Prediction of Purchase Amount for retail store customers

Capstone 1- Springboard data science career track

Problem Statement:

A retail Company wants to understand the customer purchase behaviour (specifically ,purchase amount) against various product of different categories during Black Friday sales.
Dataset contains the 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.

Now, 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.

Dataset Information

The dataset contains about 537577 observations and 12 variable.
FOLLOWING COLUMNS AND DATA TYPES ARE OBSERVED WHICH ARE CONVERTED TO SUITABLE DATATYPES FOR ML MODELS.

COLUMN NAMES	DATATYPES
User_ID	537577 non-null int64
Product_ID	537577 non-null object
Gender	537577 non-null object
Age	537577 non-null object
Occupation	537577 non-null int64
City_Category	537577 non-null object
Stay_In_Current_City_Years	537577 non-null object
Marital_Status	537577 non-null int64
Product_Category_1	537577 non-null int64
Product_Category_2	370591 non-null float64
Product_Category_3	164278 non-null float64
Purchase	537577 non-null int64

DATA WRANGLING

FOLLOWING COLUMNS AND DATA TYPES ARE OBSERVED WHICH ARE CONVERTED TO SUITABLE DATATYPES FOR ML MODELS.

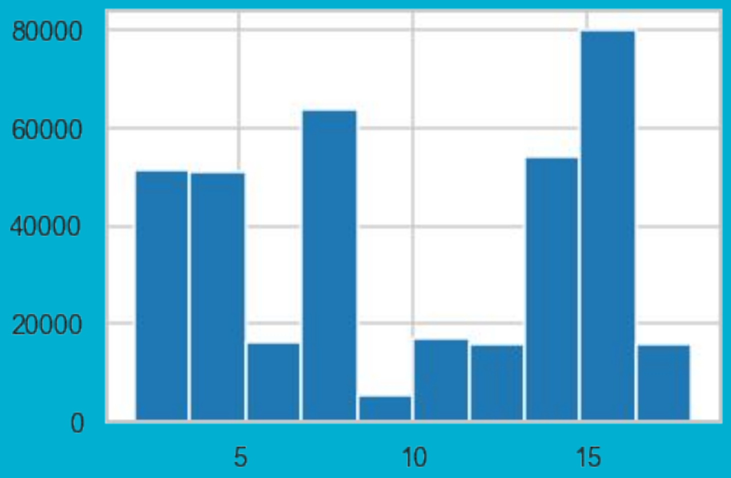
COLUMN NAMES	DATATYPES
User_ID	537577 non-null int64
Product_ID	537577 non-null object
Gender	537577 non-null object
Age	537577 non-null object
Occupation	537577 non-null int64
City_Category	537577 non-null object
Stay_In_Current_City_Years	537577 non-null object
Marital_Status	537577 non-null int64
Product_Category_1	537577 non-null int64

MISSING VALUES

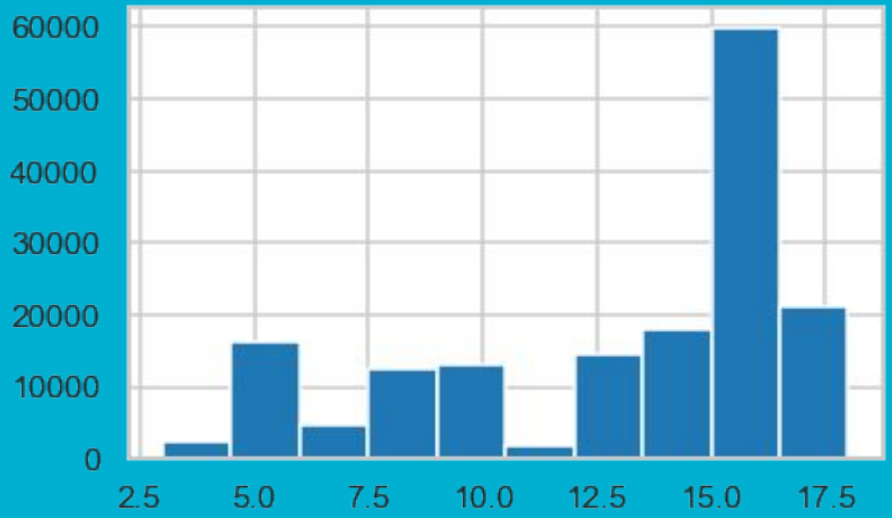
It is observed that there are missing values in both of the columns as mentioned below:

histogram distribution for Product Category is non-normal distribution.However while imputing missing value ,we can replace the 'nan' value with 0 ,considering non-purchase of perticular item from the category.

