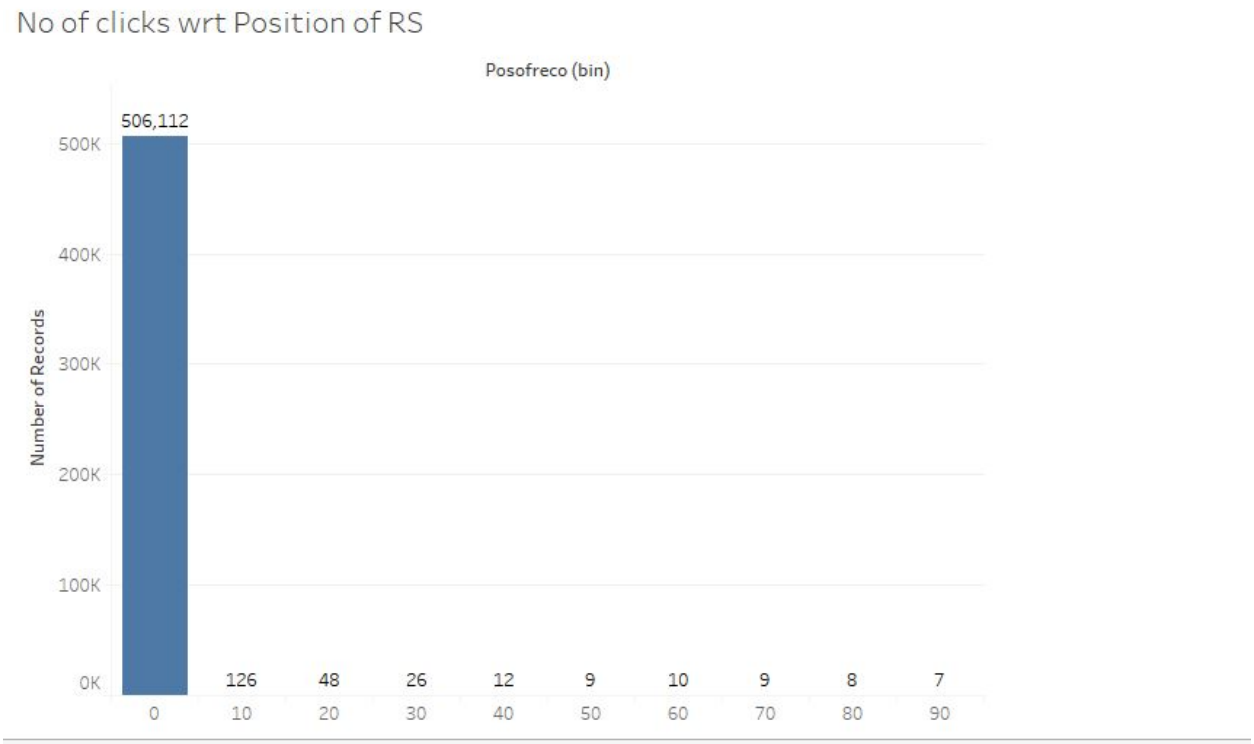
- Data is from 1st Sept 2017 - 15th Sept 2017
- There are 4 major pages - Home page, PDP, Cart Page and Search Page
- The data contains traffic from both Desktop website as well as Mobile application
- Every customer can be identified by a unique ID
- A new session is created every time and a subsequent ID is assigned to it.

Cleaning the Dataset:

Given dataset had a lot of noise and inconsistency which was resolved in the following manner -

- Regarding pagetype variable -
    a. Replaced "1_homepage" with homepage
    b. Replaced "42_searchpage" with searchpage
    c. Replaced "4_cart" with cart.
    d. Combined other variable into a new feature as "Others"

