Aegis School of Data Science,

Mumbai

Project Report II

Black Friday Sales Prediction (Regression)

Submitted By :

Nihar Shah

Pravesh Raikwar

(DF-2009-CM)

Abstract

In this project, we experiment with real world dataset for classification. We explore few machine learning classification algorithms to fit the data. We were expected to gain experience using scikit ML library and how different algorithms works over a specific data. We have to explore the dataset using EDA processes then fit multiple models and tune the hyper parameters to get the maximum accuracy. A retail company wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. After performing required tasks, herein lies our final project report.

1.Introduction :

We have selected regression challenge that is hosted on www.analyticsvidhya.com. In this challenge, we have to predict the purchase amount of customer against various products which will help to create personalized offer for customers against different products. In order to do this, complete information about 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

2.Problem Definition

The task is to predict the purchase amount of customer against various products for black Friday sale. In order to do this, complete information about the data set 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 is given.

Now, we have to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products. The output of a prediction is a continuous variable.

Official website : https://datahack.analyticsvidhya.com/contest/black-friday/

3.Exploratory Data Analysis

3.1 Descriptive Analysis :

Dataset csv files : train.csv
We have 550068 data points consisting 12 features.

Data Description :

Variable	Definition
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins
Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

train.csv contain columns as follows :

```
#    Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   User_ID                     550068 non-null  int64
 1   Product_ID                  550068 non-null  object
 2   Gender                      550068 non-null  object
 3   Age                         550068 non-null  object
 4   Occupation                  550068 non-null  int64
 5   City_Category               550068 non-null  object
 6   Stay_In_Current_City_Years  550068 non-null  object
 7   Marital_Status              550068 non-null  int64
 8   Product_Category_1          550068 non-null  int64
 9   Product_Category_2          376430 non-null  float64
 10  Product_Category_3          166821 non-null  float64
 11  Purchase                    550068 non-null  int64
```

Missing Values :