Product_Category_2 0.310627%

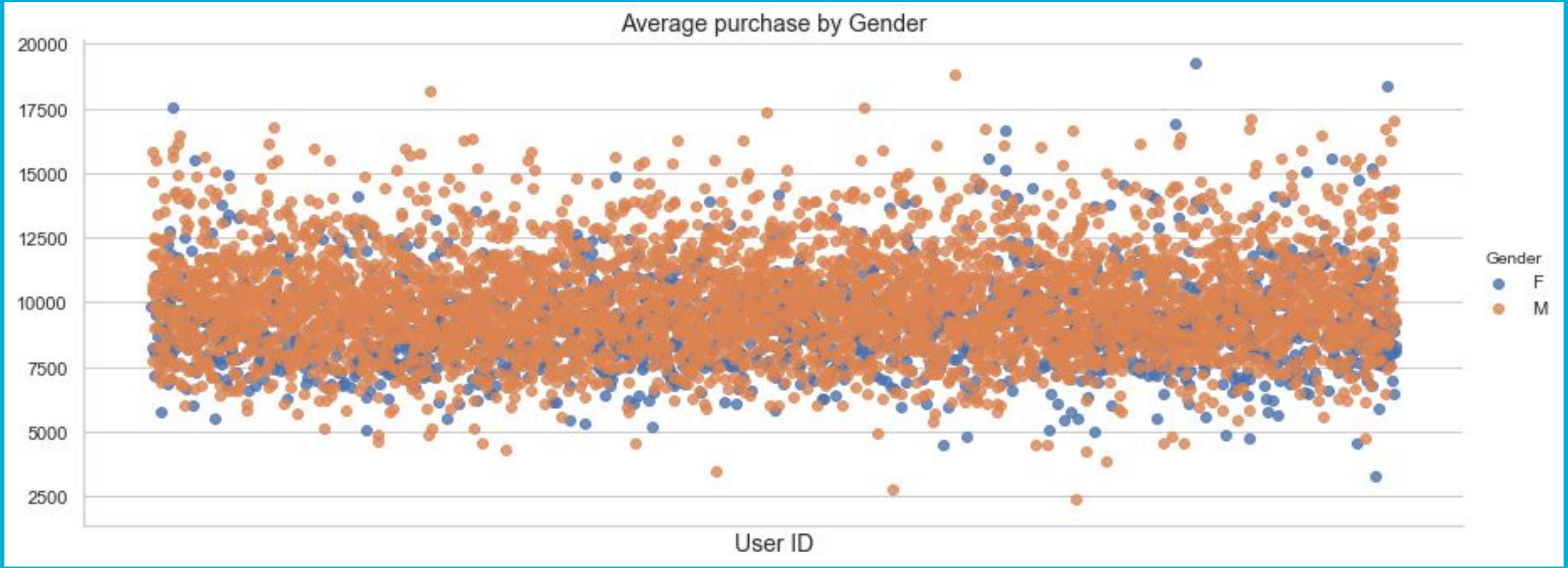


Product_Category_3 0.694410%

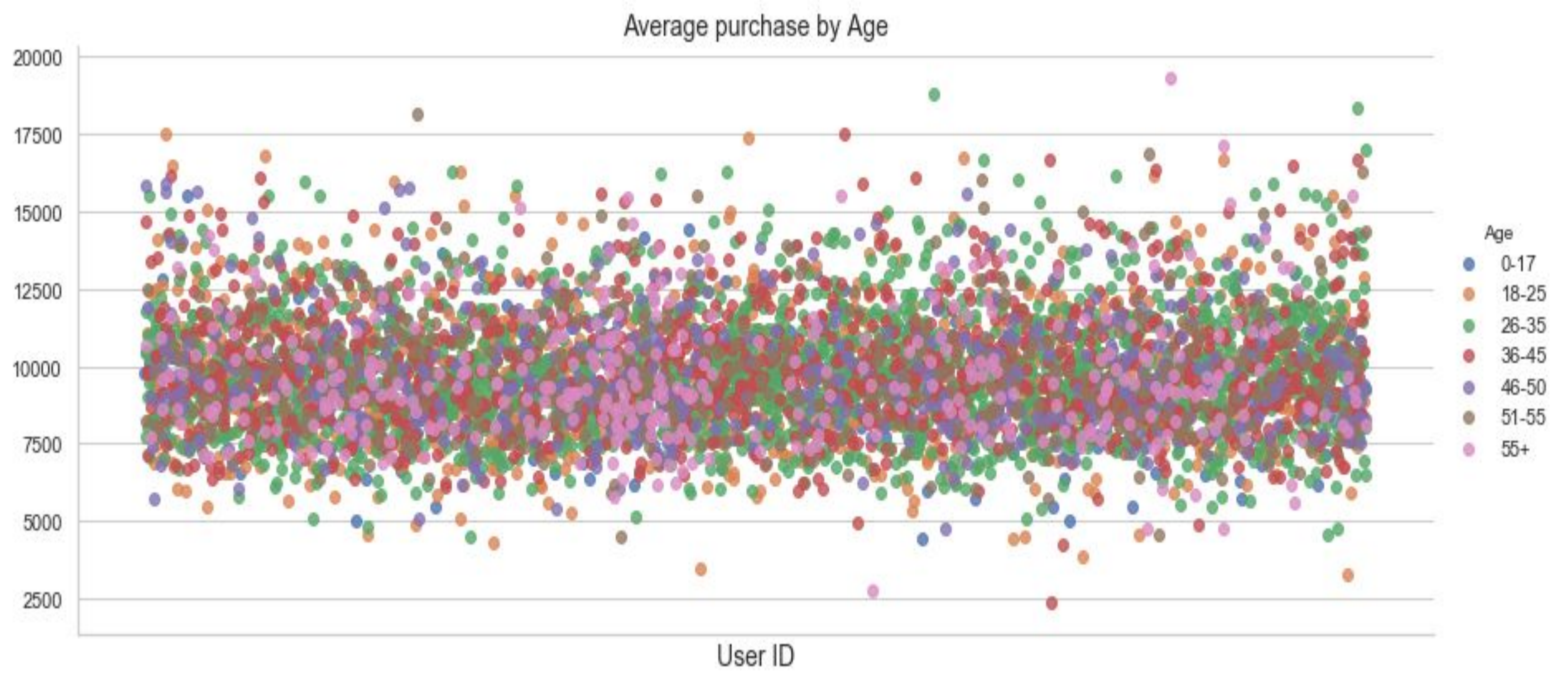


