# E-Commerce Churn Assignment

# Report

The aim of the assignment is to build a model that predicts whether a person purchases an item after it has been added to the cart or not. Being a **classification problem**, we can use the any of the classification model here in this assignment; the model development is done by logistic regression, decision tree and random forest algorithms. The most robust model is selected finally and an optimal solution is provided to predict the churn in the most suitable manner.

For this assignment, we are provided the data associated with an e-commerce company for the month **of October 2019**. The dataset stores the information of a customer session on the e-commerce platform. It contains **42 million** data points. It records the activity and the associated parameters with it.

- **event_time**: Date and time when user accesses the platform
- **event_type**: Action performed by the customer
    - - View
    - - Cart
    - - Purchase
      - Remove from cart
- **product_id**: Unique number to identify the product in the event
- **category_id**: Unique number to identify the category of the product
- **category_code**: Stores primary and secondary categories of the product
- **brand**: Brand associated with the product
- **price**: Price of the product
- **user_id**: Unique ID for a customer
- **user_session**: Session ID for a user

# Data Exploration

Data exploration of the data set is done and the findings are given below

- Dimension of the Dataset is: (42448764, 9)
- The Product with product id **1004856** is the most popular product sold.
- The top 5 popular products sold are:

| product_id | count |
|---|---|
| 1004856 | 28944 |
| 1004767 | 21806 |
| 1004833 | 12697 |
| 1005115 | 12543 |
| 4804056 | 12381 |

- The top 5 popular brands are:

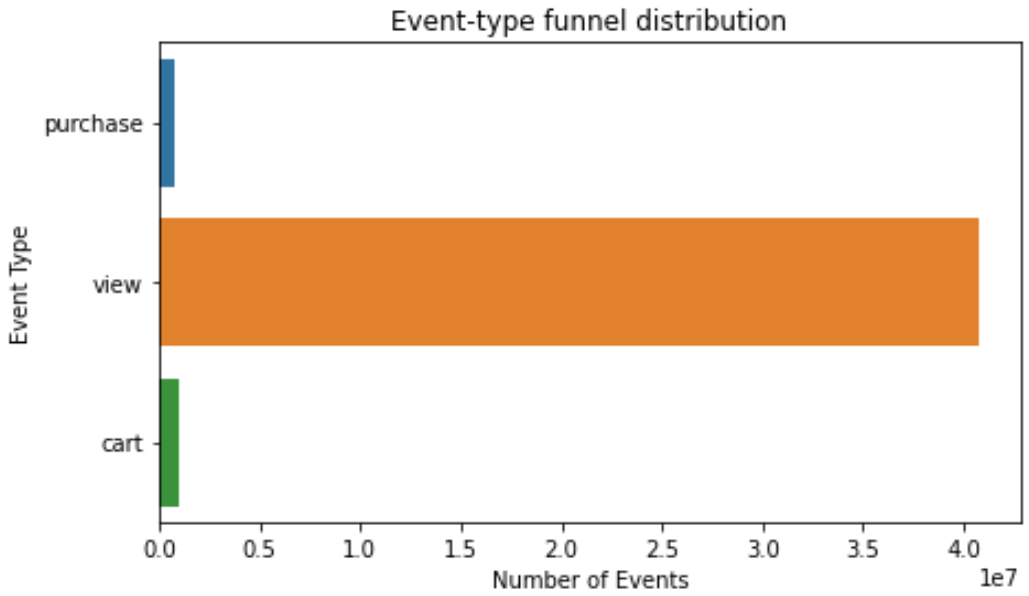
| brand | count |
|---|---|
| null | 6113008 |
| samsung | 5282775 |
| apple | 4122554 |
| xiaomi | 3083763 |
| huawei | 1111205 |
| lucente | 655861 |

- It is clear that most of the products sold on don't have a brand associated with them. The five most popular brands are**: Samsung, Apple, Xiaomi, Huawei, Lucente**.It i s clear that the most popular product is mobile phone.
- The 5 most popular product categories are:

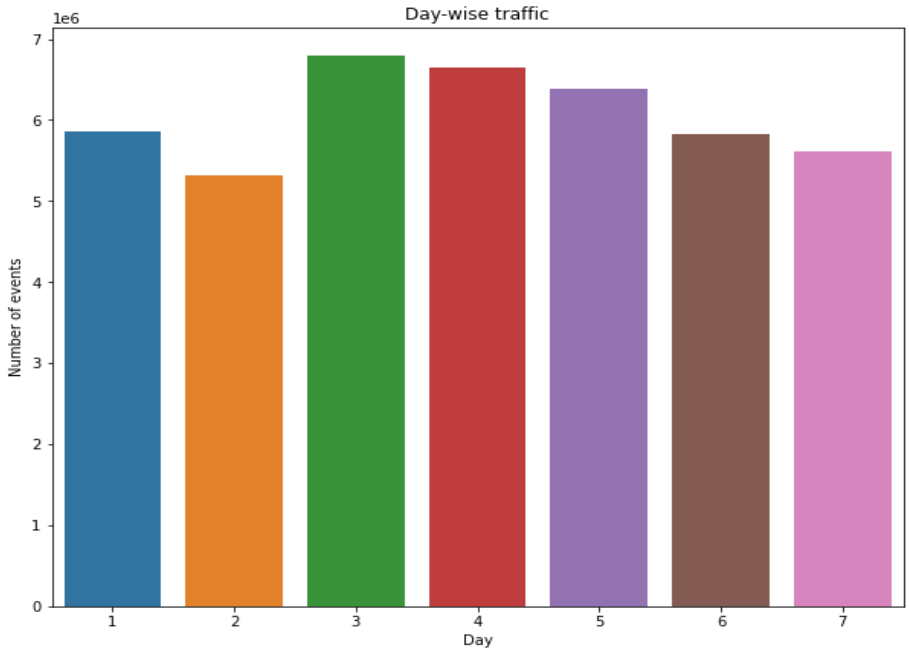
| category_code | count |
|---|---|
| null | 13515609 |
| electronics.smart... | 11507231 |
| electronics.clocks | 1311033 |
| computers.notebook | 1137623 |
| electronics.video.tv | 1113750 |
| electronics.audio... | 1100188 |