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
User_ID                            0
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            173638
Product_Category_3            383247
Purchase                           0
dtype: int64
```

Here, we can see that, Product_Category_2, Product_Category_3 has many missing values i.e. 173638, 383247 respectively. We impute it using a constant.

Description of data :

In [5]:
```
df.describe()
```

Out[5]:

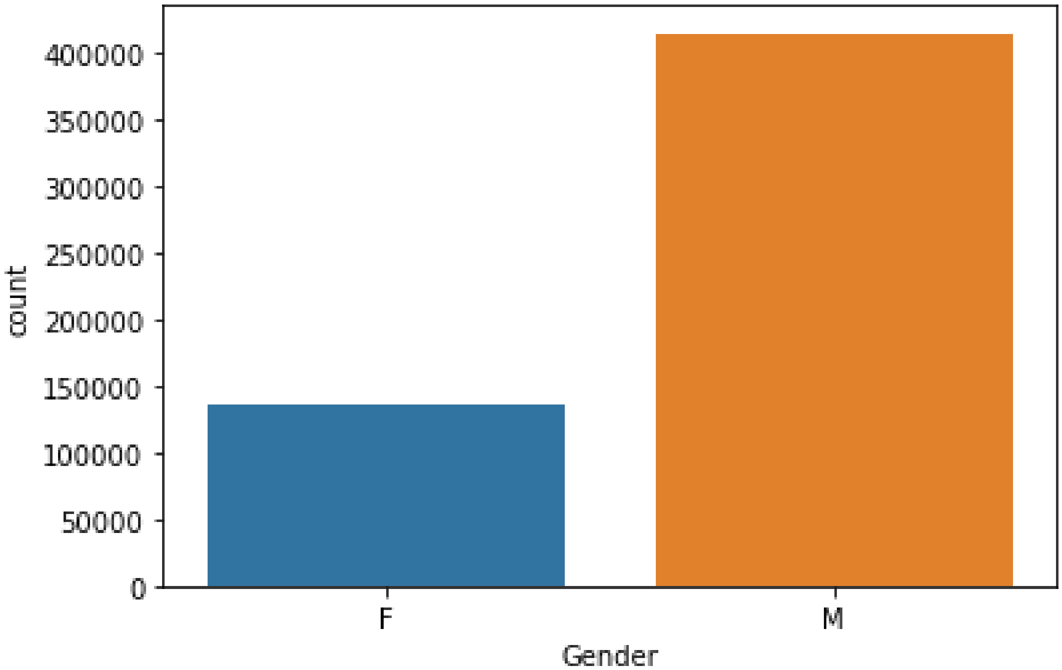