- Regarding posofreco variable -



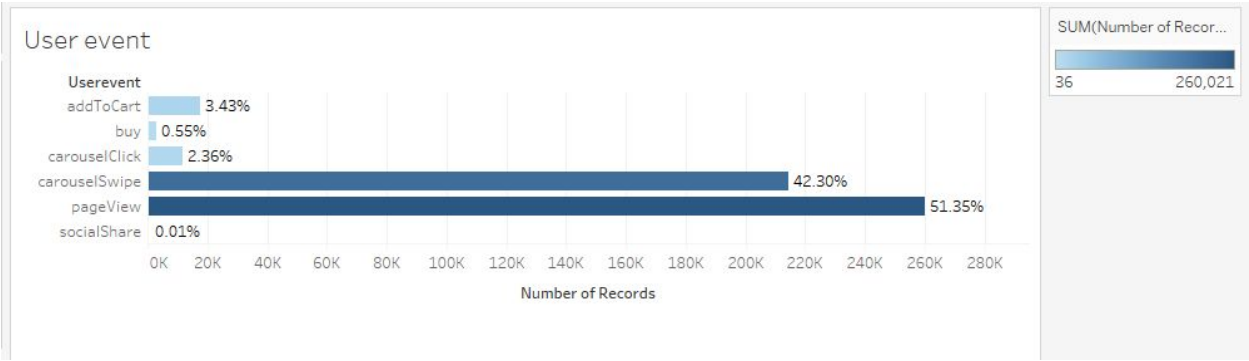
    - As you can see, 0th - 10th position has about 99% of interactions. So I've combined the rest as a new feature called as ">10"
    - I'm also assuming that 0th position is the first product on the recommendation carousel (similiar to array norm)

- Regarding noisy variables -
    - I've removed clicked_epoch & userevent from my analysis since I feel that they contain redundant information

Exploratory Data Analysis:

In Order to discover potential trends, segments and KPI's, let us carry out an ad hoc analysis on our dataset

1. User event

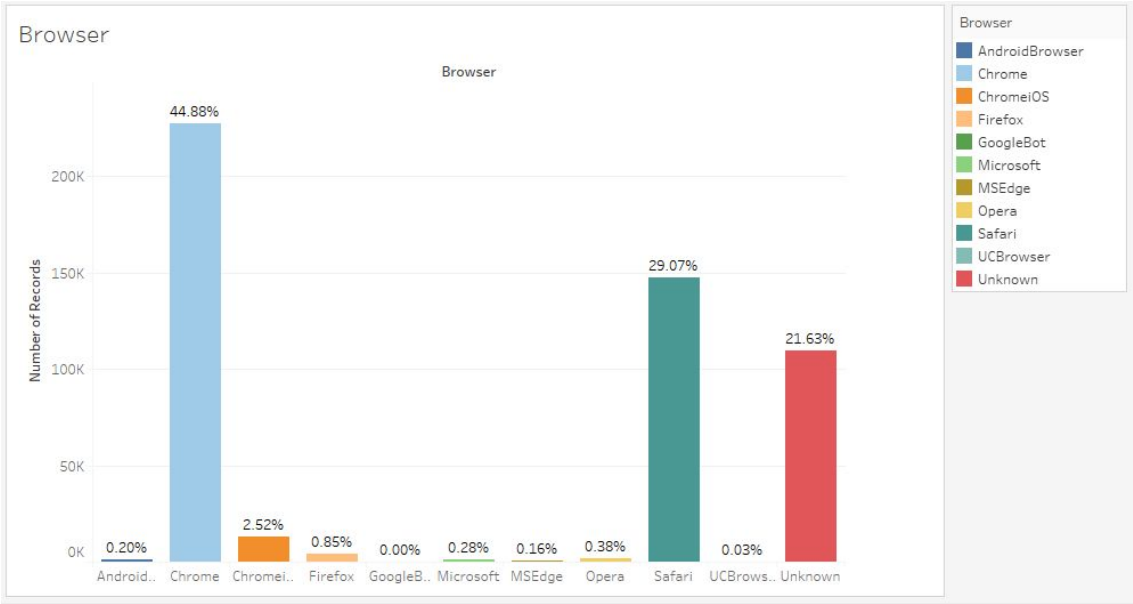


It is observed that about 93% of our clickstream data contains user events such as pageView & carouselSwipe while only a fraction of users are making an actual purchase