- There are **3022290** unique active users on the platform.
- Most active user is **512475445**. He/She has registered 7436 user sessions

- Average price of purchased smartphone : 464.61911309456883
- Maximum price of purchased smartphone : 2110.45
- The average and maximum value of prices indicate that the prices are not in rupees.
- The sales funnel can be evidently seen from the below plot




- The plot for the traffic on different days of the week is as given below




- Maximum traffic is on 3rd day of the week. we can find that it Tuesday.
- The expectation that the traffic will be more on weekends is not true according to this plot.

# Model Develpoment

## Feature Engineering

- There are 13515609 null values in the catrgory_code column and 611300 null values in b rand column. Since, there is no other column to identify them and thus to impute values. Hence, we will drop the null values. Eventhough, product_id can be checked to identify s ome, we have a huge amount of data since dropping won't affect much
- The duplicate rows are also dropped, if any.
- To capture the user activity in different stages of the sales funnel, the data is categorized using window function.
- Event_time variable is changed to event_hour to understand the peak hours of the day in the site.
- The brand column in binned by selecting the top 20 brands and the rest under "others" category to facilitate the encoding of these variables.
- Once the preprocessing is done the target variable "is_purchased" is generated.
- Reductant columns are dropped and the cleaned dataset is saved in the parquet format for model building.

## Model Building, Evalauation and Selection

- 3 model of classification – Logistic Regression, Decision trees, Random Forests
- The features are transformed- Categorical features are first encoded using string indexer and then both numerical and encoded categorical features are encoded using the onehotencoder
- The transformed dataframe is saved in parquet format to avoid repetition of the feature transformation steps.
- The features are scaled and the dataset is split to train-test sets
- The split is done in the 70:30 ratio.
- **Logistic Regression:**
    - The train data is fitted to logistic regression model with default values and the model is evaluated on test data
    - The threshold is found and is optimized by plotting the different metrics and the intersection value is taken as the threshold.
    - **Recall** is the appropriate metric since here we are concentrating on the churn. ie the 1's.
      Here we can see that the recall is about 0.76

- **Decision tree:**
    - The transformed dataframe after feature selection is split into train and test set in the ratio of 70:30.
    - The features need not be scaled since it has little to no effect on tree and ensemble forest models.
    - The train data is fitted to decision tree model and is the model is evaluated on test data
    - For obtaining the optimum parameters the parameter grid builder is used.
    - The hyper parameter tuned model is fitted and evaluated on test data.
    - **Recall** is the appropriate metric in this scenario..
      Since we are concentrating on the number of churns, that is 1's, recall is the best metric and here we have around **0.85** as recall value.
    - Decision tree model is better than logistic regression which had recall of **0.76**

- **Random Forest:**
    - The transformed dataframe after feature selection is split into train and test set in the ratio of 70:30.
    - The features need not be scaled since it has little to no effect on tree and ensemble forest models.
    - The train data is fitted to decision tree model and is the model is evaluated on test data
    - For obtaining the optimum parameters the parameter grid builder is used.
    - The hyper parameter tuned model is fitted and evaluated on test data.
    - **Random Forest** model gives the best performance out of all three. Eventhough, decision tree model gave a recall of about 0.85, this one gives a better recall of **0.92** and all other metrics do come up as better with comparison to other two models.

# Conclusion

Best performed model: **Random Forest**

Decision Tree and Random forest models performed better than the logistic regression. The evaluation metrics reflect the same. If a choice has to be mad between random forest and decision trees, we can go with random forest. The model performed well on both training and test sets. The area under the ROC curve shows that.

Also the **recall** of the Random Forest model is by far the best value.

| Metrics/ Model | Logistic Regression | Decision trees | Random Forests |
|---|---|---|---|
| Precision | 72.27 | 76.41 | 71.14 |
| Recall | 75.94 | 84.67 | 92.15 |
| F Score | 74.06 | 80.33 | 80.29 |
| Accuracy | 67.02 | 74.30 | 71.9 |

The important features are extracted from the best performed model ie: the random forest model. A custom function is developed for the extraction of important features. Below are the important features.

Out[45]:

| | idx | name | score |
|---|---|---|---|
| 1 | 70 | user_product | 0.479572 |
| 5 | 74 | user_activity_count | 0.109220 |
| 2 | 71 | user_category_2 | 0.098883 |
| 4 | 73 | user_session_count | 0.059240 |
| 0 | 69 | price | 0.042779 |
| 3 | 72 | user_mean_spend | 0.041940 |
| 26 | 20 | category_2_enc_smartphone | 0.018303 |
| 9 | 3 | brand_enc_xiaomi | 0.010335 |
| 6 | 0 | brand_enc_samsung | 0.009834 |
| 72 | 66 | hour_bin_enc_3 | 0.009604 |

.-