	User_ID	Occupation	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase
count	5.500680e+05	550068.000000	550068.000000	550068.000000	376430.000000	166821.000000	550068.000000
mean	1.003029e+06	8.076707	0.409653	5.404270	9.842329	12.668243	9263.968713
std	1.727592e+03	6.522660	0.491770	3.936211	5.086590	4.125338	5023.065394
min	1.000001e+06	0.000000	0.000000	1.000000	2.000000	3.000000	12.000000
25%	1.001516e+06	2.000000	0.000000	1.000000	5.000000	9.000000	5823.000000
50%	1.003077e+06	7.000000	0.000000	5.000000	9.000000	14.000000	8047.000000
75%	1.004478e+06	14.000000	1.000000	8.000000	15.000000	16.000000	12054.000000
max	1.006040e+06	20.000000	1.000000	20.000000	18.000000	18.000000	23961.000000

Skewness in data:

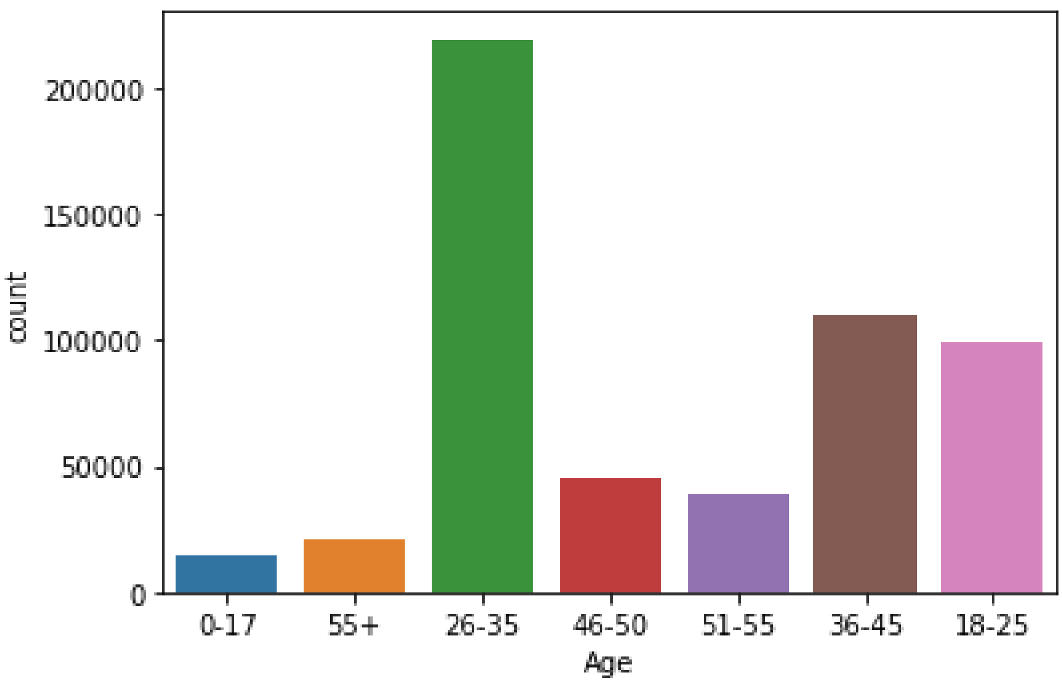
```
User_ID               0.003066
Occupation            0.400140
Marital_Status        0.367437
Product_Category_1    1.025735
Product_Category_2   -0.162758
Product_Category_3   -0.765446