2. Date of Events

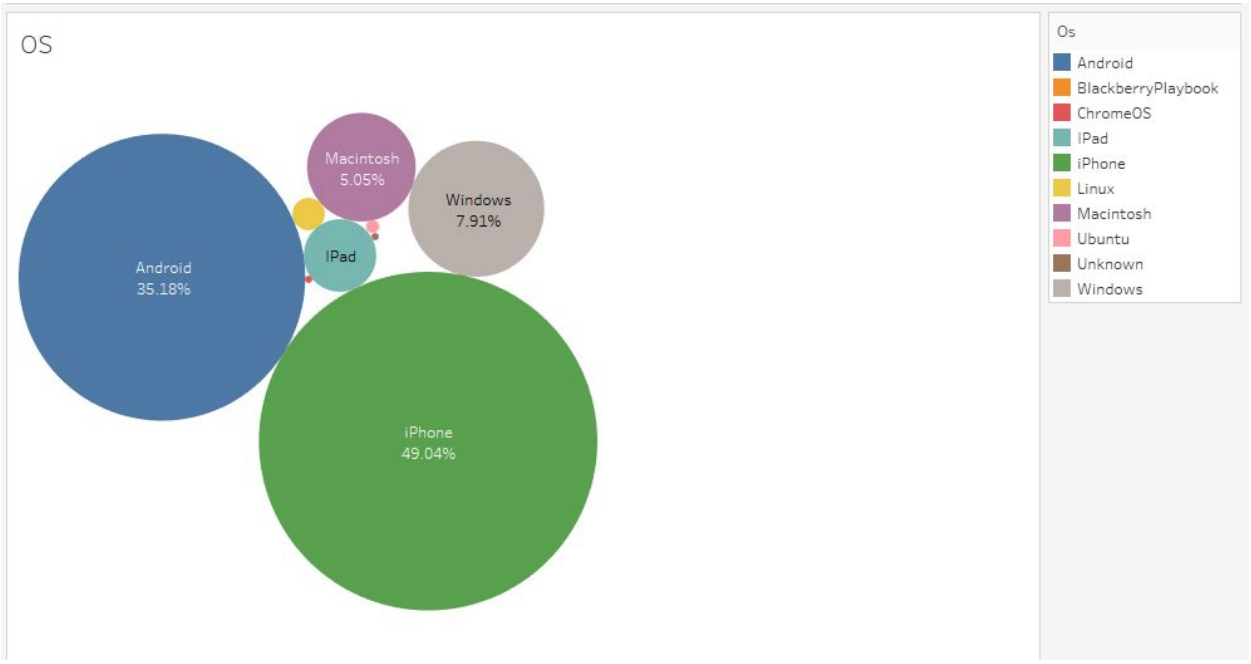
Traffic across all the 15 days

3. Browser Details

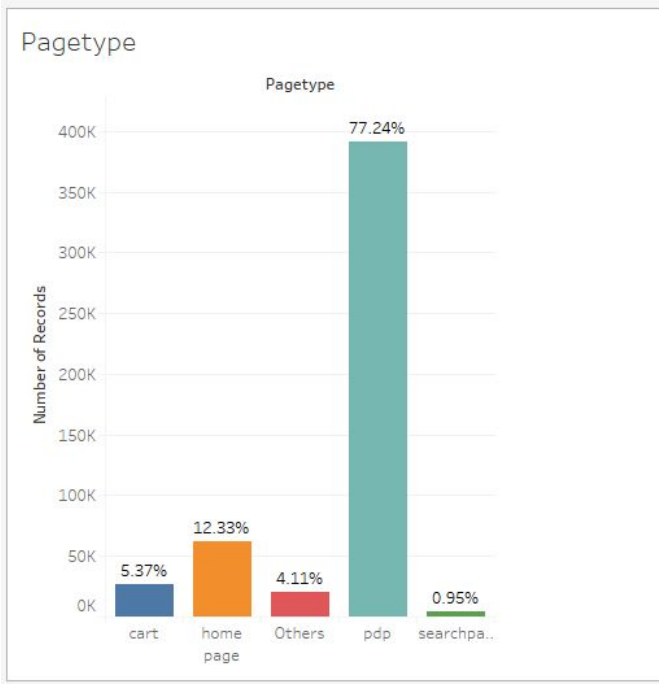


Chrome is clearly dominating all the other browsers. Company can take benefit of this fact and develop a Chrome extension that sends browser notification for a price drop alert (for eg)