DATA VISUALISATION

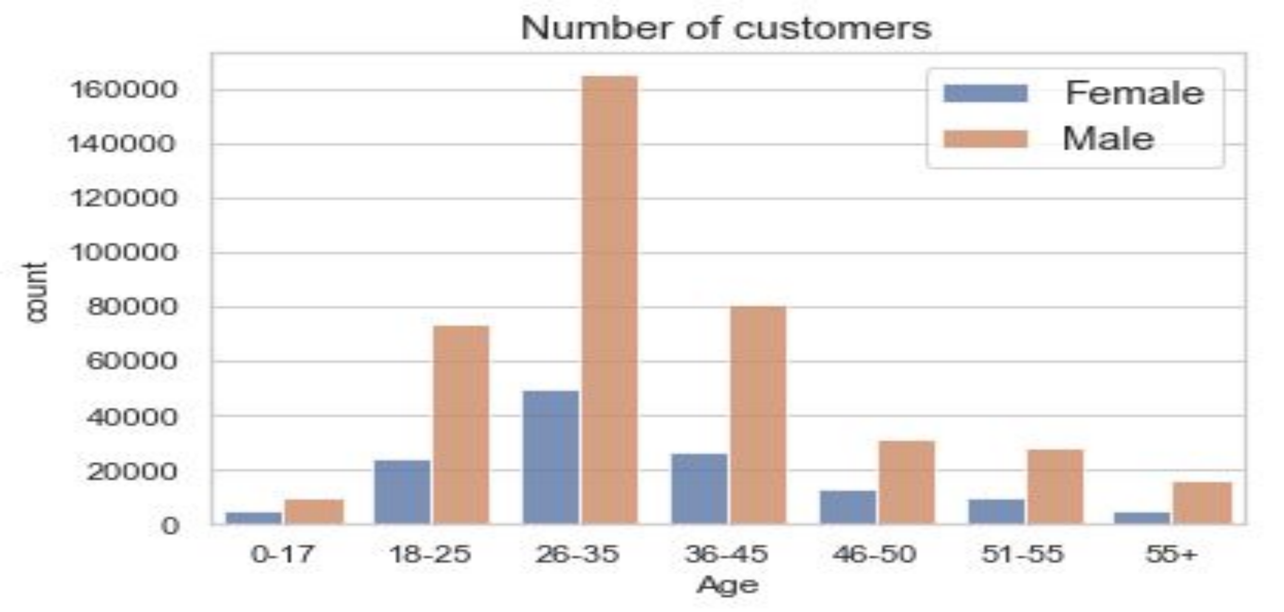
We can see the dominance of Male Buyers:



