Machine Learning Engineer Nanodegree
Capstone Proposal
Joe Udacity
December 31st, 2050
****

Proposal
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 a selected high volume products from last month.
They want 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.
Before looking at the data it is important to understand how does the company expect to use and benefit from this model? This first brainstorming helps to determine how to frame the problem, what algorithms to select and measure the performance of each one.

* black Monday was the sell-off the day before the stock market crash of 1929
*Black Friday Touted as the biggest shopping event of the year, it can get a little overwhelming when trying to find the best bargains.,
While a relatively new concept in South Africa, it's been a big deal in the United States for years when retailers took advantage of the day after the traditional Thanksgiving Day holiday. It is often characterized by amazing deals and sometimes crazy behaviour by bargain hunters.
*Finally we want to predict the Purchase by getting some features of the product .
The dataset is suitable for applying a classification problem to predict which segment “range of age” more effective on future sales/marketing technique to maximize their profits .

*this is research made with the a similar data set with other features ,with model accuracy = 42%. Through our model we would get a better result
https://medium.com/diogo-menezes-borges/project-3-analytics-vidhya-hackaton-black-friday-f6c6bf3da86f

****

Domain Background
This is a multivariate regression problem ,"as a supervised problem " that need a prediction of quantity with real valued or discrete input variables by an input data ,
****

Problem Statement
*the data set is a sample of transactions in store ,our target is to know the customer Purchase Behaviour according to different products .
then predict the dependent variable " purchase value" according to other features of the product ”new/old”.

This will help for more effective future sales/marketing technique to maximize their profits .

****

Datasets and Inputs
*I have Dataset of 550 000 observations about the black Friday in a retail store ,i get it through "kaggle" website https://www.kaggle.com/mehdidag/black-friday/version/1?login=true
*the data set contain several features that help us predict the price of the product like the gender of the customer as women may buy more than men so * Age and Gender: Men with ages ranging from 25 to 40 should spend more on techlogical products.
the store would increase products for them than men and this would appear through our analysis .
* the martial status data would help us to know which category increase our sale as Families should be more contained on spendings , just buying the best offers and only needed products. so we increase the suitable products for them .and so on other features as appear through our analysis
* and through analysis we would know which product category increase the sale ,so the next year the store would increase them and apply the suitable care for this egment of customers .
* Purchase History: Customer with a purchase history should be more willing to purchase more products on this day.
* City category would affect as Urban or Tier 1 cities should have higher sales because of the higher income levels of people there.

*the data set contains 12 column with different kinds of variables either numerical or categorical .
*features :
1-User_ID
2-project -ID
3-Gender : female-male
4- Age : with three cat " 0-17 / 26-35/ 36-45 /46-50 / 46-50 / 51-55 / +55
5-Occupation :
6-City_Category
7-Stay_In_Current_City_Years : contains four values " 0 / 1 /2 /3 / +4
8-Marital_Status :contains two values: "1" mean married / "0"mean single
9-Product_Category_1
10--Product_Category_2
11-Product_Category_3
12-Purchase

```
In [1]: import pandas as pd
        import os
        print(os.listdir("../Black Friday"))
        dataset=pd.read_csv('../Black Friday/BlackFriday.csv')
        dataset.info()
        print(dataset.head())
```

```
['.ipynb_checkpoints', 'Black Friday 1.ipynb', 'BlackFriday.csv', 'visuals.py', '__pycache__']
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 537577 entries, 0 to 537576
Data columns (total 12 columns):
User_ID                     537569 non-null float64
Product_ID                  537577 non-null object
Gender                      537577 non-null object
Age                         537577 non-null object
Occupation                  537576 non-null float64
City_Category               537577 non-null object
Stay_In_Current_City_Years  537577 non-null object
Marital_Status              537577 non-null int64
Product_Category_1          537575 non-null float64
Product_Category_2          370591 non-null float64
Product_Category_3          164278 non-null float64
Purchase                    537577 non-null int64
dtypes: float64(5), int64(2), object(5)
```

4-preprocessing

```
print(dataset.isnull().sum())
```

```
User_ID                          8
Product_ID                       0
Gender                           0
Age                              0
Occupation                       1
City_Category                    0
Stay_In_Current_City_Years       0
Marital_Status                   0
Product_Category_1               2
Product_Category_2          166986
Product_Category_3          373299
Purchase                         0
dtype: int64
```

****

Solution Statement

*1-first we upload our data and apply some exploration technique

2-splitting our data to features and target variables

3-cleaning our data by normalizing our data :

-applying feature selection

(as some of inputs data don’t affect our prediction like "user-id \ Product_iD)