Purchase              0.600140
```

Product_Category_1 has the positive skewness. All other columns are not so skewed.

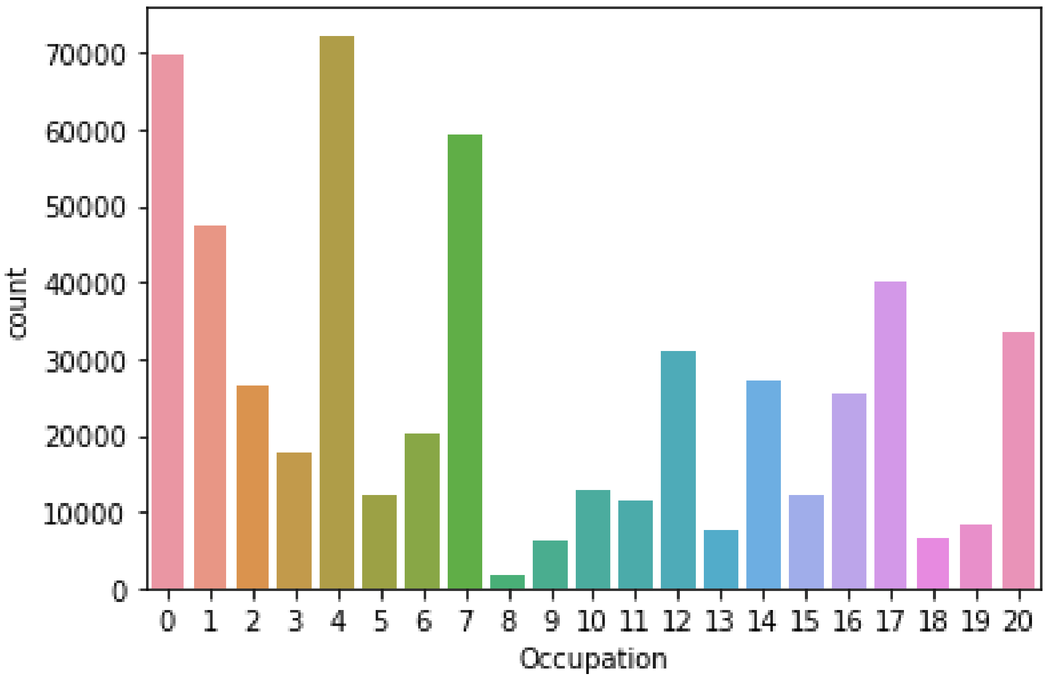
3.2 Data Visualizations:



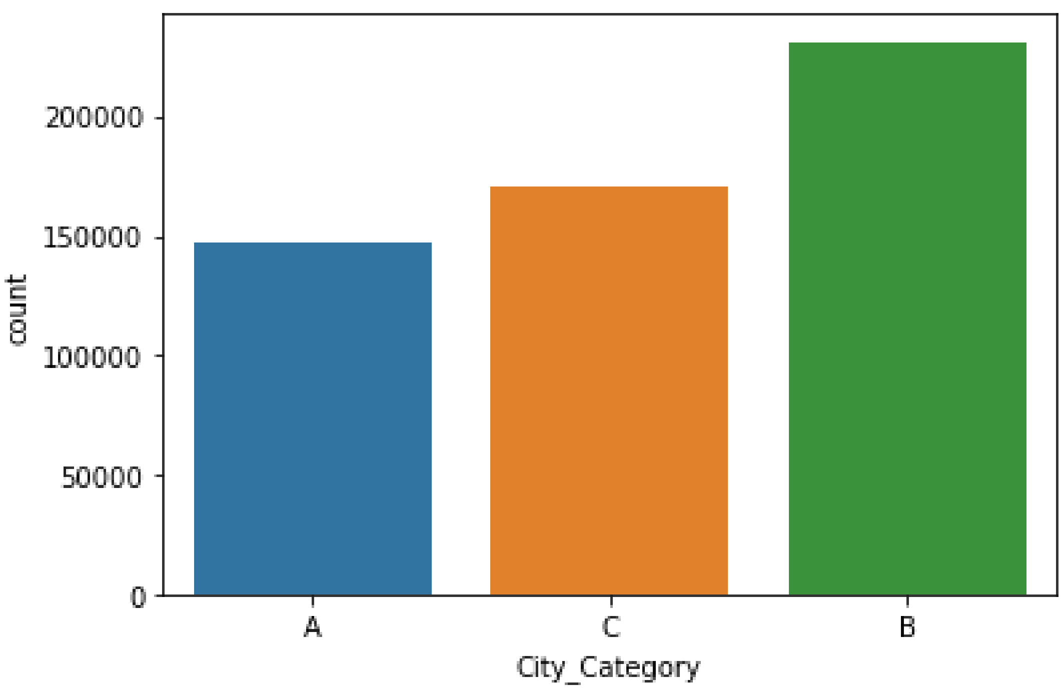
Male customer are more in our dataset compare to female.



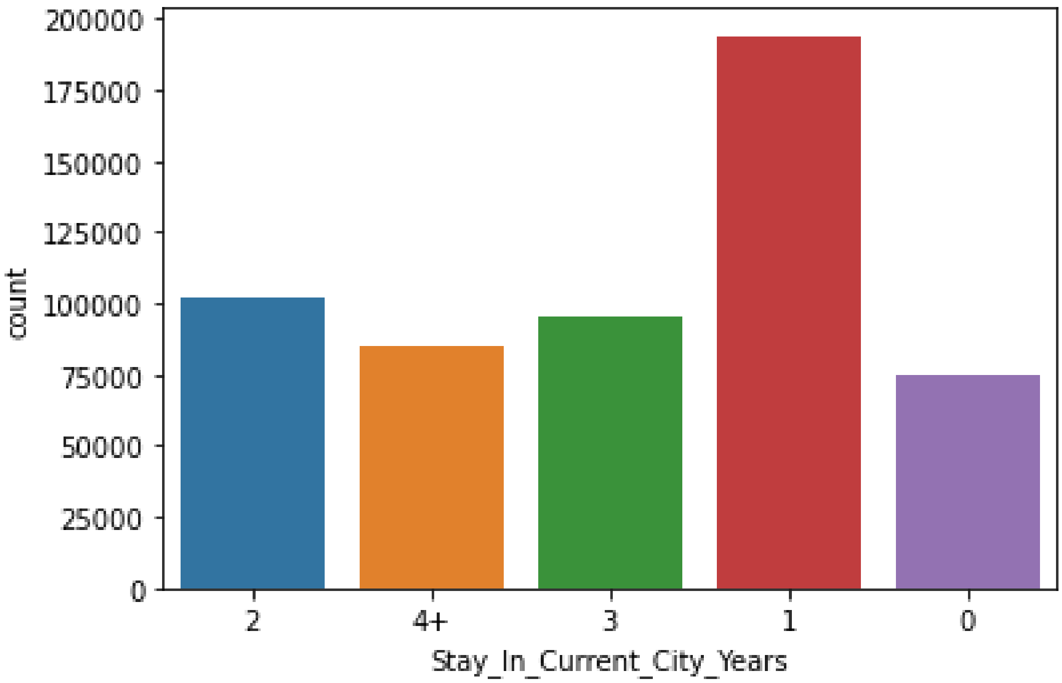
More people aged 26-35years are purchasing more.



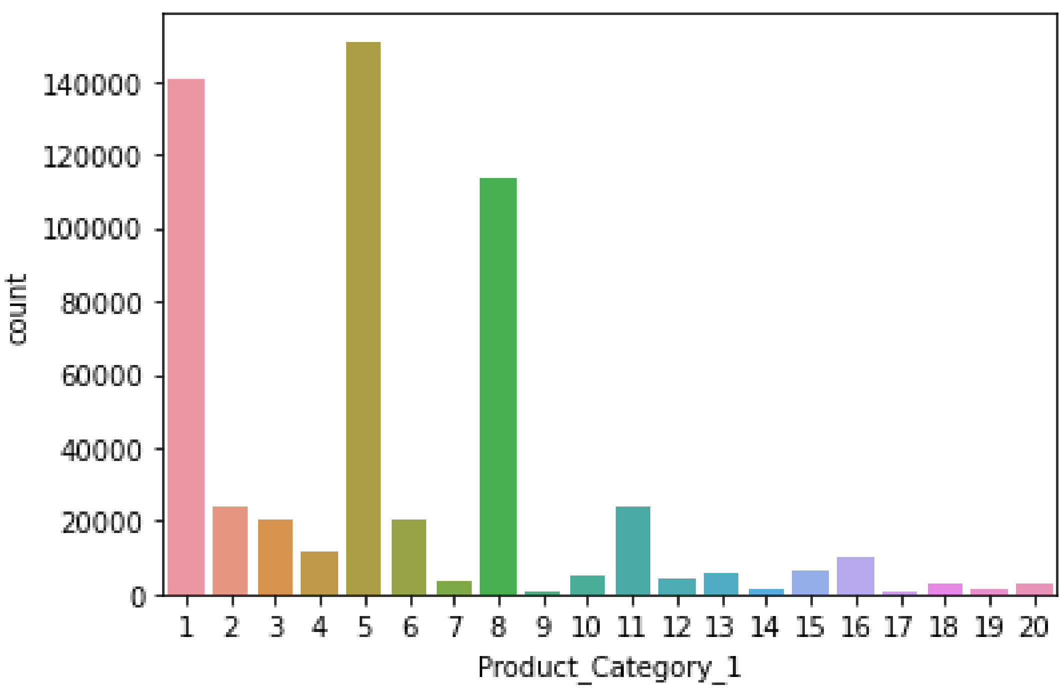
People having occupation masked [0,4,7,1] are usually purchase more during sales.