Genderwise purchase distribution



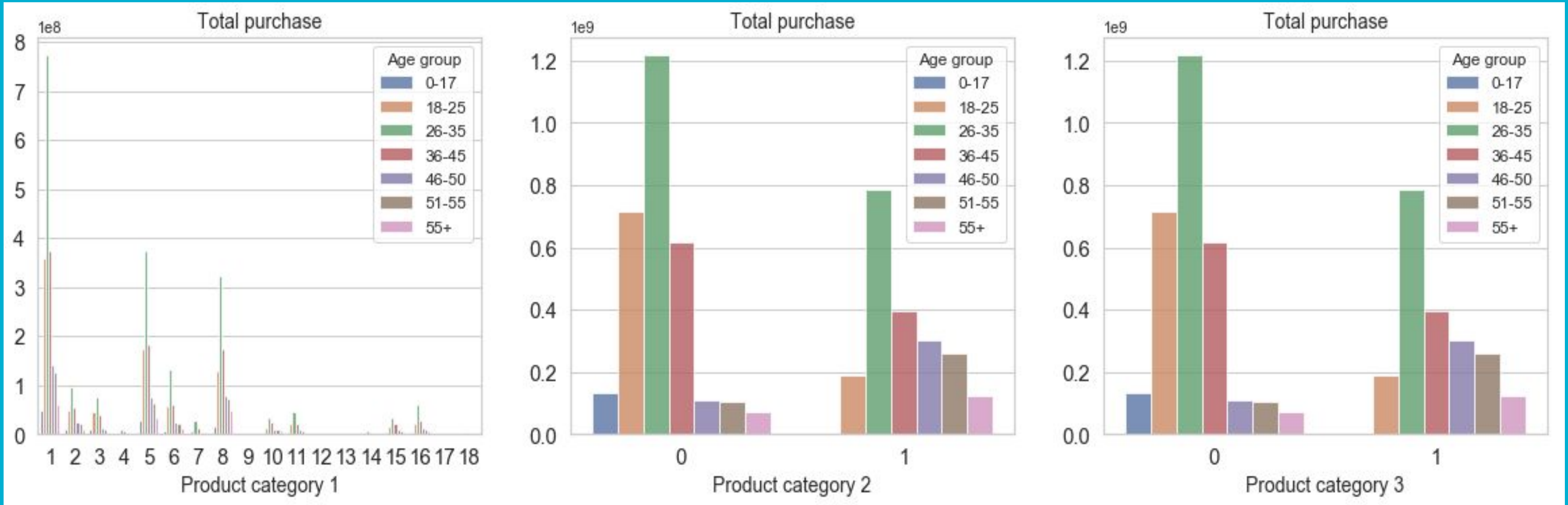
Gender Count between different Age Group



Total purchase
1e8
2.5
2.0
1.5
1.0
0.5
0.0
A
B
C
City
Occupation
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Total purchase
1e9
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Married
Single
A
B
C
City
1e8
Residency duration
0
1
2
3
4
7
6
5
4
3
2
1
0
A
B
C
City

