**Introduction :-**

**Topic :- Black Friday Sales prediction**

**Dataset link :-** [**https://www.kaggle.com/c/gb-black-friday-sales/overview**](https://www.kaggle.com/c/gb-black-friday-sales/overview)

**Python file** :-  , **Best performing model** :- XGBOOST

**What is Black Friday ?**

Black Friday is the largest shopping day of the year in America following the Thanksgiving holiday. It is recognized as the ignition of one of the busiest shopping seasons in a year.

**Why have I chosen Black Friday as my assignment for Machine Learning ?**

From the computer science point of view, one of the most interesting applications of machine learning in the retail industry is to effectively predict how much a customer is probably to spend at a store based on historical purchasing patterns. If retailers comprehensively understand their customers in terms of characteristics, behaviours, and motivations in the previous shopping seasons, they can implement and develop more effective marketing strategies for specific customers categories. This study proposes an empirical implementation of extreme gradient boosted trees algorithm for addressing an interesting challenge in the retail industry. From the experimental results, the authors can conclude that the applications of bagging and boosting techniques can achieve great performance and be further improved by a proper combination of models' hyperparameters tuning and feature engineering. ***Considering all the above facts, I have chosen Black Friday as my final topic****.*

**How am I using Black Friday Sales for predicting ?**

This project is analysing the data of Black Friday sales and I have tried to get more insights to answer the key business questions as below. Please refer my python file and visualisation for information :-

* **What is the purchase amount by product categories?**
* **Who are the top buyers of products between age group 18-45 ?**
* **Who are the most buyers from city category B?**
* **What is the purchasing power of users of each category- A,B,C ? Which Category has maximum number of Sales ?**
* **Who are the top buyers in the city ? Are they newcomers or people living in the city from longer time ?**
* **Is the purchasing power dependent on years of staying in current city?**
* **How is purchase variable distributed ?**

**How am I using Black Friday for prediction ? What is the problem statement ?**

Ans:- Using Black Friday dataset ,I wanted to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high-volume products from last month. The data set also contains customer demographics (age, gender, marital status, city\_type, stay\_in\_current\_city), product details (product\_id and product category) and Total purchase\_amount from last month. In this assignment , I am trying to build a model to predict the purchase amount of customer against various products which will help to predict a personalized offer for customers against different products. **Model performance** is evaluated based on your prediction of the purchase amount for the test data (test.csv), which contains similar data-points as train except for their purchase amount. Submissions are scored on the root mean squared error (RMSE).**RMSE** is very common and is suitable general-purpose error metric. Compared to the Mean Absolute Error, RMSE punishes large errors

***I am using RMSE metric for evaluating the model.***

**Data Preparation :-**

In this competition, I am trying to predict the purchase amount in test.csv ,Kaggle Competition (Black Friday Sales) has two datasets: train.csv (contains all features and Purchase) and test.csv (only features).

**Data Preparation or (Data Pre-processing)** is the manipulation of the data, I am performing Data Pre-processing so that further analysis and processing can be carried out further. I am using train data from Kaggle dataset .Below are few of the steps :-

*Step 1*:- Importing required libraries.

*Step 2*:- loading of train and test data ,Finding shape of train and test data.

*Step 3* :- I am trying to list unique products each user has bought.

*Step 4* :- Finding Data types in train.csv, test.csv and description of train data .

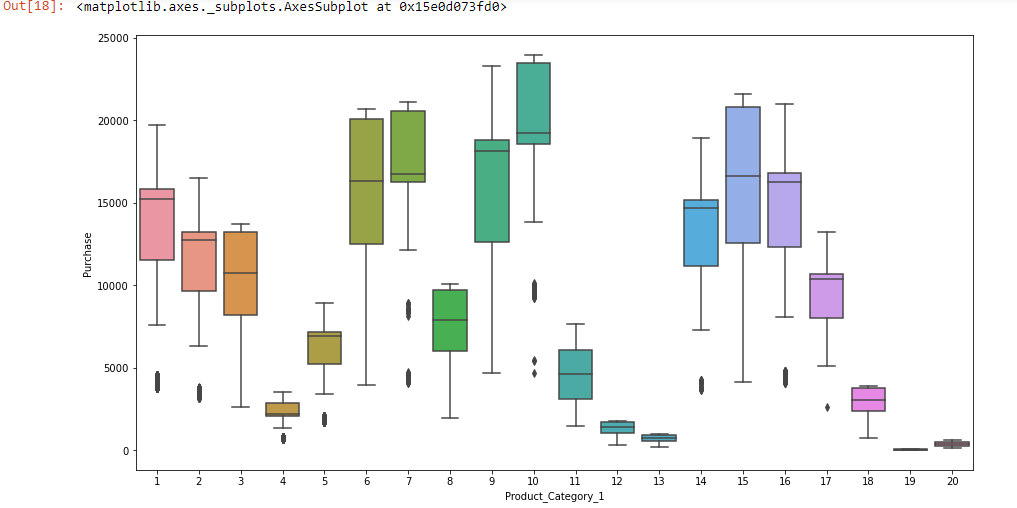
*Step 5* :- Finding of unique elements in Age column (train data)