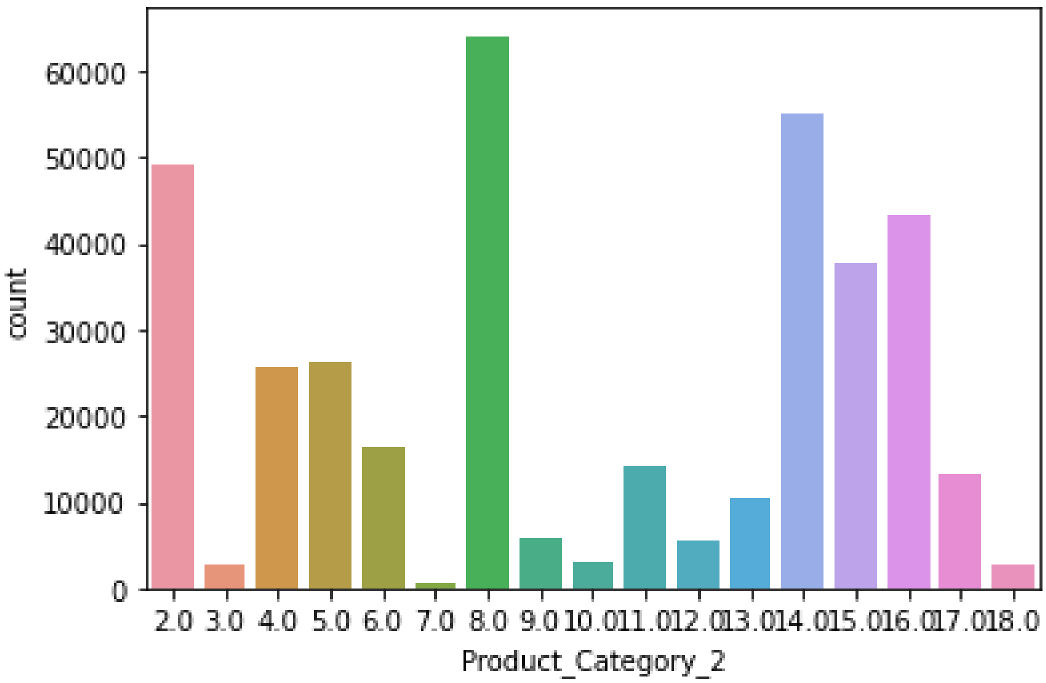
We have more people from city B who has max numbers in dataset.



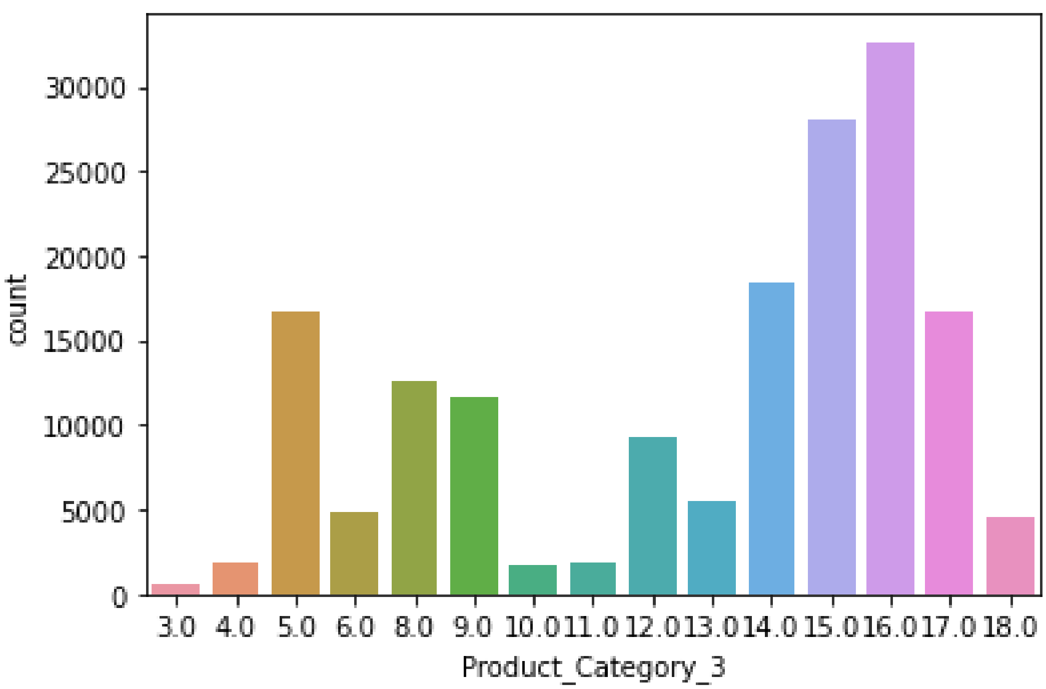
People who are residency in city for 1 year are tent to purchase more.



Product category_1's products are more often purchased. List of these type of products are 1,5,8 (masked).

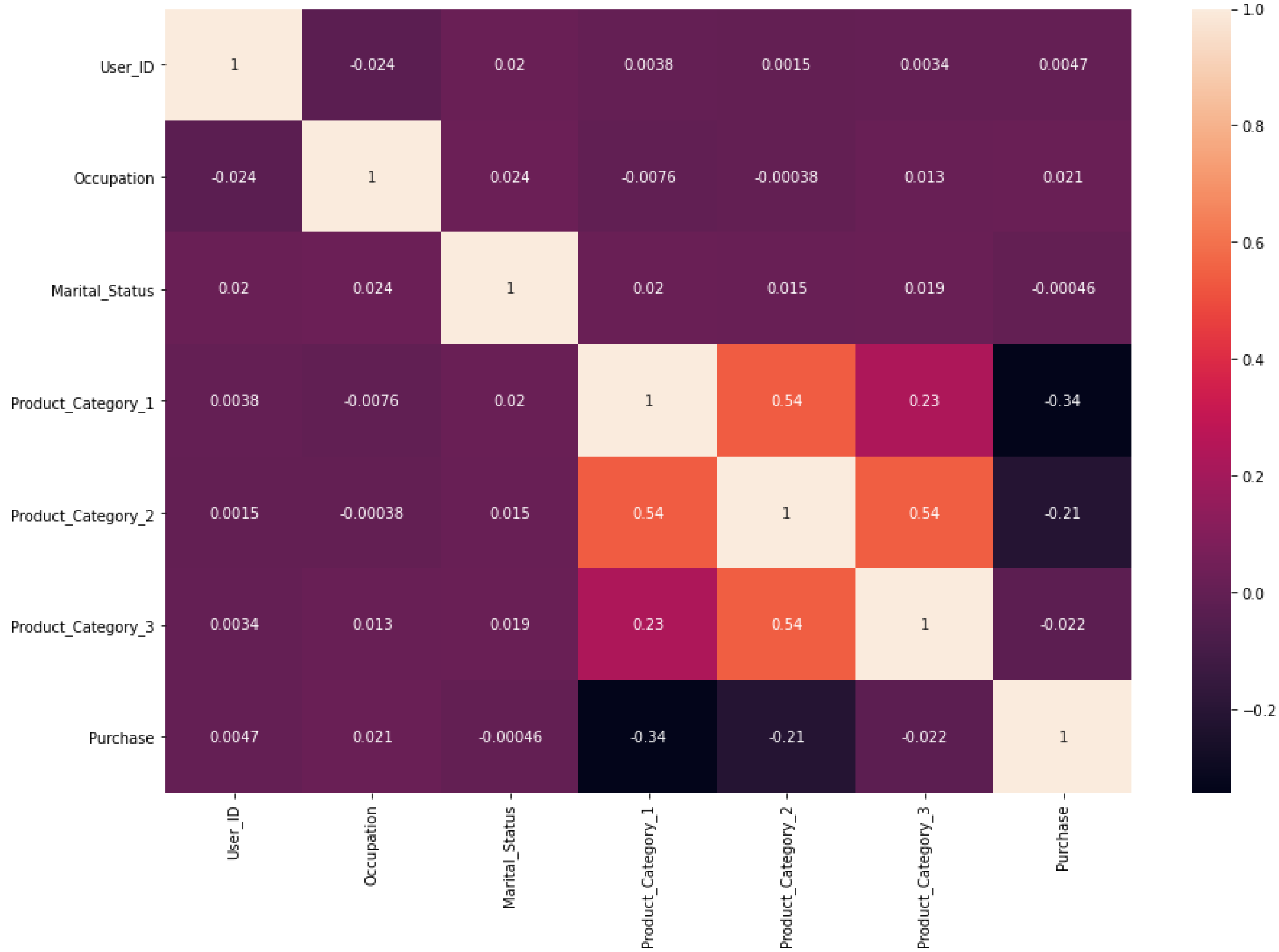


Product category_2's products are more often purchased. List of these type of products are 2, 8, 14, 15, 16(masked).