We can observe that Items from Product Category1 are purchased more



Observations based on Exploratory data analysis

1.There are more numbers of male shoppers than female

2.The maximum selling item belongs to product category

3.More numbers of shoppers are single

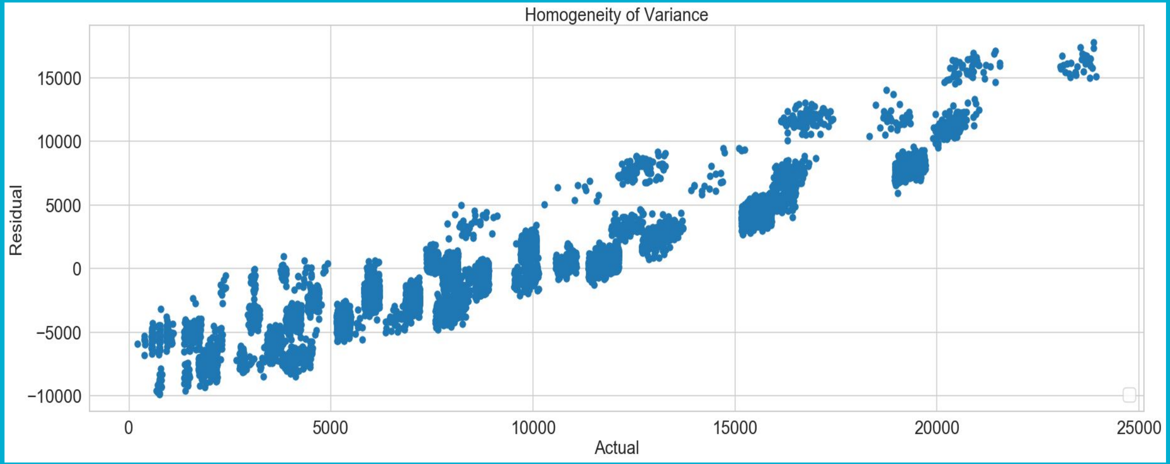
4.The shoppers in the range 26-35

5.Customers from City B shopped the most

6.Customers who has resided in their city for 1 year shopped the most

Machine Learning Models

This is Supervised Model and we need to perform regression Analysis to predict Purchase Amount by Customer.
Based on below plot ,it is clear that data does not follow the linearity assumption such as homogeneity of variance which is essential for linear regression.



Tree Based Models

As our model fails to perform on Linear Regression with very low r2 of 0.107.

Its better to try tree based ensemble models.

I have tried following models:

1.Decision Tree

2.Random Forest

3.Gradientboosting Regressor

4.XGBoost

ML Model building Process

As we have already done data preprocessing and data wrangling stage .

I have tried to do feature selection and removing constant feature if any.

I also stratified   dataframe in order to remove any sampling biases while splitting

Dataframe using train_test_split.

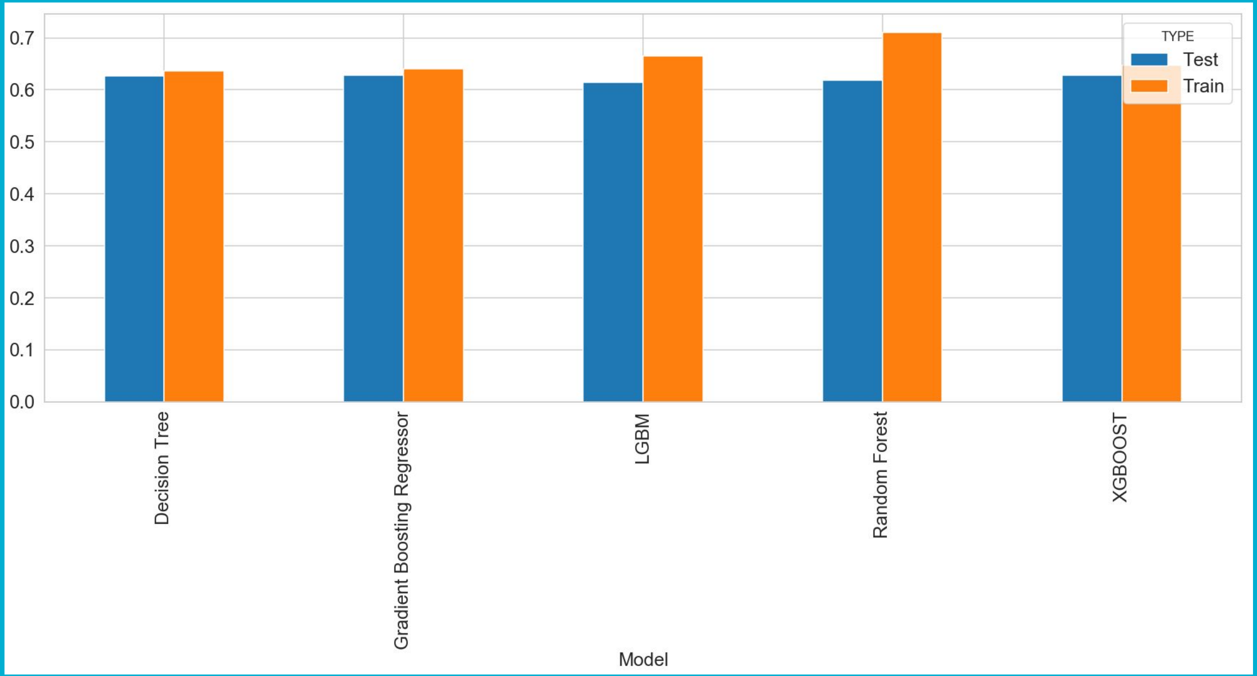
Due to limiting computing capacity,the models are built on only 5% of data which

