FLIP ROBO

BLACK FRIDAY PROJECT

Submitted by:

Pooja Rajpal

Internship Batch 33

Acknowledgement

I would like to express my appreciation to my institute

(Data Trained Education)

And to my respected teacher(Shankar sir) for his teachings and notes he provided to the students.I would like to thank my institute which is always available to answer our queries.

This project has been source of learning and bring our theoretical knowledge into real time projects .

I would really acknowledge my teacher’s help and guidance.

Introduction

Business problem statement:

It is a regression problem where we have to predict the sale of an product based on various aspects .The aim is to know the sale of a particular product.In this retail statement a retail company wants to understand the purchase of a particular product.This dataset comprises of customers demographics,details of the product and the purchase amount of the last month.

Background of the problem:

To understand the project it is necessary to understand what the project is about .This project is about a retail company who wants to understand the behaviour of the customer purchase against various products.This project is about the retail industry who needs to effectively predict how much a customer is probably to spend at a particular store .if retailers are able to understand the behaviour of their customers,they can develop the marketing strategies in a more effective manner.

Review of literature:

Black Friday is the Christmas shopping festival across US.Customers are offered discounts and deals on various products .The categories of the products varies from electronic items to kitchen appliances,from clothing to décor.

Research is carried out on the analysis and the prediction of sales.models such as Linear Regression,Decision tree with bagging are used.RMSE(Root Mean Squared Error)isw used to evaluate the models .This dataset consists of 550068 rows and 12 columns .This dataset consists of various attributes such as user id,product id,gender ,age , occupation,city_category,marital_status etc.

Motivation behind the problem:

Building project will improve my skills and make things better.Building successful projects gives the

feeling of pride .And for that I need to build more and more projects passionately and try to do it

with more accuracy.

However completing the project requires complex steps.Things get really difficult at some stages.but

we need to adjust our vision .And try once again.Also its my passion to build a model with high

accuracy and will definitely try to learn more n more for that.

Analytic Problem Framing

- ***Mathematical/ Analytical Modeling of the Problem:***

  The first step here is to import necessary libraries.And load the dataset in jupyter notebook to analyse it.

  Statistical Modelling is necessary because it summarizes the results can be observed by the evaluators .It is relationship between variables.

  Checking for nulls is another step in this dataset and impute them using imputers to clean the data.

  Statistical techniques such as mean,median ,standard deviation ,interquartile ranges .These are simple and we can start it as starting point for EDA.

The goal is to collect the data and make predictions about real world.

Two types statistics (Descriptive and inferential) Eg(weather forecasting ,stock market,loan approvaland fraud detection ,housingetc)

Mean ,median and mode are the measures of central tendency and is descriptive statistics.

Sampling of data and infer the results to describe entire population.

Label encoder to encode the data because computer only understands numeric data .

- **Data Sources and their formats**

A data source is the location where data that is being used originates from.A data source may be a database,a flatfile,scrapped web data.etc.

The data collected here is raw data which is not useful ,cleaning of raw data and utilizing the data for further analysis.Data collection is the process of collecting data whether in structured format or unstructured format.

There are three types of data sources

1. Relational

2. Multidimensional
3. Dimensionally modelled relational

   Sources of data collection:

   Primary and Secondry

   Primary data that is collected through interviews and surveys.

   Secondry data includes internal data (organisations)and external data(govt agencies).

This data is internal data in the form of CSV file given to us by our organization which need to be cleaned to make a model .

- Data Preprocessing Done

Step1:collection of data and load the dataset in jupyter notebook using pandas and for that we need to import pandas library .

Step2:data cleaning that include checking for nulls and treat them using imputers.

Step3:checking for datatypes and if object datatype convert them into integer as computers only understand numeric data

Step 4:data visualisation

Step 5:Countplot to visualize the data.

Step 6:plotting distplot for columns to check for outliers and treating them using outlier detection

Step 7:checking for skewness and treat them using Power transformer if there is skewness

Step 7:when the data is cleaned we will check for collinearity using heatmap .

