Black Friday Practice Problem Competition

Institute of Computer Science, University of Tartu
Anastasiia Shalygina
nshal3879@gmail.com

Problem

Background

A retail company "ABC Private Limited" wants to understand the customer purchase behaviour against various products of different categories. They have shared purchase summary of various customers for selected high volume products from the Black Friday event. The data set also contains customer demographics (age, gender, marital status, city type etc.), product details (product id and product category) and a total purchase amount.

Goals

* To study customer purchase behaviour
* To predict the purchase amount of customer against various products.

Data

Training dataset contains 550 068 observations and different kinds of variables either numerical or categorical. Test dataset contains 233 599 instances.

Evaluation

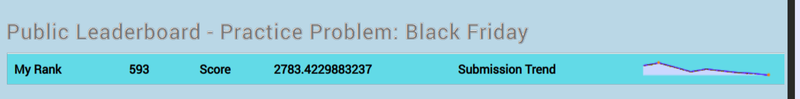
Models performance evaluated on the basis of a prediction of the purchase amount for the test data, which contains similar data-points as train except for their purchase amount. Submissions are scored on the root mean squared error (RMSE).

Prediction

I tried to use 4 different algorithms for this regression problem:

***Random Forest**
***AdaBoost**
***Gradient Boosting**
***XGBoost**

Best Score



Best results on this dataset gives XGBoost regressor. My best score is 2783.42 (RMSE) and currently my rank is 593 out of 13 559 participants.