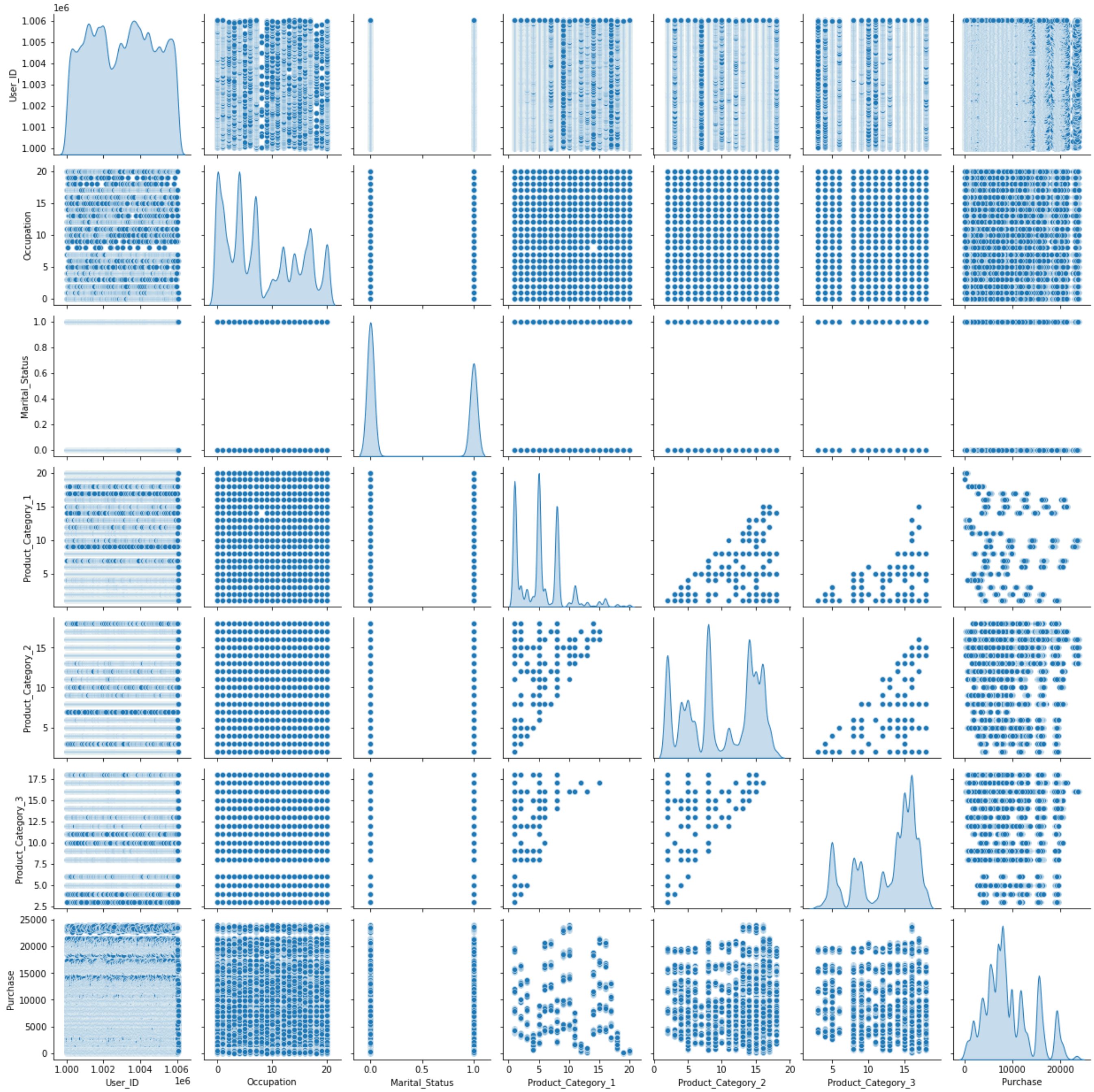
Product category_3's products are more often purchased. List of these type of products are 5, 14, 15, 16 (masked).

Heatmap :



There is no clear interpretation of feature that are linearly corelated. But we can infer via lightly shaded area, that there is some collinearity in the dataset.

Pairplot :



Diagonally we can see the distribution of the features.

4.Evaluate Algorithms

4. 1 **Linear Regression :**

Score :
model_1_LinearRegression.score(X_test, y_test)
0.14497179163973384

4.2 **Decision Tree Regressor:**

Scores :
model_2_DecisionTreeRegressor.score(X_test,y_test)
0. 5673416953318338

4.3 XGboost:

Models parameters :
{ objective ='reg:linear', n_estimators = 25, seed = 10}

Scores:
model_3_XGBRegressor.score(X_test,y_test)
0. 6638326837907499

4.4 Random Forest Regressor:

Model's Best paramters :
{ max_features='sqrt', n_estimators=500, n_jobs=-1, warm_start=True}

Scores :
0.6381038302184262