Black Friday Sale Prediction

Submitted by:

Arpita Rai

Int_33

ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies, for giving me this opportunity to work on this project. I got to learn more from this project about Data Scraping, and practical implementations of using machine learning modules.

I take this opportunity to express my gratitude and regards to my mentor Mr. Shwetank Mishra for his guidance, monitoring and constant encouragement by giving new projects. The help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

Lastly, I thank almighty, my parents, brother, sister and friends for their constant encouragement without which this assignment would not be possible.

INTRODUCTION

- ## Business Problem Framing

  A retail company “ABC Private Limited” wants 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.

- ## Conceptual Background of the Domain Problem

  A sales forecast helps every business make better business decisions. It helps in overall business planning, budgeting, and risk management. Sales forecasting allows companies to efficiently allocate resources for future growth and manage its cash flow. Sales forecasts help sales teams achieve their goals by identifying early warning signals in their sales pipeline and course-correct before it’s too late Sales forecasting also helps businesses to estimate their costs and revenue accurately based on which they are able to predict their short-term and long-term performance.

- ## Review of Literature

  Black Friday has been recognized as the largest shopping day of the year. For most retailers, it is the busiest day of the year. Black Friday is traditionally known for long lines of customers waiting outdoors in cold weather before the open hours. Sales are so high for Black Friday that it has become a crucial day for stores and the economy in general with approximate 30% of all the annual retail sales occurring in the time from Black Friday making it the kick-off day for the busiest and most profitable season for many businesses.

- Motivation for the Problem Undertaken

  In order to compete with Online Shopping Platforms, Brick and Mortar based Retailers need to figure out how to boost Sales during the most important Shopping Day of the Year. By understanding the Purchase Patterns of the Customers Retailers can provide improved Service Quality. Improve Staffing and Inventory of the Retail Store.

  Reveal and understand the most important factors from predictors such as Age, Gender, City of Residence etc., that influence the spending of a customer. Establish a quantitative impact of the revealed factors and how they influence Purchase by a Customer on a personal level i.e., whether they have a positive or negative contribution on the Purchase.

Analytical Problem Framing

- Data Sources and their formats

  This dataset comprises of sales transactions captured at a retail store. This is a regression problem.

  The dataset has 550,069 rows and 12 columns.

  Problem: Predict purchase amount. Data Overview Dataset has 550068 rows (transactions) and 12 columns (features) as described below:

  ➢User_ID: Unique ID of the user.

  ➢Product_ID: Unique ID of the product.

  ➢Gender: indicates the gender of the person making the transaction.
  ➢Age: indicates the age group of the person making the transaction.
  ➢Occupation: shows the occupation of the user, already labeled with numbers 0 to 20.

  ➢City_Category: User's living city category. Cities are categorized into 3 different categories 'A', 'B' and 'C'.

➢Stay_In_Current_City_Years: Indicates how long the users has lived in this city.

➢Marital_Status: is 0 if the user is not married and 1 otherwise.

➢Product_Category_1 to _3: Category of the product. All 3 are already labaled with numbers.

➢Purchase: Purchase amount

- Data Preprocessing Done

Most of the raw data contained in any given Dataset is usually unprocessed, incomplete, and noisy.

In order to be useful for data mining purposes, the Dataset needs to undergo pre-processing, in the form of 'Data Cleaning' and 'Data Transformation'.

Handling Missing Values Handling Outliers Dealing with Categorical Variable In our dataset the only predictors having missing value are Product_Category_1, Product_Category_2 and Purchases.

We can either try to impute the missing values or drop these predictors. We can text both approaches to see which returns the best results.

- Data Inputs- Logic- Output Relationships

➢Dropping Irrelevant Variables

➢For the purpose of data preprocessing we have used the following tools :

• Mean and Mode

• Concatenate

• Standard Scaler

- Hardware and Software Requirements and Tools Used

1. Windows

2. Jupter Notebook

3. Anaconda Navigator

4. Python

Model/s Development and Evaluation

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

Note : Data Scientists have to apply their analytical skills to give findings and conclusions in detailed data analysis written in jupyter notebook . Only data analysis is required. Need not to create machine learning models /but still if anybody comes with it that is welcome.