*Step 6* :- Cleaning of Data (missing values evaluation in both train and test data and removing them)

***Please refer my python file for further questions***.

**Implementation :-**

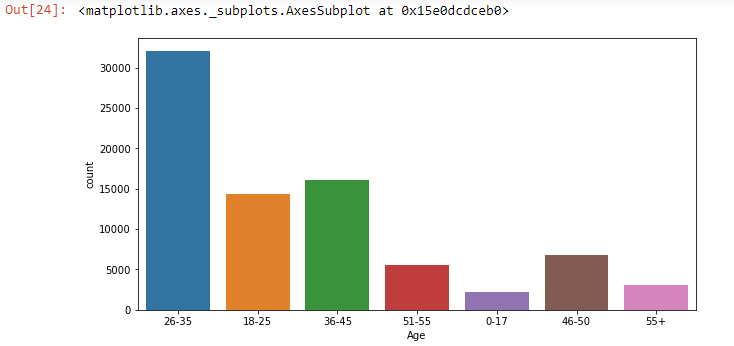
In train data , I have tried to calculate the frequency /Count of each age group which helps me to understand the maximum number of purchases of each product for all ages from (0-55), I noticed that people between age groups 26-35 have a maximum number of purchase count of 32067. See below :-Missing values evaluation in train data :-During Data cleaning process, I have tried to identify missing values in both train and test data , for training data I noticed that Product\_Category\_3 has maximum number of **missing values 69.58625.**

**Visualisation :-** Using Data visualization technique, I have tried to implement interactive visuals and answer series of questions using train data ,I have also tried to display data in a story format that visualizes patterns, trends, and correlations.

**Visualisation 1:- What is the purchase amount by product categories 1,2, 3?**

**For product 2 and 3 visualisation , please refer my python file.**

**Reference : -** [**https://www.machinelearningplus.com/plots/python-boxplot/**](https://www.machinelearningplus.com/plots/python-boxplot/)

**Visualisation 2 :- We can see that most of the buyers are between the age of 26-35**

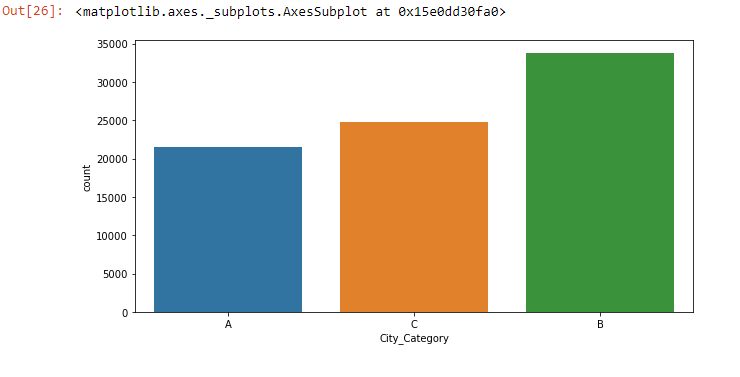
**Reference : -** [**https://stackoverflow.com/questions/46623583/seaborn-countplot-order-categories-by-count**](https://stackoverflow.com/questions/46623583/seaborn-countplot-order-categories-by-count)

