Springboard Data Science Career Track

Capstone Project 1

Prediction of Purchase Amount for Customers of Retail Store

Sushama Jadhav

Table of Contents

Sr No	Description	Page Number
1	Introduction	1
2	Overview of Data Set	3
3	Data Wrangling	3
4	Initial Findings	5
5	Exploratory Data Analysis	6
6	Machine Learning Models	14

Introduction

As we are aware,everyday ,million of retailers ,from all over the world,offer up products and services to billions of their consumers.In most countries ,retailers and retailing are proven to be backbone to the economy.e.g.Amazon ,Walmart etc.However to achieve this position,retailers have to utilize the resources at their disposal to offer strategies that provides unique customer experience.Also researchers have worked to understand a range of retailing issues,through host of t theoretical and methodological processes,and using vast variety of data collection methods,lab experiments and field studies.

One crucial part of retailing business is effective Marketing.Creative expression develops marketing campaigns that catch the eye and capture the imagination but behind every marketing strategy are theories which relies solidly in psychology, economics, and studies in human behavior. Marketing strategy based on consumer behaviour or why consumers buy, pattern ,age and gender .

Problem Statement:

For this hypothetical dataset, a retail Company wants to understand the customer purchase behaviour (specifically ,purchase amount) against various product of different categories.

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, it is required to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers.

OverView of Dataset:

The original dataset, which was acquired from Kaggle.com, contains 537577 numbers of rows and 12 variables.
The data types for different variables are : float64(2), int64(5), object(5).

Later some of them have converted to suitable data types and imputed using suitable techniques and also checked and corrected for null values wherever needed.The final dataset is used to perform data wrangling and exploratory data Analysis.(Refer .ipynb files)

Data Wrangling:

Data Wrangling involves the checking of Data consistency and data types for the variables.

While exploring data types for each variable following information about datatype is observed and columns are transformed to the appropriate category to make the more sense about the data.

Original Data types are as follows:(Reference .ipynb file)

```
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
```

Following code describes the steps taken to transform the data types to appropriate types to help process the data without any type casting error while using Numpy/Pandas/Scikit-learn etc.

```
# variable type casting to suitable type

df['Product_ID'] = df.Product_ID.astype('category')
df['Gender'] = df.Gender.astype('category')
df['Age'] = df.Age.astype('category')
df['Occupation'] = df.Occupation.astype('category')
df['City_Category'] = df.City_Category.astype('category')
df['Marital_Status'] = df.Marital_Status.astype('category')
df['Product_Category_2'] = df.Marital_Status.astype('int')
df['Product_Category_3'] = df.Marital_Status.astype('int')
```

REMOVAL OF SPECIAL CHARACTER FROM COLUMN VALUES

Also removed the special character from variable values of columns ‘'Stay_In_Current_City_Years'

The code snippet is as below:

df['Stay_In_Current_City_Years'] = (df['Stay_In_Current_City_Years'].str.strip('+').astype(int))

MISSING VALUES:

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

Product_Category_2	0.310627
Product_Category_3	0.694410

However in this particular case it is sensible to impute missing values with 0 considering that the customer might not have brought any particular item where there is ‘NaN’.

Initial Findings:

While preliminary analysis of dataset reveals following descriptive statistics about dataset.

	COUNT	MEAN	STD	MIN	MAX
Total Number of Users	537577	-	-	-	-
Stay in Current City	537577	1.859	1.289	0.00	4.0
Product_Category_1	537577	5.295	3.75	1	18
Product_Category_2	537577	6.784	6.21	0	18
Product_Category_3	537577	3.871	6.26	0	18
Purchase Amount(In Indian Rs)	537577	9333.85	4981.022	185.0	23961

We can observe that there are there is significant difference between Max and Min value of Purchase Amount column. Also observe that customers does not buy Products in category 2 and category 3 as min value appears to be 0.

Though the Product Category 2 has highest mean value,it would be interesting to get more insight about the most profitable product during detailed exploratory data analysis.

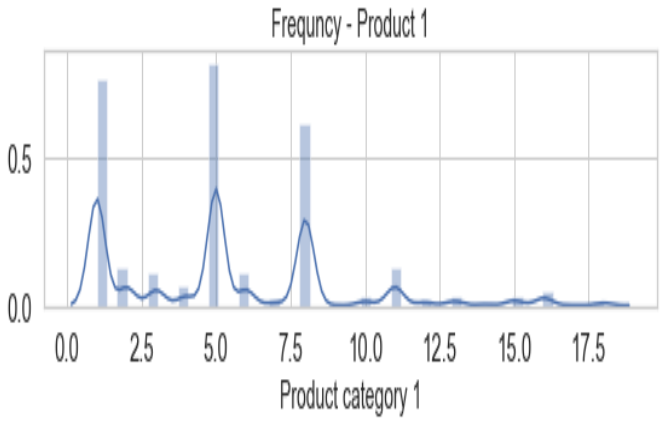
Exploratory Data Analysis:

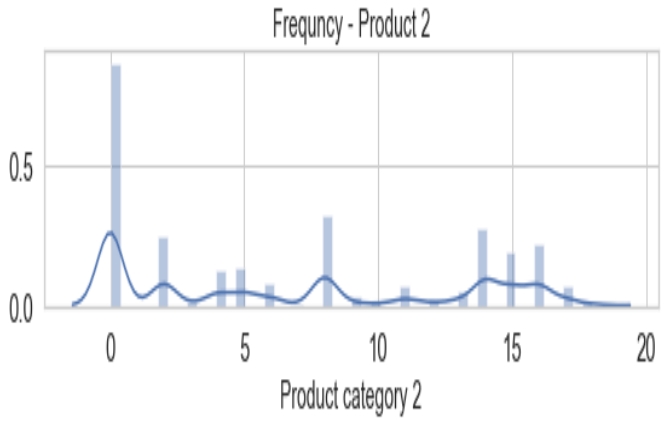
Normality of the Variables:

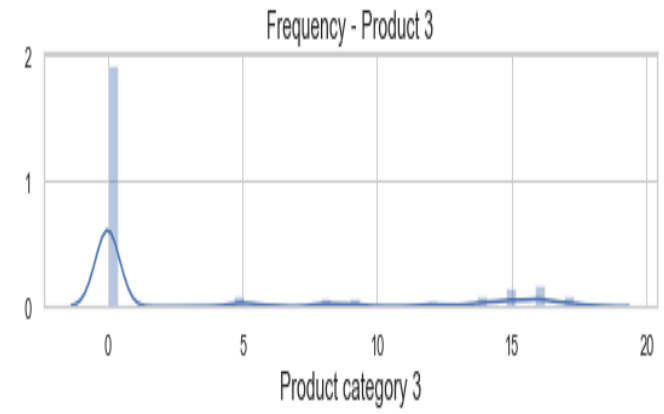
After observing the following frequency distribution

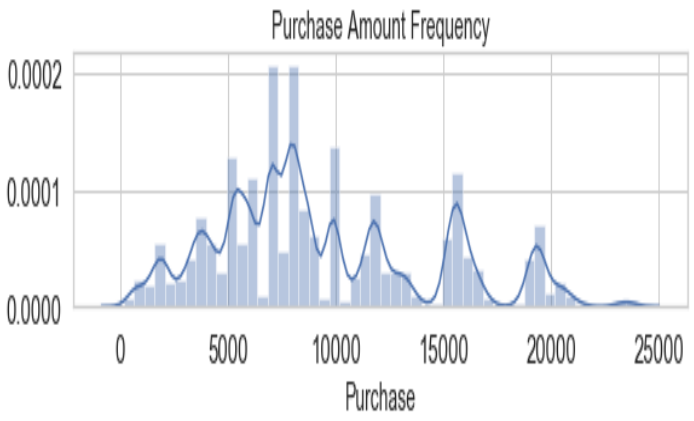
graph , we can see that the data is not normally distributed for the numerical features such as

Purchase Amount and Ordered value for Product 1,Product 2 and Product 3.









Gender specific Purchase Pattern reveals that Mens shopped more than females.It can be observed from the graph as mentioned below.