"So I have not created any model"

- Visualizations

We have joined two data i.e., Train and Test dataset.

```
df=df1.append(df2,sort=False)
df.shape
```

```
C:\Users\Arpita\AppData\Local\Temp\ipykernel_3620\4229194643.py:1: FutureWarning: The frame.append method is deprecated and wil
l be removed from pandas in a future version. Use pandas.concat instead.
  df=df1.append(df2,sort=False)
```

```
(783667, 12)
```

```
df.head()
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	Product_Category_2	Product
0	1000001	P00069042	F	0-17	10	A	2	0	3	NaN	
1	1000001	P00248942	F	0-17	10	A	2	0	1	6.0	
2	1000001	P00087842	F	0-17	10	A	2	0	12	NaN	
3	1000001	P00085442	F	0-17	10	A	2	0	12	14.0	
4	1000002	P00285442	M	55+	16	C	4+	0	8	NaN	

Checking for null values

```
df.isnull().sum()
```

```
Product_ID                        0
Gender                            0
Age                               0
Occupation                        0
City_Category                     0
Stay_In_Current_City_Years        0
Marital_Status                    0
Product_Category_1                0
Product_Category_2           245982
Product_Category_3           545809
Purchase                     233599
B                                 0
C                                 0
dtype: int64
```

Filling NAN values with mean and mode

```
df['cat2']=df['cat2'].fillna(df['cat2'].mode()[0])
df['cat3']=df['cat3'].fillna(df['cat3'].mode()[0])
df['Purchase']=df['Purchase'].fillna(df['Purchase'].mean())
```

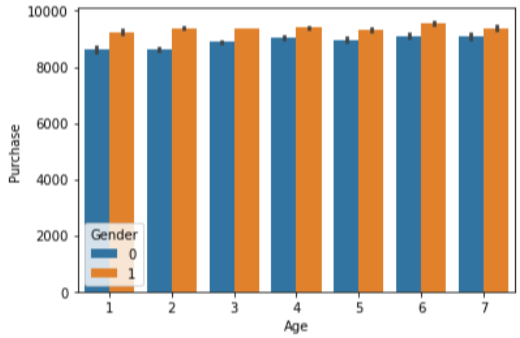
```
# Again checking for null values
df.isnull().sum()
```

```
Product_ID                    0
Gender                        0
Age                           0
Occupation                    0
City_Category                 0
Stay_In_Current_City_Years    0
Marital_Status                0
Product_Category_1            0
cat2                          0
cat3                          0
Purchase                      0
B                             0
C                             0
dtype: int64
```

```
#Age v/s Purchase
sns.barplot('Age','Purchase',hue='Gender',data=df_i)
```

```
C:\Users\Arpita\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as k
rgs: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an
t keyword will result in an error or misinterpretation.
  warnings.warn(
```

```
<AxesSubplot:xlabel='Age', ylabel='Purchase'>
```

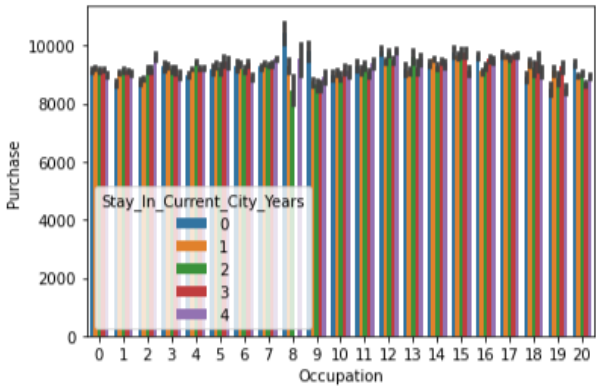


Purchase of goods at each range of age are almost equal. We can conclude that the percentage of purchasing goods of men over women is higher.

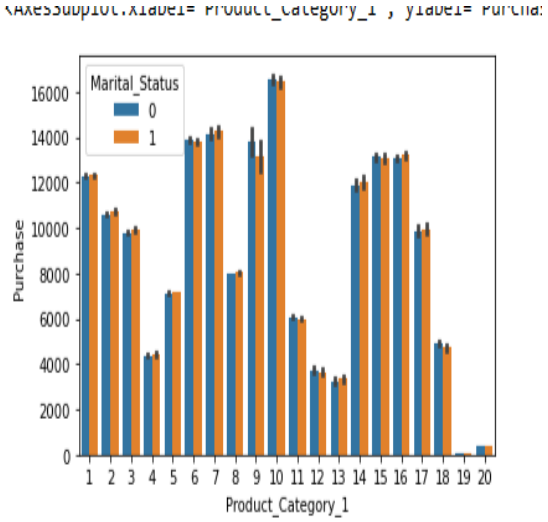