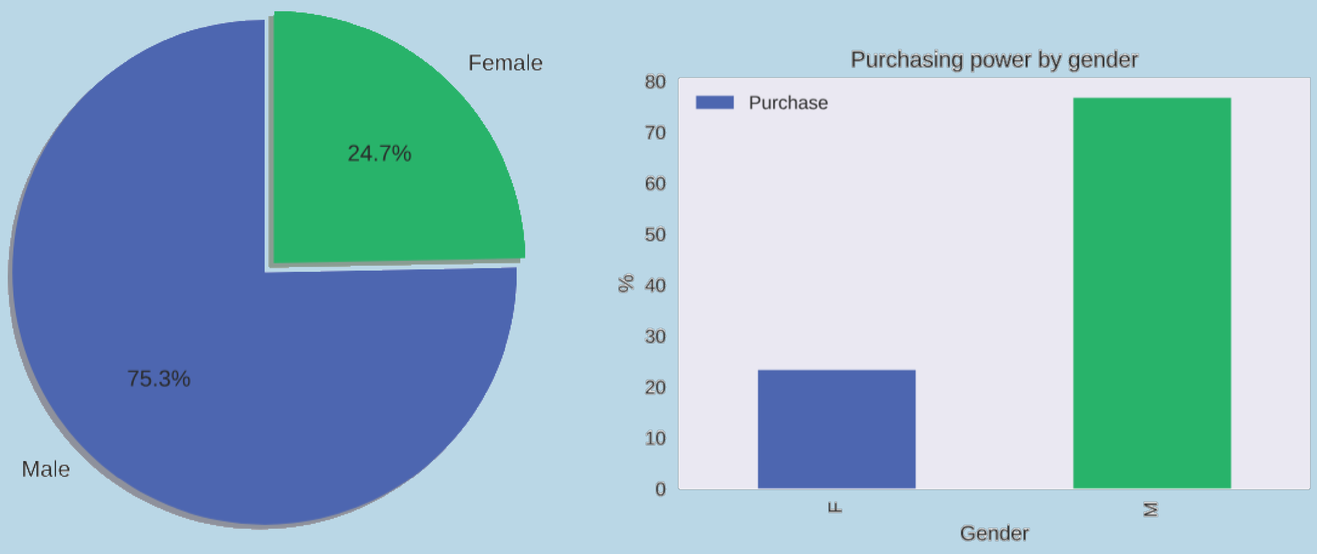
Links

Contest:
https://datahack.analyticsvidhya.com/contest/black-friday/

Github:
https://github.com/ShalyginaA/black-friday-practice-problem

Age, Gender and Purchasing Power

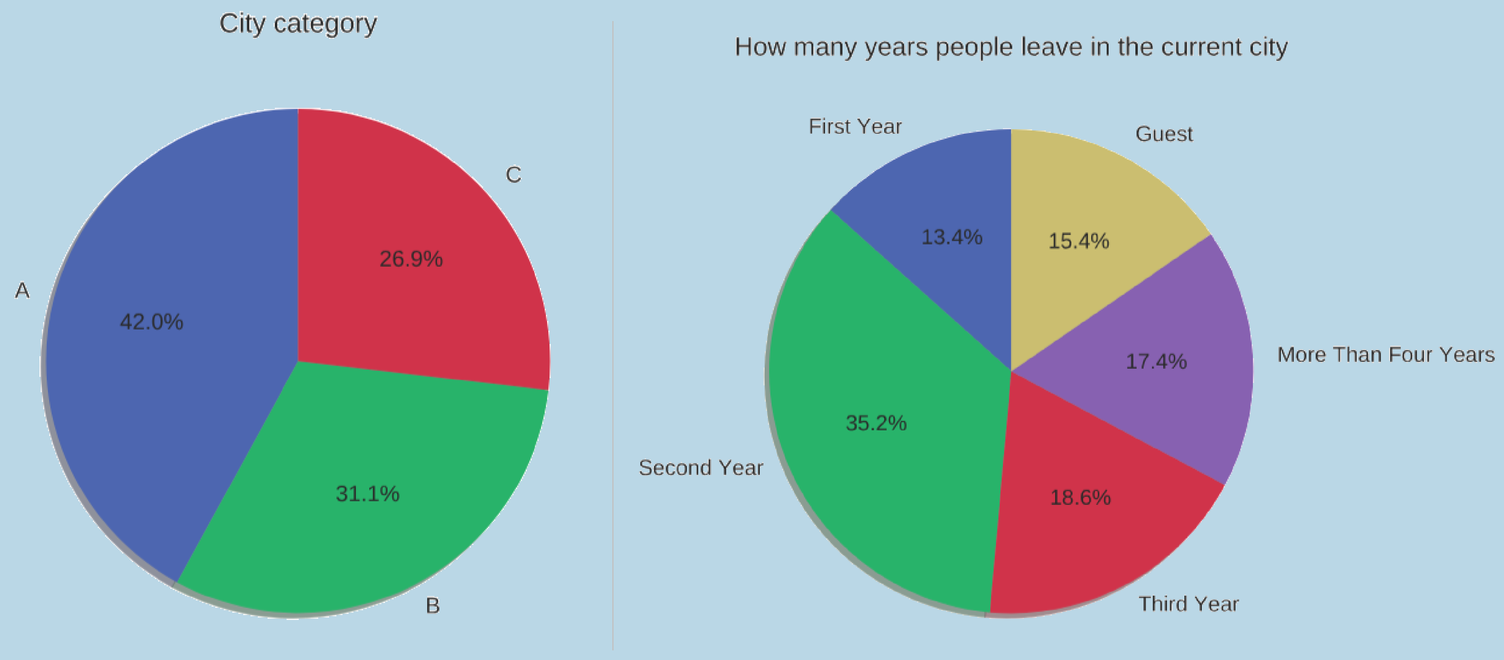
Looking at the plot below we can conclude that female less affected Black Friday sale than male and male purchasing power is bigger.



Also, we can see that most of the purchases are made by people from 25 to 30 years old and people of the same age category are the most frequent visitors of the store.

City Category, Marital Status and Amount of Purchase

The majority of the store visitors live in the cities of category A. From the right chart we can conclude that visitors of the store tends to change their place of living quite often.



Most purchases are made in cities of category B and people from the area B have higher purchasing power than others.