4. Phone OS -

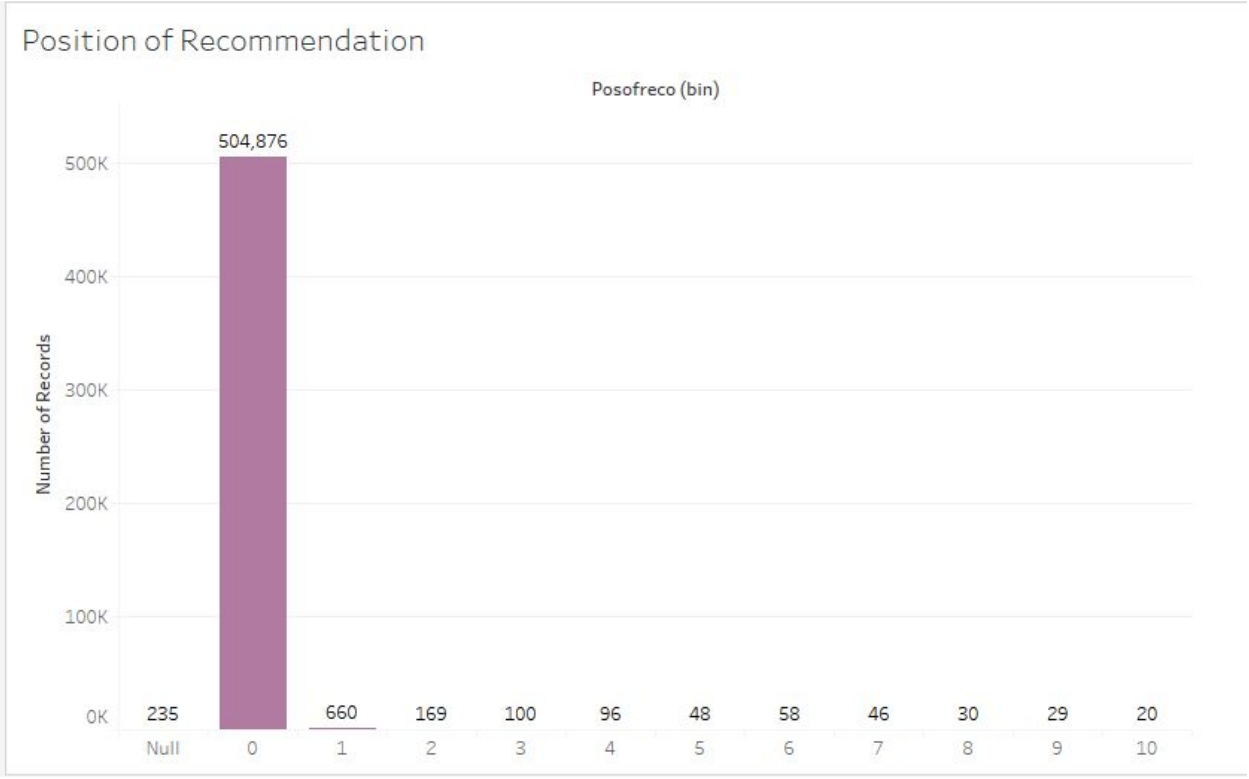


5. Pagetype



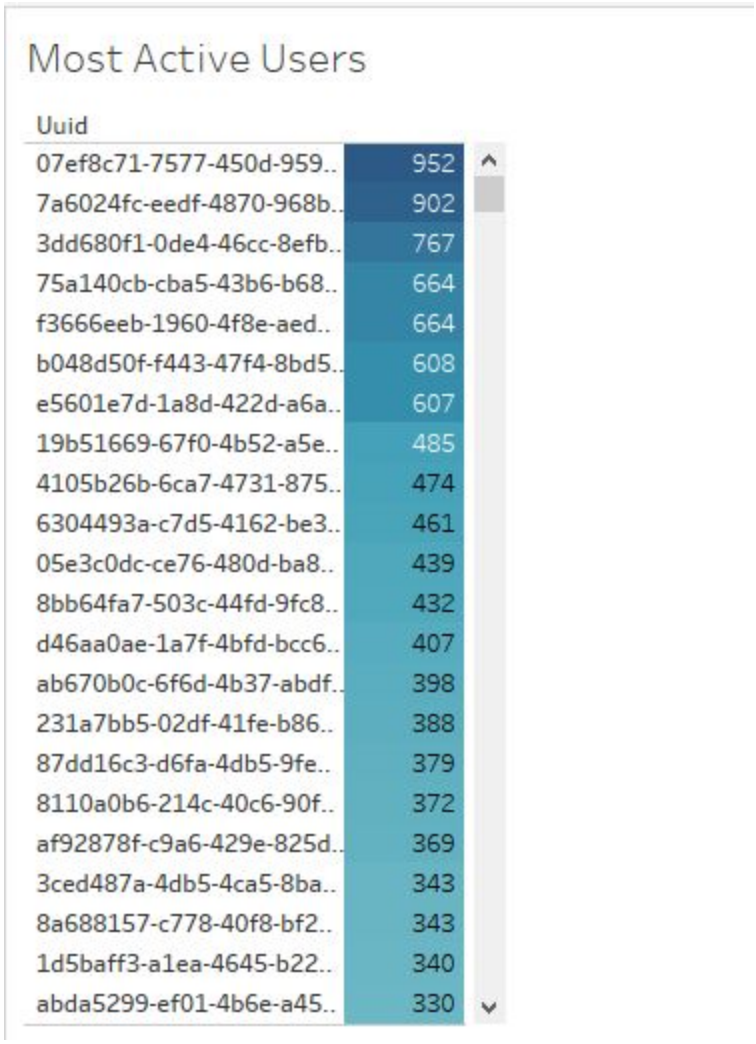
Most of the RS clicks are happening on PDP and Home page, this indicates that we can save costs by taking