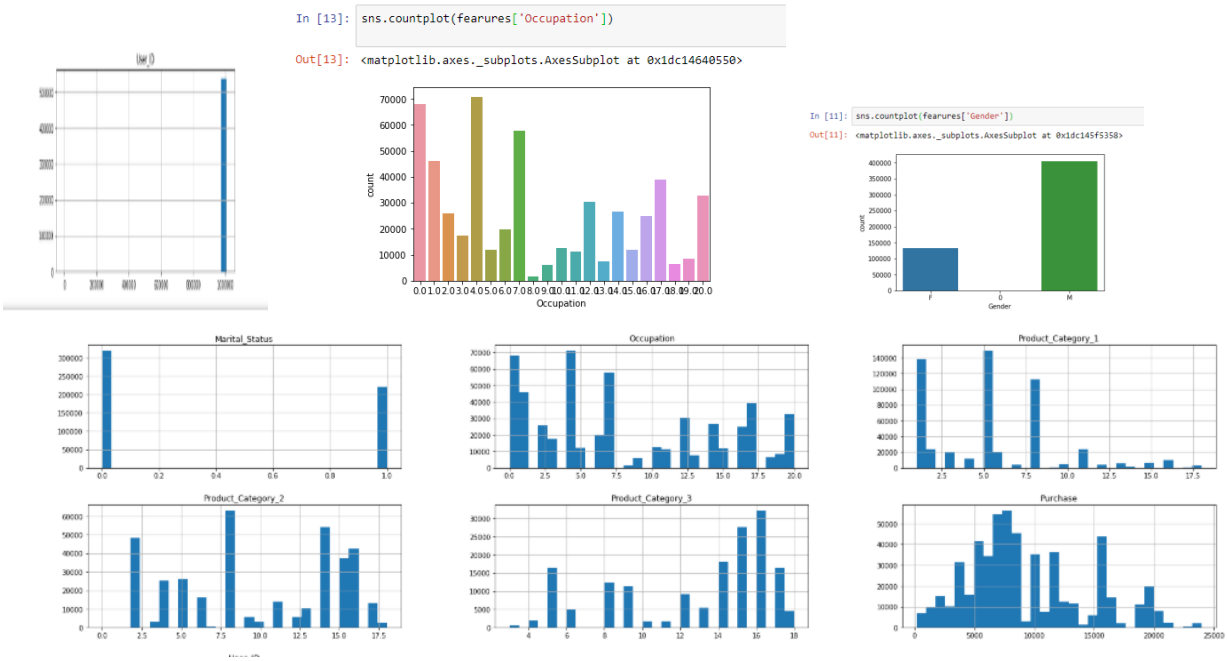
-solving the missing data .

-applying one-hot-encoding

4-setting our prediction model .using random forest model

5-model tuning “grid search”

5-evaluate our model using root mean squared error (RMSE)



This grapgh show features ditripution according to purches

****

Benchmark Model

*This is a similar data set applied linear regression with a similar data set and get accuracy with 42% ,
https://medium.com/diogo-menezes-borges/project-3-analytics-vidhya-hackaton-black-friday-f6c6bf3da86f

*through our model we would increase our accuracy to 80% and apply another structure Using decision tree and SVM/random forest and cross validation technique

****

Evaluation Metrics

As a regression problem I will use RMSE (Root Mean Square Error)
It represents the sample standard deviation of the differences between predicted values and observed values (called residuals). Mathematically, it is calculated using this formula:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

****

###Project Design:
In this final section, summarize a theoretical workflow for approaching a solution given the problem.
Provide thorough discussion for what strategies you may consider employing,
what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

*first we apply some analysis like (max-min-median……)
*then clean our data by delating some columns .
*we apply the prediction model and applying the grid search
* I will apply a random forest model .
Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges
we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate
* It is effective in high dimensional spaces., It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
parameters:
* kernel: ,, kernel trick, which allows us to sidestep a lot of expensive calculations., we have various options available with kernel like, "linear", "rbf","poly" and others (default value is "rbf"). Here "rbf" and "poly" are useful for non-linear hyper-plane.
gamma: Kernel coefficient for 'rbf', 'poly' and 'sigmoid'. Higher the value of gamma, will try to exact fit the as per training data set i.e. generalization error and cause over-fitting problem.
C: Penalty parameter C of the error term. It also controls the trade off between smooth decision boundary and classifying the training points correctly.

We can categorize our Machine Learning (ML) system as:
Supervised Learning task: we are given labeled training data (e.g. we already know how much a customer spent on a specific product);
Regression task: our algorithm is expected to predict the purchase amount a client is expected to spend on this day.
Plain batch learning: since there is no continuous flow of data coming into our system, there is no particular need to adjust to changing data rapidly, and the data is small enough to fit in memory, so plain batch learning should work.

****

Resources:
https://www.kaggle.com/dalalmanish/black-friday-python

https://www.freelancer.com/community/articles/algorithms-machine-learning-engineers-need-to-know

https://medium.com/diogo-menezes-borges/project-3-analytics-vidhya-hackaton-black-friday-f6c6bf3da86f