Accounts to be Train Data size -(18815,39)

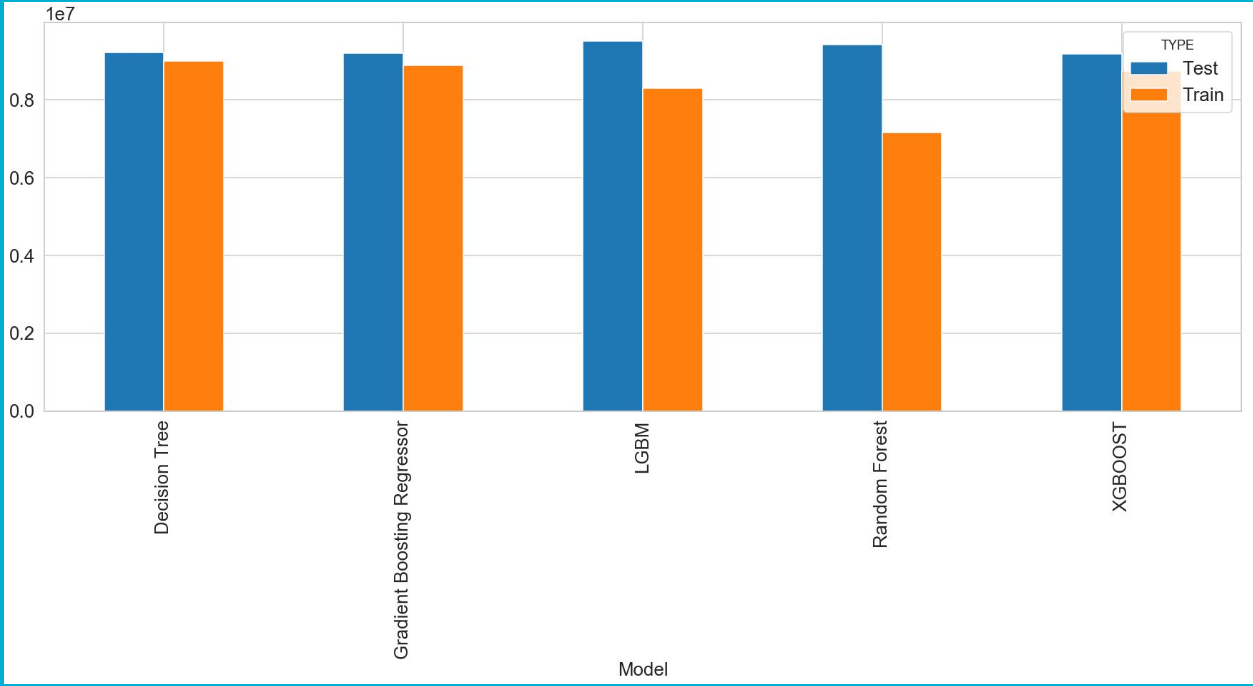
Test Data - (8064,39)

Gridsearch and hyperparametrs tuning are tried to achieve optimum model performance.

R2 Values of Various Algorithms:



MSE Score for various Algorithms:



Conclusion

All the algorithms works equally though the R2 is not very good but we have learned

How the performance of ensembles is improved dramatically from 0.107 r2 values to 0.64 approx. when compared with linear model.

Future work:

By working on whole dataset ,we may try to achieve more robust R2 and better performance by trying different combinations of hyperparameter values.