At the end split the dataset into features and label.

- **Data Inputs- Logic- Output Relationships**:

An input is whatever you used to predict the output.

There can be negative or can be positive relationship between input and output. In this dataset columns such as user id , product id, gender, age , occupation,marital_status are input data which are used to predict output variable purchase.

All the input variables have more or less relationship with output variable.

- Hardware and Software Requirements and Tools Used

  Laptop and

  Anaconda navigator is desktop graphical user interface that allows to launch applications

  Jupyter notebook is open source software.

1. Pandas library is used to import the dataset
2. Matplotlib and seaborn to plot for visualizing the dataset.
3. From sklearn.preprocessing import LabelEncoder to encode the data.
4. For train and test importing train_test_split from sklearn.model_selection.

Model/s Development and Evaluation

- **Identification of possible problem-solving approaches (methods)**

Statistical approach is a method for removing the biasness from evaluating the data by employing numeric analysis.Mean,Median,Mode are used to get the central value for the dataset.Here we have used descriptive statistical analysis to solve the problem.

- **Testing of Identified Approaches (Algorithms)**

As it is a regression problem linear regression, random forest regression can be good for this type of dataset.

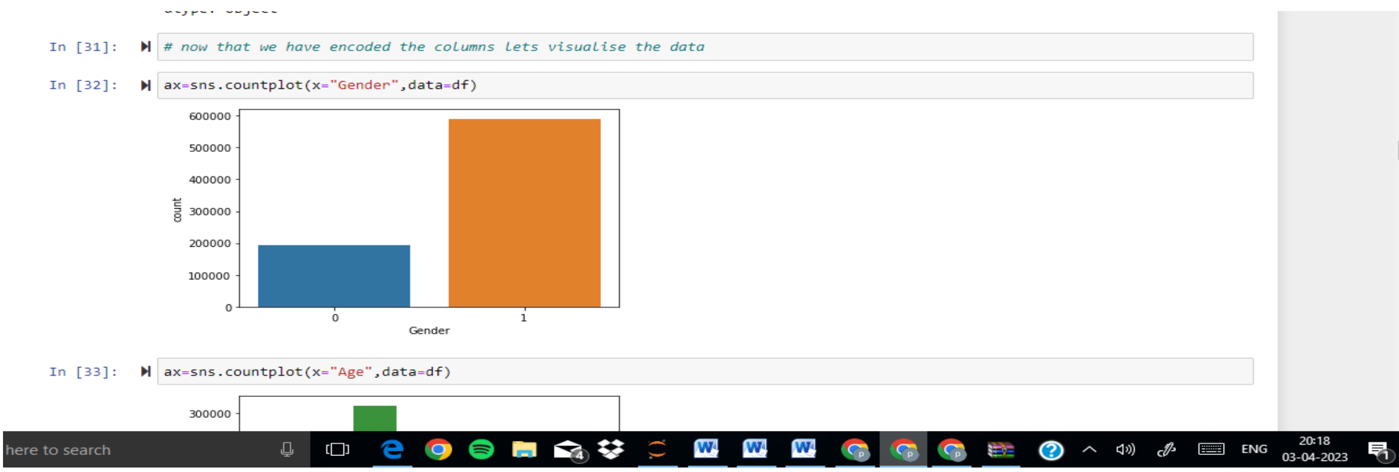
- ***Key Metrics for success in solving problem under consideration***

Metrics are used to measure the performance of the model,Key metrics (MSE,MAE) good for this dataset.

