**Project Report**

**Exploratory Data Analysis and Price Prediction**

**of Black Friday Sales Data**

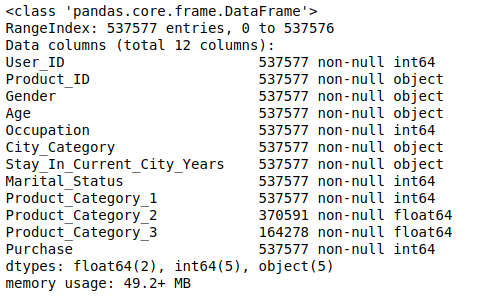
**Course Title: Data Science Course Code: CS-481**

**Name : Pawan Kumar Submitted To:**

**Roll No : 16K3965 Dr. Zeeshan Ahmed**

**Section : GR1**

**Introduction:**

The main idea of this project was to predict the purchase field with available features. Black Friday data set was taken from the website kaggle. Data set contains 537577 entries or rows and 12 features that is columns, almost half a million data set. Data set format is CSV file. By using pandas information of data set gather, and pandas also describes the numerical values with a correlation matrix. Initial data set contains null values and the columns are Product\_Category\_2 and Product\_Catergory\_3. Data science process is followed and shown in the project as well.

**Data Science Process:**

1. **Setting Research Goal**

The Idea behind this project is to understand the purchasing behavior of customers on Black Friday.This project can be utilized to increase Stores sales by looking into insights of generated Transactions.Final Goal is to predict the purchase

amount of customer by looking at age and other features present in the data Set.

**2.** **Retrieving Data**

The data set is download from kaggle website in CSV format. <https://www.kaggle.com/mehdidag/black-friday>

**3.** **Data Preparation**

Pythons pandas library is used for creating data frames of

csv file. All operations are applied on data frame for useful

Results. Missing values are analyzed by functions is null.

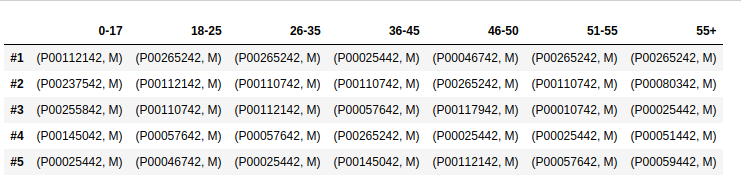
As all the variables were written correct no numerical variable present in the data set so need of changing them.

Data cleansing performed on two variables Product\_Category\_2 and Product\_Category\_3 by replacing null values with zero. After EDA more variables required cleansing. Categorical variables Gender, Stay\_In\_Current\_City\_Years and City\_Category replaced with integers for example female with 0 and male with 1 etc.

**4.** **Exploratory Data Analysis (EDA)**

EDA is performed deep insights of data set. Prediction variable is Purchase every other feature relation shown with purchase. User\_ID gender relation not gives much information. [1,3] It results that 71.72 percent were male buyers and 28.28 percent female buyers. Unmarried customers buy more than married. New residents at least one year in the town are more willing to buy on black friday than Old Residents. 26-35 Age group was highest buyer. That all answers our first and third use case [1,3]. User\_ID grouped with 9 relatable features, but that also not provide clue about which final variables to use for prediction.Same process applied for Product\_ID, but unfortunately that also not performed as desired.Encoding of categorical variables done. Finally a correlation matrix formed that shows relation of features with output variable Purchase. It shows that Product\_Category\_3 was highly related or important to Purchase. Product\_Category\_3 and Product\_Category\_1 are reciprocal to each other one increases other decreases. Product\_Category\_2 negative relation with purchase, customers buying less products for Product\_Category\_2 due to high price assuming as no other information about products and user. Marital Status and Age were correlated positively that makes sense.

**Use case (2) and (4)**



**5.** **Data Modeling**

Three algorithms applied on the data set.

**Purchase Amount Prediction:**

1. **Linear Regression:**

Encoded data set is provided to the model and results reported.

Data split in training and test with 80:20 and 70:30 ratio. Results

Not quite satisfactory so, normalization of data set is done. All columns normalize to (0-1) Still results only 15 to 16 percent accuracy.

**2. Random Forest:**

Same Split performed on Random Forest on first run gives good result 50%. Changing parameters increase the accuracy by 10 to 15 percent further changed reports 20 percent increment in accuracy. Normalization does not boost accuracy.

**Gender Prediction:**

Gender prediction done with occupation and purchase amount with KNN Classifier. Gender predicted with an accuracy of 74 %.

**Output** **Results**

|  |  |  |
| --- | --- | --- |
| Model Name | Training Accuracy | Testing Accuracy |
| Linear Regression | 15% (26 on kfolds =1000) | 14% (25 on kfolds = 1000) |
| Random Forest | 65% (approx 70) | 64% (approx 68 ) |
| KNN Classifier | 78 %  best k = 27  75.6% | 70 %  best k = 27  75.1% |