down RS on other pages. Afterall, the 80/20 rule

6. Position of Recommendation

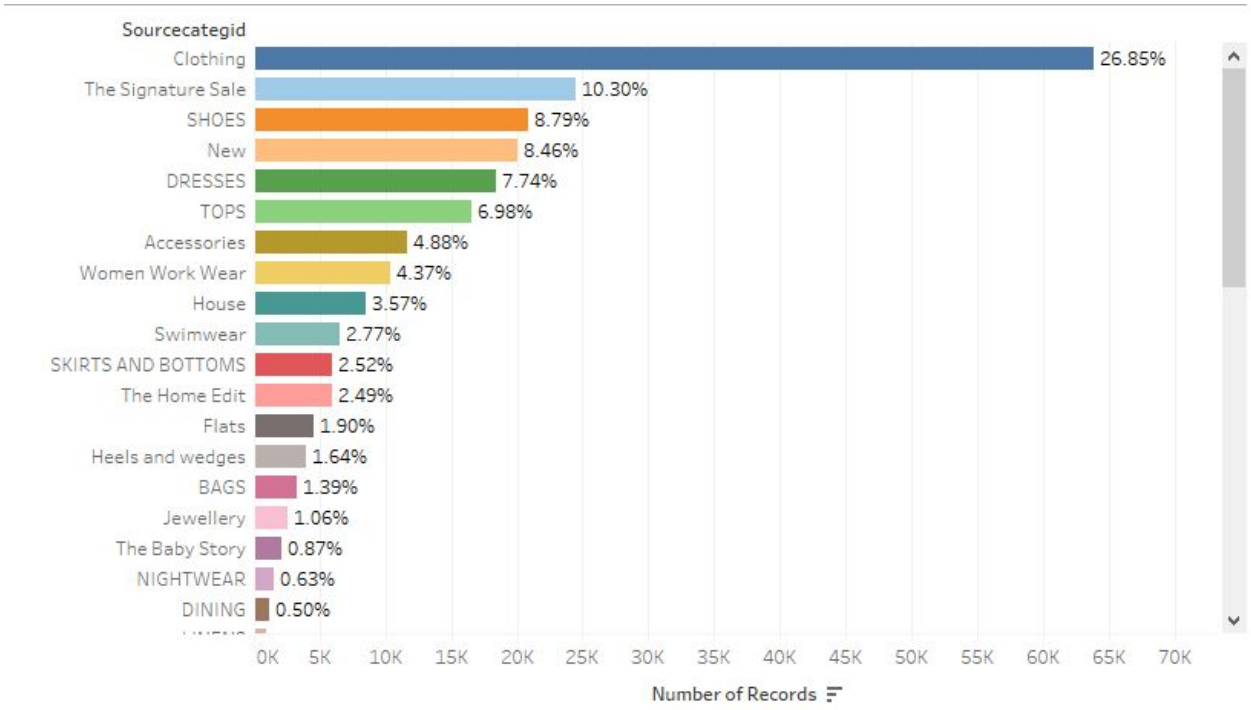


The first product on the carousel is more likely to get clicked on when compared to others.

