E-COMMERCE CUSTOMER SEGMENTATION & PRODUCT RECOMMENDATION

Group 5

Debatra Chatterjee (0342/61)

Pranav Bahl (0371/61)

Priyanka Janged (0054/61)

Saloni Bansal (0388/61)

Group 5 - E-Commerce Customer Segmentation and Product Recommendation

Business Context/Background:
In the e-commerce space there is a growing need of moving towards hyper-personalized marketing targeted to individual existing customers. While customer acquisition is important for e-commerce platforms, revenue is ultimately generated through customer conversion (from viewing a product to purchasing the product) and customer retention. Because traditional electronic marketing campaigns like emails and SMS are generic, e-commerce platforms are looking for more personalized marketing campaigns targeted to individual customers based on their preferences, which will increase both customer retention (as the existing customers will be more willing to come back from the tailored ad) and customer conversion (the existing customer will be more likely to purchase the relevant product)

Business Objective:
To increase the conversion rate of existing customers (currently ~2%) by conducting hyper-personalised marketing campaigns for existing customers based on analyzing and predicting customer behaviour from e-commerce transaction data.

Analytics objectives:

Bucketing of existing customers dependent on their purchasing behaviour and preferences using transaction data we got from the E-commerce segment. The analytics will help us create hyper-personalized marketing campaigns and increase both conversion and retention of customers.

Specific questions that you seek to answer using Data Mining

1. What are the different clusters created based on event_type, price and category_id?
2. Which user cluster will be more likely to come back to our platform?
3. What products to recommend to the selected user cluster to impact the conversion?

Overview of the data set – source, no. of records, fields, etc.

Data - https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

(We are only using Oct-2019 data)

Data Source - Kaggle

No of records - 42448764

No of fields - 9

Fields - ['event_time', 'event_type', 'product_id', 'category_id', 'category_code', 'brand', 'price', 'user_id', 'user_session']

Blank fields -

df.isna().sum()

```
PS C:\Users\DEBATRA\Downloads\archive> python test.py
event_time              0
event_type              0
product_id              0
category_id             0
category_code    13515609
brand             6117080
price                   0
user_id                 0
user_session            2
dtype: int64
PS C:\Users\DEBATRA\Downloads\archive>
```

Data Overview -

df.head()

```
                 event_time event_type  product_id          category_id  \
0  2019-10-01 00:00:00 UTC       view    44600062  2103807459595387724
1  2019-10-01 00:00:00 UTC       view     3900821  2053013552326770905
2  2019-10-01 00:00:01 UTC       view    17200506  2053013559792632471
3  2019-10-01 00:00:01 UTC       view     1307067  2053013558920217191
4  2019-10-01 00:00:04 UTC       view     1004237  2053013555631882655

                       category_code     brand    price    user_id  \
0                               None  shiseido    35.79  541312140
1  appliances.environment.water_heater      aqua    33.20  554748717
2          furniture.living_room.sofa      None   543.10  519107250
3                computers.notebook    lenovo   251.74  550050854
4              electronics.smartphone     apple  1081.98  535871217

                         user_session
0  72d76fde-8bb3-4e00-8c23-a032dfed738c
1  9333dfbd-b87a-4708-9857-6336556b0fcc
2  566511c2-e2e3-422b-b695-cf8e6e792ca8
3  7c90fc70-0e80-4590-96f3-13c02c18c713
4  c6bd7419-2748-4c56-95b4-8cec9ff8b80d
```

df.info()

```
PS C:\Users\DEBATRA\Downloads\archive> python test.py
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42448764 entries, 0 to 42448763
Data columns (total 9 columns):
 #   Column         Dtype
---  ------         -----
 0   event_time     object
 1   event_type     object
 2   product_id     int64
 3   category_id    int64
 4   category_code  object
 5   brand          object
 6   price          float64
 7   user_id        int64
 8   user_session   object
dtypes: float64(1), int64(3), object(5)
memory usage: 2.8+ GB
```

df.describe() [Only applicable for price column]

```
None
         product_id   category_id         price       user_id
count  4.244876e+07  4.244876e+07  4.244876e+07  4.244876e+07
mean   1.054993e+07  2.057404e+18  2.903237e+02  5.335371e+08
std    1.188191e+07  1.843926e+16  3.582692e+02  1.852374e+07
min    1.000978e+06  2.053014e+18  0.000000e+00  3.386938e+07
25%    1.005157e+06  2.053014e+18  6.598000e+01  5.159043e+08
50%    5.000470e+06  2.053014e+18  1.629300e+02  5.296965e+08
75%    1.600030e+07  2.053014e+18  3.585700e+02  5.515788e+08
max    6.050001e+07  2.175420e+18  2.574070e+03  5.662809e+08
```

Which techniques do you plan to use? Why?

1. K-Means Clustering: K-Means is fast for large numeric datasets; and as the dataset that we are using is large and numeric, so K-Means clustering method is appropriate to use here to create clusters of user segments.
2. Frequent Pattern Growth (FP-Growth): FP-Growth is more efficient in forming associations between data in large datasets hence FP-Growth is appropriate to use for our dataset to find out product associations while purchasing.

How will results from your analytics plan help you solve your Business Problem?

By having user clusters that are more likely to be repeat customers or purchasing customers, and recommending products that they are more likely to purchase, we can plan hyper-personalized marketing campaigns which should see an increase in customer retention and conversion.