**Visualisation 3 :-:-** **Does Age highly impact the purchasing power of user ?**

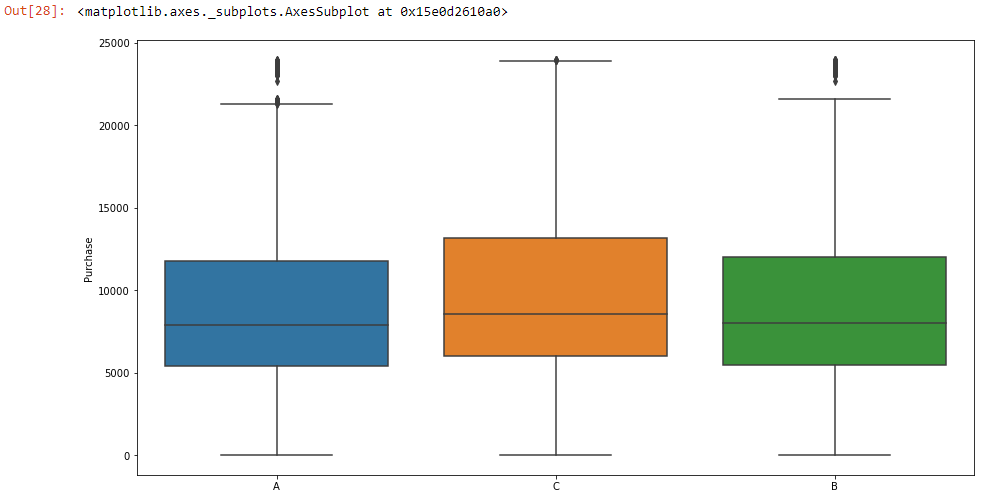
No, it does not , refer my python file for boxplot visualisation .

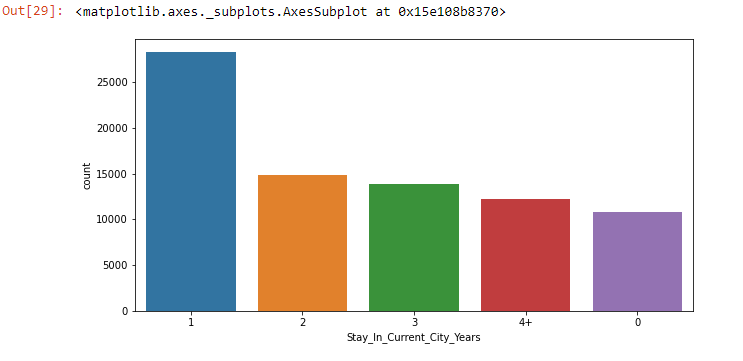


**Visualisation 4** :- Out of city category (A,B,C) ,Most of the buyers are from city category B



**Visualisation 5** :-Purchasing power of users of category C city is highest followed by B and A

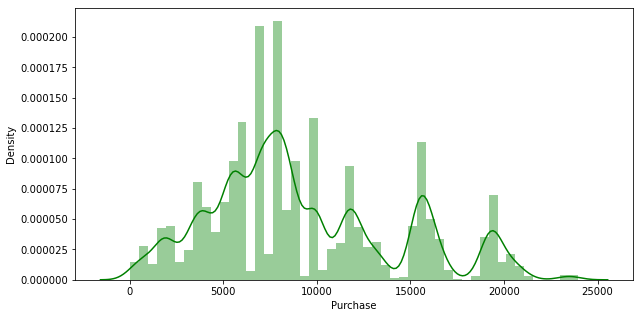




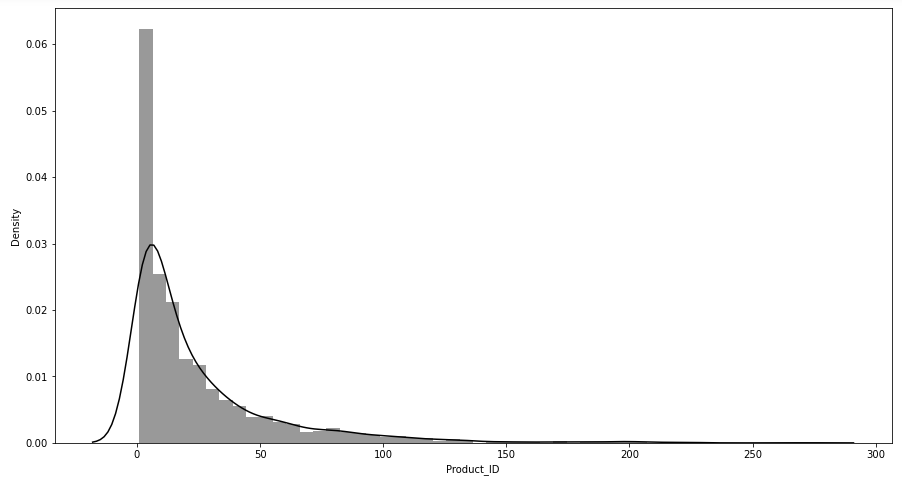
**Visualisations 6** :-Most of the buyers are newcomers in the city probably newcomers want to buy home appliances, furniture’s etc.

