**Black Friday Sales Prediction**

## **Table Of Contents**

* [Project Introduction](https://github.com/nanthasnk/Black-Friday-Sales-Prediction#project-introduction)
* [Dataset Description](https://github.com/nanthasnk/Black-Friday-Sales-Prediction#dataset-description)
* [EDA](https://github.com/nanthasnk/Black-Friday-Sales-Prediction#eda)
* [Data Preprocessing](https://github.com/nanthasnk/Black-Friday-Sales-Prediction#data-preparation)
* [Modeling Phase](https://github.com/nanthasnk/Black-Friday-Sales-Prediction#modeling-phase)
* [Evaluation Metric](https://github.com/nanthasnk/Black-Friday-Sales-Prediction#evaluation-metric)
* [Conclusion](https://github.com/nanthasnk/Black-Friday-Sales-Prediction#conclusion)

### **Project Introduction**

Black Friday is an informal name for the Friday following Thanksgiving Day in the United States, which is celebrated on the fourth Thursday of November. The day after Thanksgiving has been regarded as the beginning of the United States Christmas shopping season since 1952, although the term "Black Friday" did not become widely used until more recent decades. Many stores offer highly promoted sales on Black Friday and open very early, such as at midnight, or may even start their sales at some time on Thanksgiving. The major challenge for a Retail store or eCommerce business is to choose product price such that they get maximum profit at the end of the sales. Our project deals with determining the product prices based on the historical retail store sales data. After generating the predictions, our model will help the retail store to decide the price of the products to earn more profits.

### **Dataset Description**

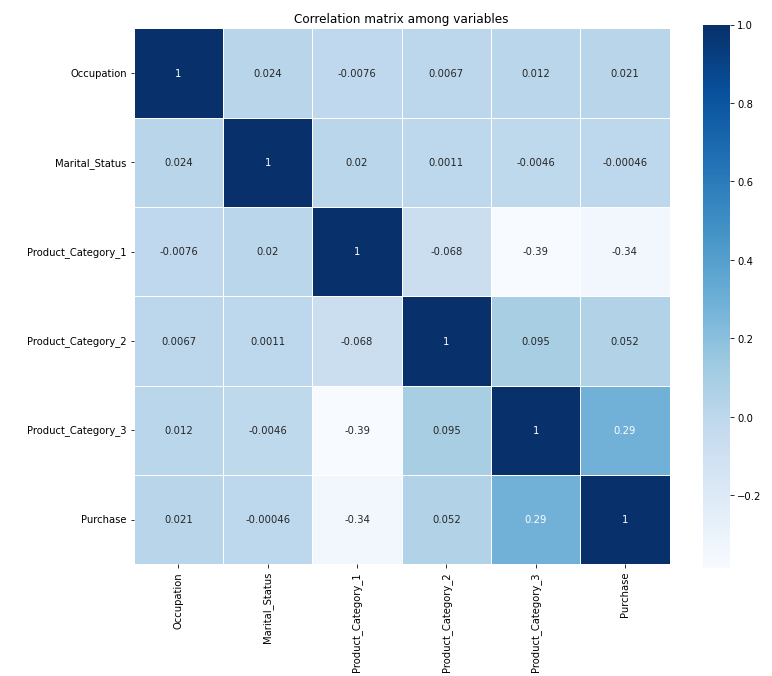
The dataset is acquired from an online data analytics hackathon hosted by Analytics Vidhya. The data contained features like age, gender, marital status, categories of products purchased, city demographics, purchase amount etc. The data consists of 12 columns and 537577 records. Our model will be predicting the purchase amount of the products.

### **EDA:**

Below are the observations which we have made from the data visualization done as part of the Data Understanding process.

* Approximately, 75% of the number of purchases are made by Male users and rest of the 25% is done by female users. This tells us the Male consumers are the major contributors to the number of sales for the retail store. On average the male gender spends more money on purchase contrary to female, and it is possible to also observe this trend by adding the total value of purchase.
* When we combined Purchase and Marital\_Status for analysis, we came to know that Single Men spend the most during the Black Friday. It also tells that Men tend to spend less once they are married. It maybe because of the added responsibilities.
* There are more unmarried customers in the dataset who purchase more during black Friday sales
* For Age feature, we observed the consumers who belong to the age group 25-40 tend to spend the most.
* The Average purchase amount of married and unmarried customers is the same, though unmarried customers purchased more than married customers.
* There is an interesting column Stay\_In\_Current\_City\_Years, after analyzing this column we came to know the people who have spent 1 year in the city tend to spend the most. This is understandable as, people who have spent more than 4 years in the city are generally well settled and are less interested in buying new things as compared to the people new to the city, who tend to buy more.
* The longest customer is living in that city are less prone to buy new things whereas Customers which are new in town are more likely to take advantage of the low prices in Black Friday Sales.
* When examining which city the product was purchased to our surprise, even though the city B is majorly responsible for the overall sales income, but when it comes to the above product, it majorly purchased in the city C.
* Male customers have done more transactions than female.

**Correlation**

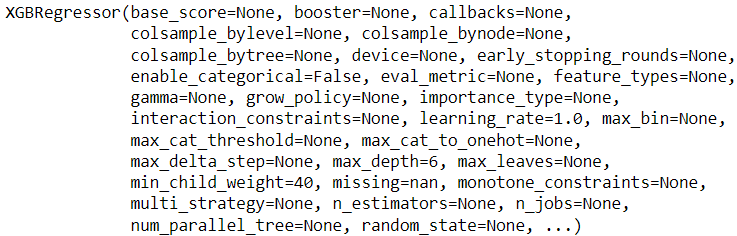


### **Data Preparation**

* Used LabelEncoder for encoding the categorical columns like Age, Gender and City\_Category
* Used get\_dummies form Pandas package for converting categorical variable State\_In\_Current\_Years into dummy/indicator variables.
* Filled the missing values in the Product\_Category\_2 and Product\_Category\_3

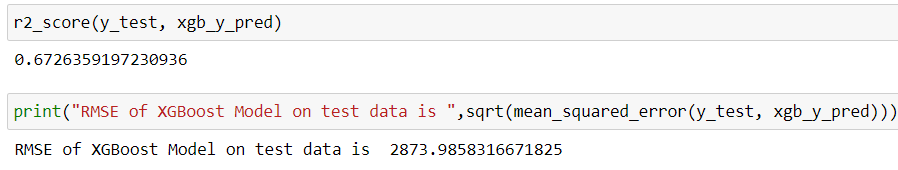
### **Model Phasing**

* Splitted dataset into into random train and test subset of ratio 80:20
* Implemented multiple supervised models such as Linear Regressor, Decision Tree Regressor, Random Forest Regressor.



### **Evaluation Metric**

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. It’s the square root of the average of squared differences between prediction and actual observation.



### **Conclusion**

Implanted multiple supervised models such as Linear Regressor, Decision Tree Regressor, Random Forest Regressor and XGBOOST Regressor. Out of these supervised models, based on the RMSE scores XGBRegressor/XGBOOST Regressor was the best performer with a score of 2873

.