7. Most Active Users



8. Most clicked Categories -



That brings us to the end of Exploration section.

Evaluation Metrics:

Our main objective is to measure effectiveness of our recommendation engines. We are going to use certain standard metrics to achieve the same.

Primarily, I've divided the metric into three major areas -

1. **Impact** - The impact measures include the places where the recommendations are presented, for example, the home page, the product list page or the shopping cart page; and the number of recommendation list.

2. **Interaction -** It can be defined as the no of clicks on the recommended items

3. **Conversion -** Are customers actually placing an order? We can check this by a few means such as determining the growth in revenue, growth in the number of orders through the RS.

Keeping this three broad categories in mind, the following are few possible metrics for evaluating our recommendation system

- REC: number of recommendations presented in a list.
- LOC: places where the recommendation lists are placed.
- CER: total of clicks in the recommendations
- CTR (%): rate of clicks in the recommendations
- TPR (%): proportion of orders with recommendations
- IR (%): increase in the revenue

Customer Segmentation & KPIs:

On the basic of our exploratory data analysis, we have understood that there are few features that have the statistical significance and so we are going to segment our population based on that.

1. Customers on Web vs Customer on Mobile -

Basically we are segregating our desktop customers and mobile customers and analyzing their clickstream data. So let's look at some numbers