Visualizations

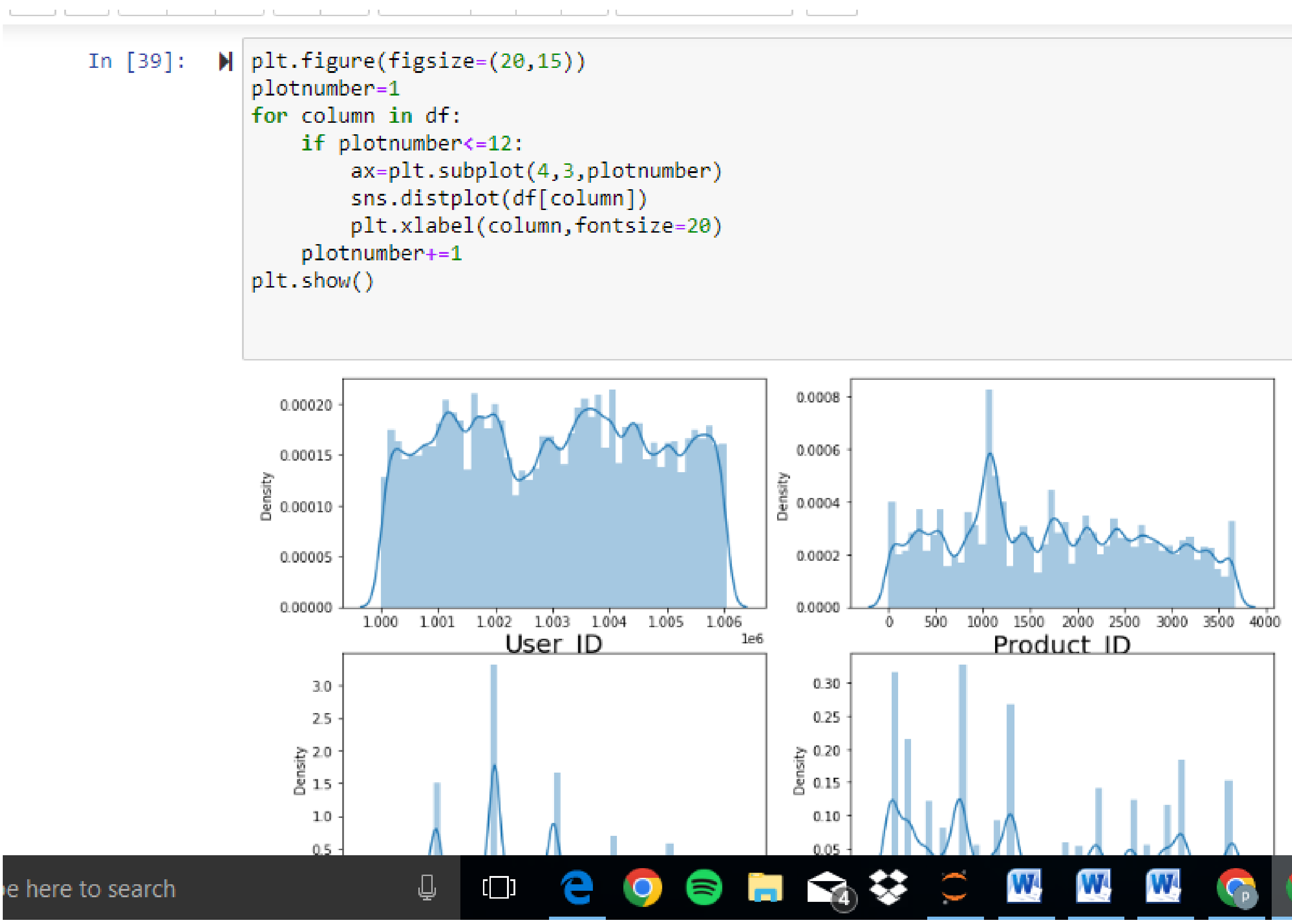
Importing necessary libraries for visualising the data is necessary .

Matplotlib,seaborn are the two necessary libraries for visualizing the data.count plot to visualize categorical data.