We can see that Purchasing power is not much dependent on years of staying in current city ,See my python file for code implementation.

**Reference** : - <https://medium.datadriveninvestor.com/statistics-for-data-science-box-whisker-plots-e5ce7e9ff92a>

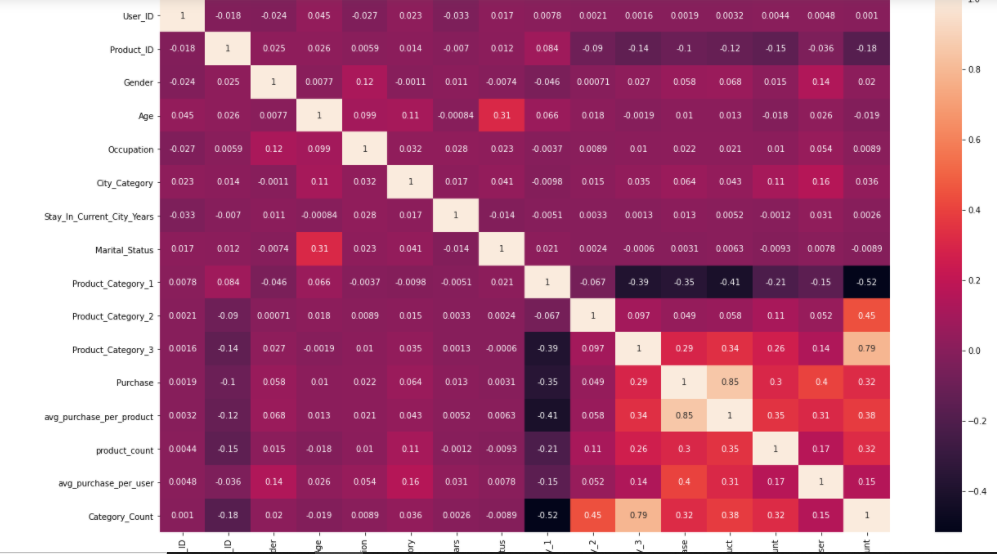
**Visualisations 7**:-Purchase variable of train seems to be normally distributed. See below :-

**Refrence** : - <https://datascience.stackexchange.com/tags/seaborn/hot>

**Visualisations 8 :-** I found that there are around 300 products that are purchased many times and then products purchased rarely . Please refer my python file for syntax.

**Refrence** : -<https://stackoverflow.com/questions/50669200/distribution-plot-in-python>

**Visualisation 9** :- Plotting a diagonal correlation matrix.



I have also tried using different combinations of feature with purchase to get more insight .We can see that males are willing to spend more the females .On further analysis , I also found that teens (i.e 0-17 age) are spending a little bit less than other age categories but not large difference **.Please refer my python file for syntax.**

* By calculating mean and standard deviation of occupation , we can see that occupation has a little bit impact on purchase.
* C Category users are spending higher followed my A and B category users.
* There is almost no difference between amount spent by males and females.

**Label Encoding :-** Using Label Encoding , I am converting labels into numeric forms (machine readable form) .I have performed label encoding of both train and test data , replaced all NAs with zero. See my python code for further explanation . I have also created two new variables :- average purchase by Product\_ID and User\_ID avg\_purchase\_per\_product,avg\_purchase\_per\_user,Category\_Count and Product\_Category\_3 is most important feature.