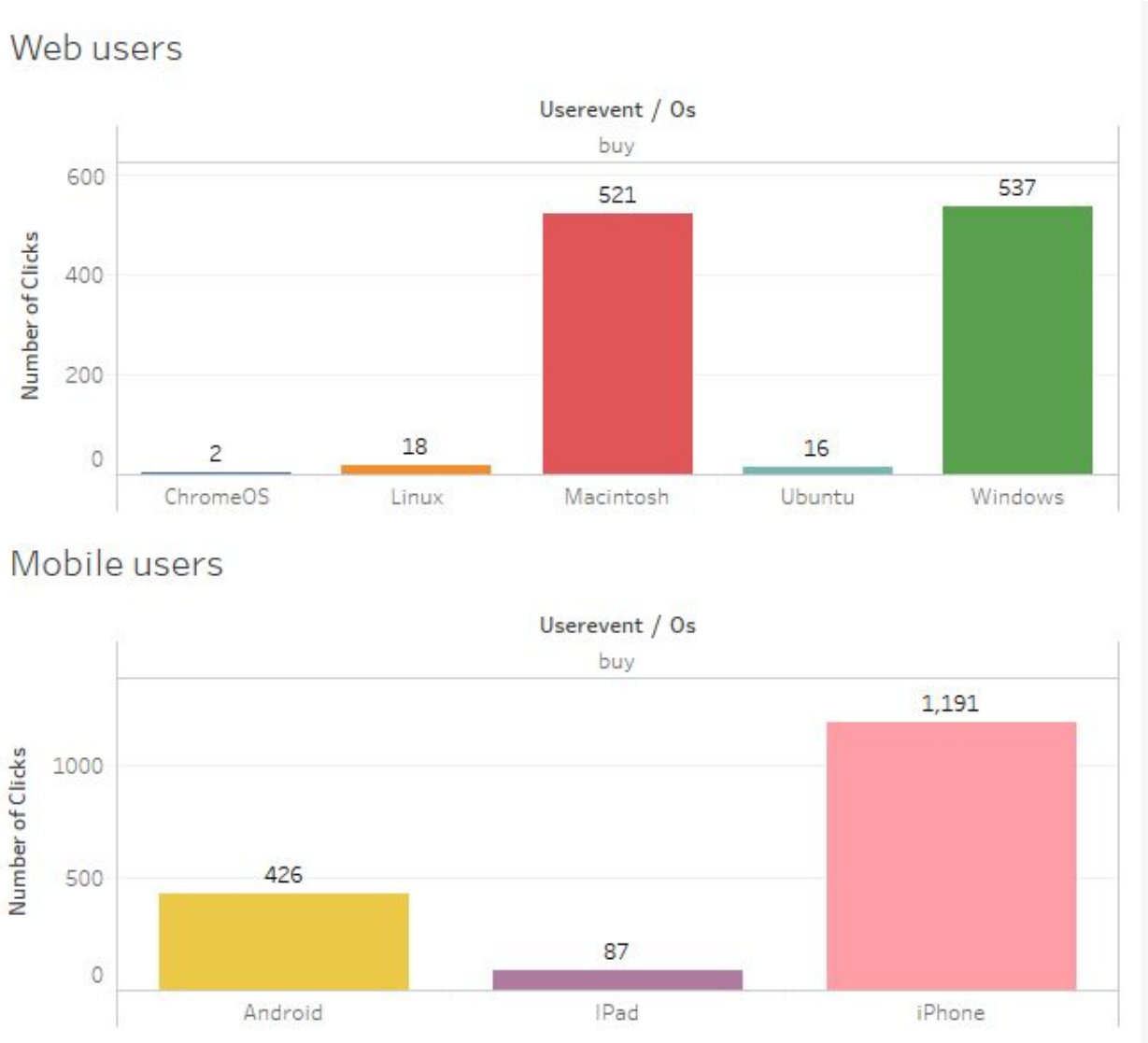
→ Calculating CTR -

Customers on Web: They majorly use OS such as Windows, Ubuntu and Macintosh. There are about 68,389 clicks made through web which accounts for 13% of total clicks

Customers on Mobile: They use OS such as Android, iOS, Blackberry. There are about 4,37,861 clicks accounting for 86% of total clicks

→ Proportion of Orders -

This can simply be calculated by Total number of Orders / Orders through recommendation carousel. However, in our scenario, we are going to segment it into Web users and Mobile users pertaining to only “buy” userevent



Purchases made by Web users = 1094 so Conversion rate = 0.21%

Purchases made my Mobile users = 1704 so Conversion rate = 0.33%

2. Clicks through PDP vs Home Page

Here we are trying to discover a trend in our incoming traffic. Is it through PDP or any other page? If so, by how much?

