Table Of Contents

- **Project Introduction**
- **Dataset Description**
- **EDA**
- **Data Preprocessing**
- **Modeling Phase**
- **Evaluation Metric**

- # 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.

Dataset:

Column ID	Column Name	Data type	Description	Masked
0	User_ID	int64	Unique Id of customer	False
1	Product_ID	object	Unique Id of product	False
2	Gender	object	Sex of customer	False

Column ID	Column Name	Data type	Description	Masked
3	Age	object	Age of customer	False
4	Occupation	int64	Occupation code of customer	True
5	City_Category	object	City of customer	True
6	Stay_In_Current_City_Years	object	Number of years of stay in city	False
7	Marital_Status	int64	Marital status of customer	False
8	Product_Category_1	int64	Category of product	True
9	Product_Category_2	float64	Category of product	True
10	Product_Category_3	float64	Category of product	True
11	Purchase	int64	Purchase amount	False

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.
- For Age feature, we observed the consumers who belong to the age group 25-40 tend to spend the most.
- 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.
- 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.

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

Modeling Phase

- 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.

```
Random Forest Regression:
RMSE  is:   3051.35541573242
R2 Score : 0.6309821516972987

Decision tree Regression:
RMSE of the Model is  3361.633452177241
R2 Score : 0.5521191505924365

Linear Regression:
RMSE of the Model is  4625.781368526566
R2 score : 0.15192944521481688
```

- In this project, we tried to build a model using various algorithms such as Linear regression, Decision tree regression, Random forest and XGB regressor to get the best possible prediction.
- The hyperparameter tuned random forest regressor gives us the best rmse value and r2 score for this problem.