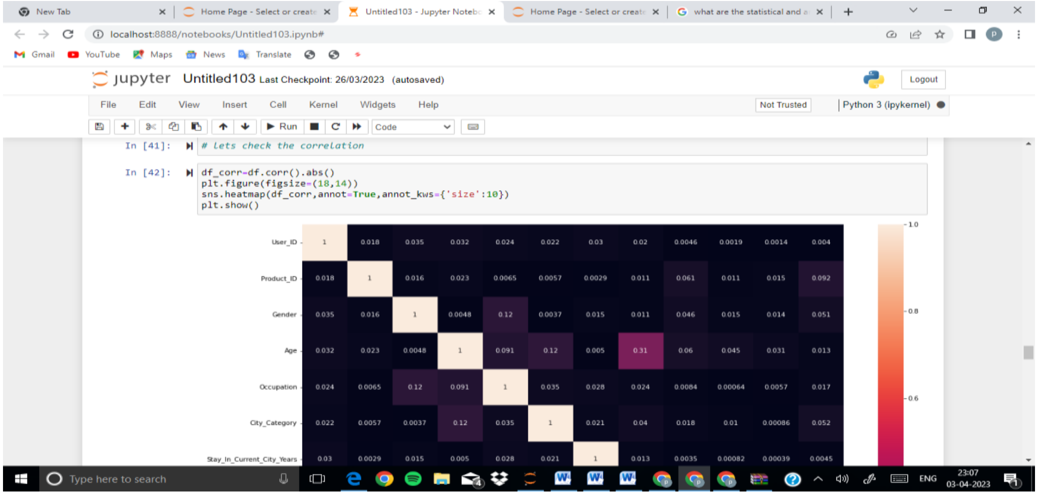
```
In [31]:  # now that we have encoded the columns lets visualise the data

In [32]:  ax=sns.countplot(x="Gender",data=df)

In [33]:  ax=sns.countplot(x="Age",data=df)
```

Distplot to check for skewness



```
In [39]:  plt.figure(figsize=(20,15))
          plotnumber=1
          for column in df:
              if plotnumber<=12:
                  ax=plt.subplot(4,3,plotnumber)
                  sns.distplot(df[column])
                  plt.xlabel(column,fontsize=20)
              plotnumber+=1
          plt.show()
```

Heat map to check multicollinearity



New Tab
Home Page - Select or create
Untitled103 - Jupyter Notebo
Home Page - Select or create
what are the statistical and a
localhost:8888/notebooks/Untitled103.ipynb#
Gmail
YouTube
Maps
News
Translate
jupyter
Untitled103 Last Checkpoint: 26/03/2023 (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
In [41]:  # lets check the correlation
```

```
In [42]:  df_corr=df.corr().abs()
          plt.figure(figsize=(18,14))
          sns.heatmap(df_corr,annot=True,annot_kws={'size':10})
          plt.show()
```

User_ID	1	0.018	0.035	0.032	0.024	0.022	0.03	0.02	0.0046	0.0019	0.0014	0.004
Product_ID	0.018	1	0.016	0.023	0.0065	0.0057	0.0029	0.011	0.061	0.011	0.015	0.092
Gender	0.035	0.016	1	0.0048	0.12	0.0037	0.015	0.011	0.046	0.015	0.014	0.051
Age	0.032	0.023	0.0048	1	0.091	0.12	0.005	0.31	0.06	0.045	0.031	0.013
Occupation	0.024	0.0065	0.12	0.091	1	0.035	0.028	0.024	0.0084	0.00064	0.0057	0.017
City_Category	0.022	0.0057	0.0037	0.12	0.035	1	0.021	0.04	0.018	0.01	0.00086	0.052
Stay_In_Current_City_Years	0.03	0.0029	0.015	0.005	0.028	0.021	1	0.013	0.0035	0.00082	0.00039	0.0045

- **Interpretation of the Results**

  .Preprocessing is necessary to clean the data Visualization because it is easily understood by us to observe the data in graphical form.

conclusion

Data cleaning and Exploratory Data Analysis to analyse the data.After that checking the correlation between the features because it can affect the accuracy of the model .

Hyperparameter tuning is necessary to increase the accuracy.

Then building model and checking the accuracy.

Learning Outcomes of the Study in respect of Data Science

I faced lot of challenges in making this project

Also learnt lot of things .Found difficulty in choosing the right metrics.

In regression problem the output variable must be continuous and in classification it must be descrete.

Limitations of this work and Scope for Future Work

I worked harder but want to do more and more practice to be perfect in every project.