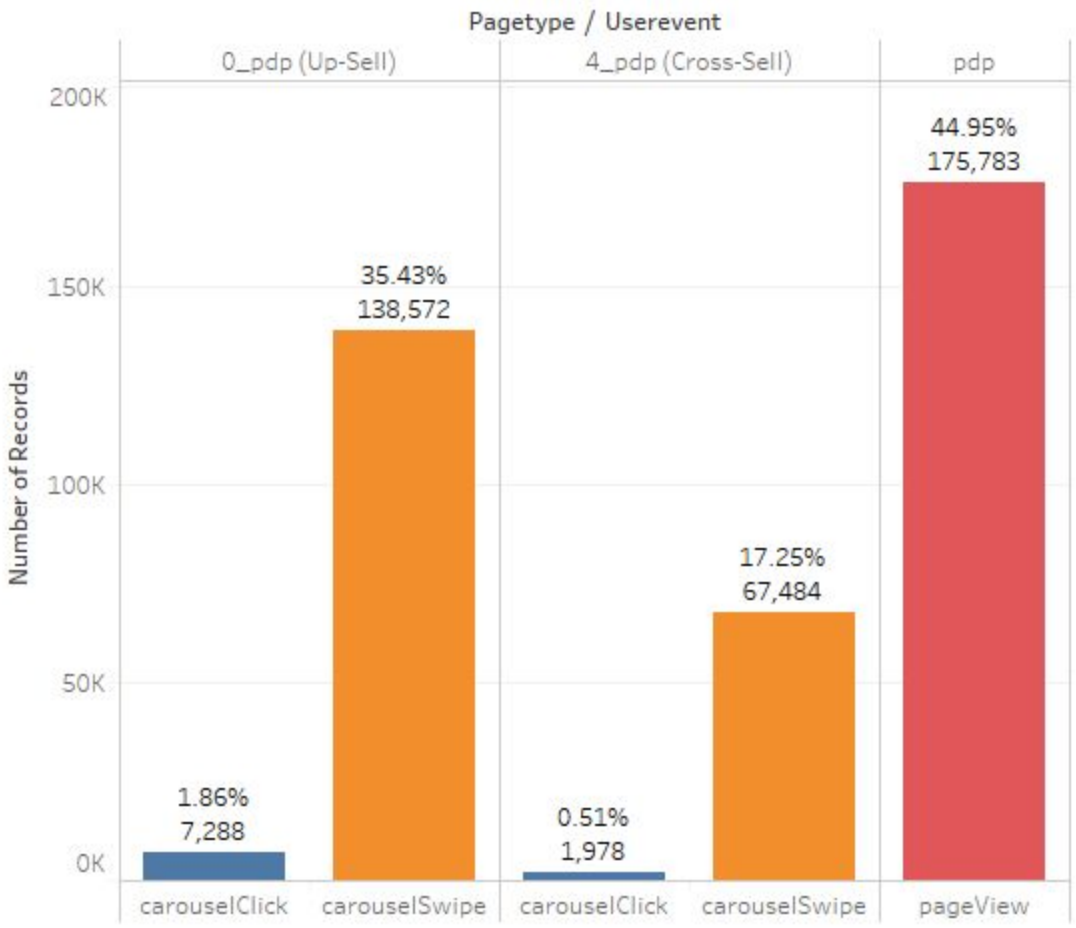
There are 3 sub categories in PDP -

- 0_pdp which represents products of similar category. This implies that here the RS does “up selling”
- 1_pdp which represents products of different category. This implies that here the RS does “cross selling”
- Generic PDP which shows recommendations from other sources

Calculating CTR -

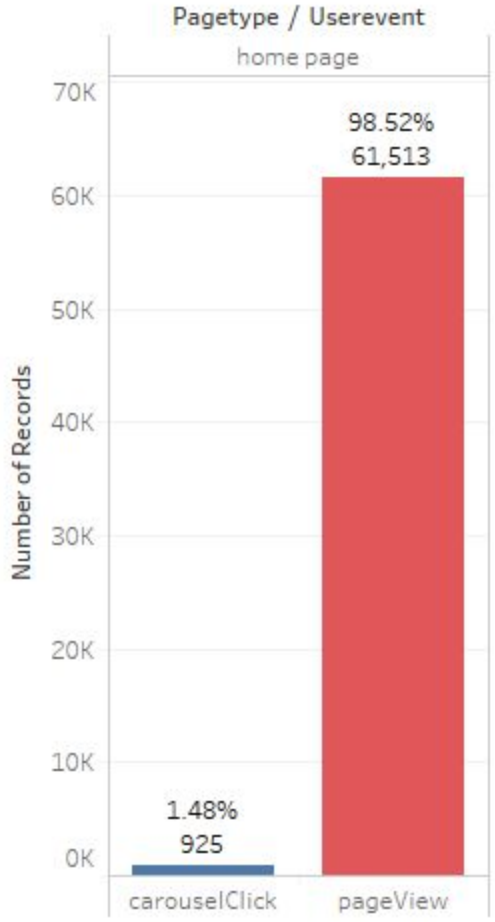
Clicks through PDP: About 77% (or) 3,91,105 of clicks are through PDP in total.

This is the distribution of different PDP’s



This shows that absolutely no purchases are made through PDP page. However, people are more likely to do a carousel swipe if they are shown products of same category

Clicks through Home Page: About 62,438 clicks are through the Home page carousel which accounts to 12.33%



3. Based on Position

About 99.7% of clicks are on first product on the carousel.

The last KPI is **Increase in Revenue (%)** which can be calculated with the help of sales data.

Since we only have data of 15 days, we cannot come to any solid conclusions but few insights can be derived

Further Approach -

1. Map the Metadata with Client data based on a common “product_id” to understand the trend with respect to prize
2. Calculate the total sales in 15 days
3. Create a content based recommendation engine using product ID and desc and try to increase the upsell.
4. Introduce Time data as well, to understand how much time a customer spends on a particular page

Insights -

1. Customers are more likely to buy a product from the Mobile application
2. Similar products must be given more importance than Cross selling
3. Home page can view “Best sellers” and “Most views items”
4. Carousel of 5 products will be beneficial
5. As of now, “Complete the Look” engine needs more prediction accuracy

This assignment was really fun to work on.

Thanks,
Naman
namanjd@outlook.com
www.namandoshi.co