**Model Evaluation**

A machine learning algorithm is a computer program that has been taught to identify certain patterns. I have used datasets from Kaggle to train my model, cleaned my dataset and later provided it with an algorithm that it can use to think about and learn from these data.  I have used 5 different algorithms to evaluate my model. These are :-

1. LINEAR REGRESSION (**RMSE Value = 2527.2749990446146**)

2. DECISION TREE (**RMSE Value = 3589.629045206134**)

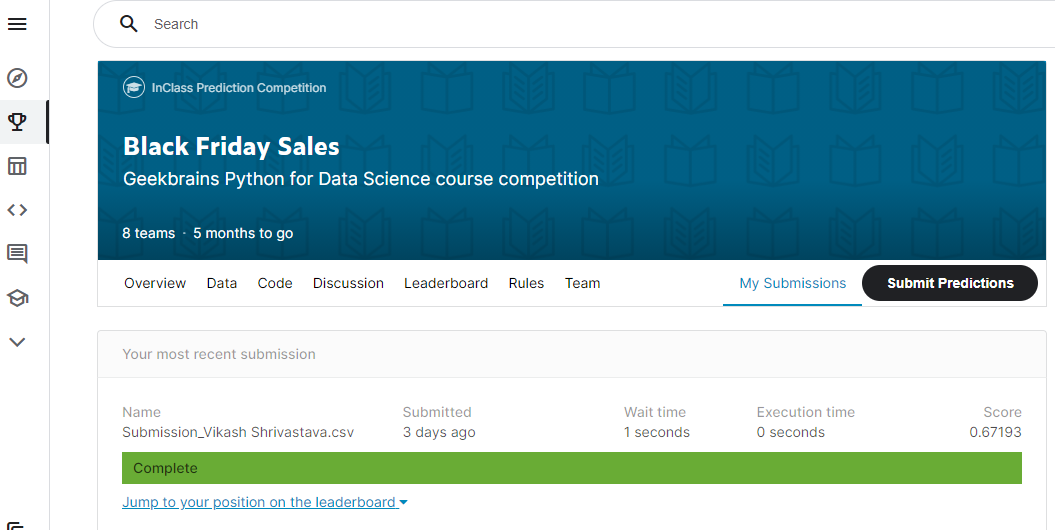
3. RANDOM FORREST (**RMSE Value = 3589.629045206134**)

4. XGBOOST (**RMSE Value = 2609.0666358339413**)

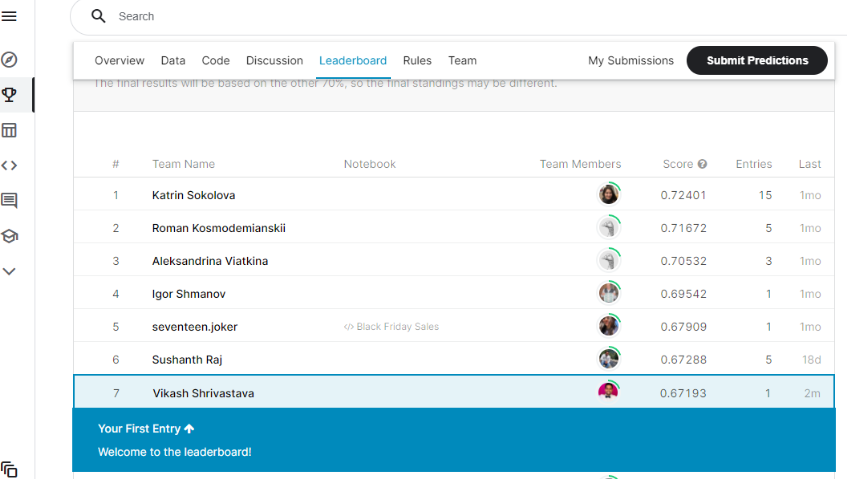
5. NEURAL NETWORK (**RMSE Value = 3589.629045206134**) .Please refer my python file for more information:- **Reference**: <https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d> . Through my above analysis, I notice that **XGBOOST has the best performance** with value **2609.0666358339413** , **hence I am using XGBOOST as my final submission at Kaggle. I am using RMSE metric for evaluating the model.**



**Kaggle Standing**

**My Kaggle submission score- 0.67193**

**I achieved a rank of 7 with score 0.67193 in my first entry.**



**Discussion**

**Question** :- Suggestions for what I might do to continue the work if I had time ?

**Answer** :- If I had more time, I could have worked on model Neural Network .We can see there is No improvement with neural network-based model, but can be improved by tuning hidden layer sizes, Due to infrastructure issue , I was unable to build large complex neural nets

Through my project , I have tried to predict Black Friday Sales . Using these techniques or models , any organisation can be benefited. They can use these models to **predict the purchase amount of customer** against various products which will help **them to create personalized offer for customers against different products.** Please refer my python file for more information. 