```
#Occupation v/s Purchase
sns.barplot('Occupation','Purchase',hue='Stay_In_Current_City_Years',data=df_i)
```

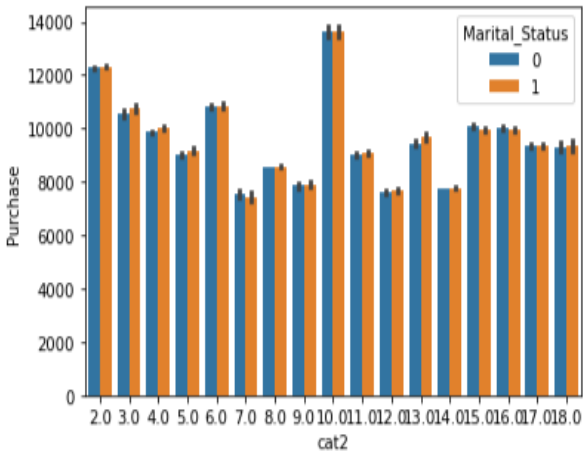
```
C:\Users\Arpita\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the follow
rgs: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other ar
t keyword will result in an error or misinterpretation.
  warnings.warn(
```

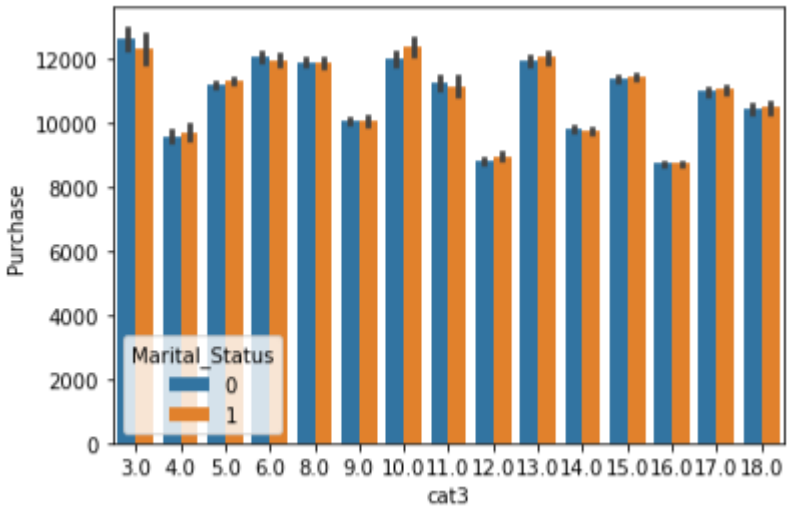
```
<AxesSubplot:xlabel='Occupation', ylabel='Purchase'>
```



All the occupation contributes almost same in purchasing rates and it won't affect alot that how many years you live in a city.







CONCLUSION

- *The count of Male gender is higher as compared to the female.*
- *The female gender is slightly higher compared to the male gender when compared to marital status category.*
- *Higher purchases have been done by the male gender as compared to the female.*
- *Occupation has a direct effect on the purchases done by the customer*
- *The female gender in the occupation 18 with higher purchases compared to others.*
- Now we have features for both training and testing.

- The data can now be converted to a data frame, if necessary
- It can be fed to a machine learning model.

```
X_train
```

```
array([[ 0.57275431, -0.36745197,  0.6008837 , ...,  0.36937114,
         1.17365495, -0.67228678],
       [ 0.57275431, -0.36745197, -1.23913919, ...,  0.36937114,
        -0.85203918, -0.67228678],
       [ 0.57275431,  1.10995723, -0.16579251, ...,  0.36937114,
         1.17365495, -0.67228678],
       ...,
       [ 0.57275431,  1.84866184,  1.67423038, ...,  0.36937114,
        -0.85203918,  1.48746045],
       [ 0.57275431, -1.10615657, -0.93246871, ...,  0.36937114,
        -0.85203918, -0.67228678],
       [ 0.57275431, -0.36745197, -1.23913919, ...,  0.36937114,
        -0.85203918,  1.48746045]])
```

```
X_test
```

```
array([[ 0.57275431, -0.36745197, -0.62579823, ...,  0.36937114,
        -0.85203918, -0.67228678],
       [-1.74594931, -1.10615657, -0.62579823, ...,  0.36937114,
         1.17365495, -0.67228678],
       [ 0.57275431, -1.10615657, -0.62579823, ...,  0.36937114,
        -0.85203918, -0.67228678],
       ...,
       [ 0.57275431, -1.10615657,  0.90755418, ..., -3.64065155,
         1.17365495, -0.67228678],
       [ 0.57275431, -1.10615657,  0.29421322, ...,  0.36937114,
        -0.85203918,  1.48746045],
       [-1.74594931,  1.10995723,  0.6008837 , ...,  0.36937114,
        -0.85203918,  1.48746045]])
```

- Limitations of this work and Scope for Future Work

Forecasting is not always intended to be a realistic projection of anticipated sales and not a depiction of desired sales. The challenge for company marketing and sales reps in preparing forecasts is that internal bias is hard to avoid