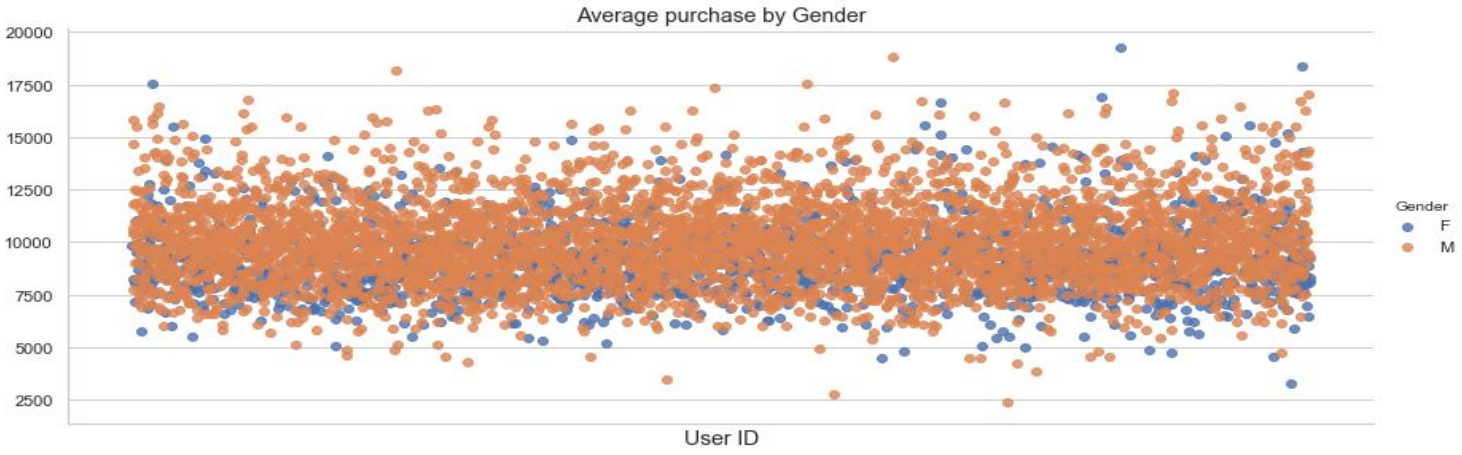


Fig: Scatterplot of Gender wise Average Purchase

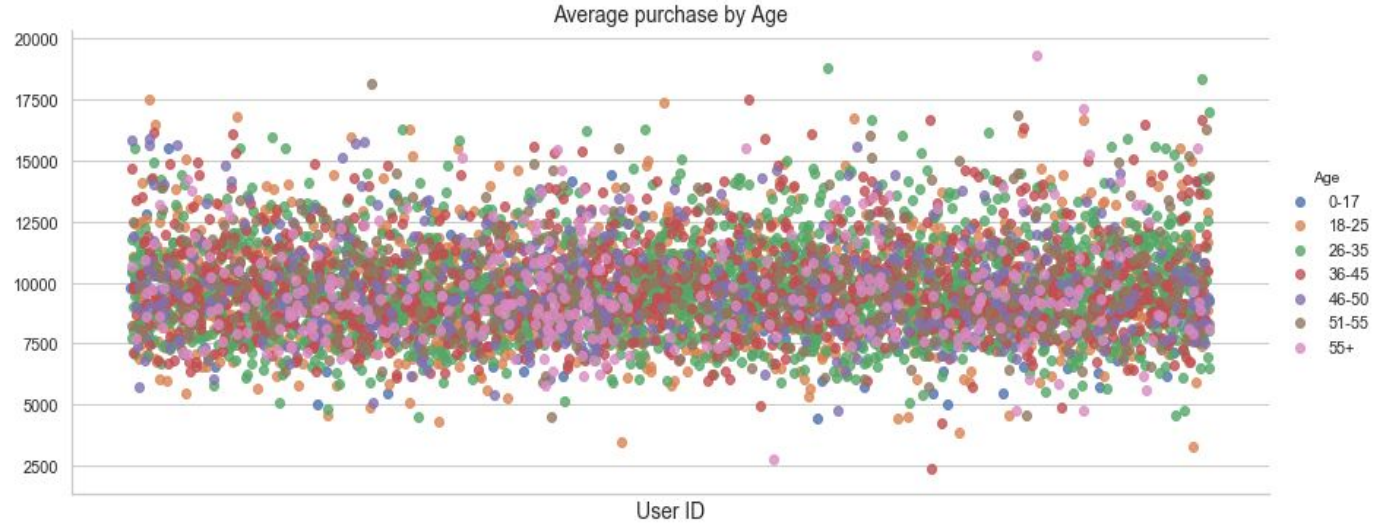
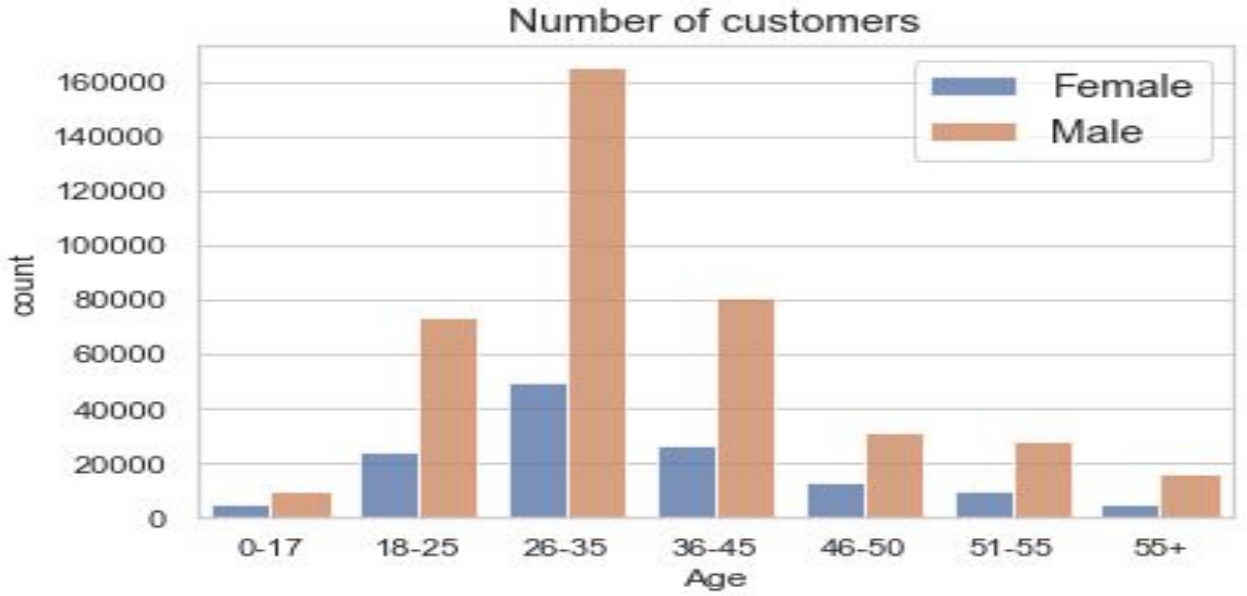


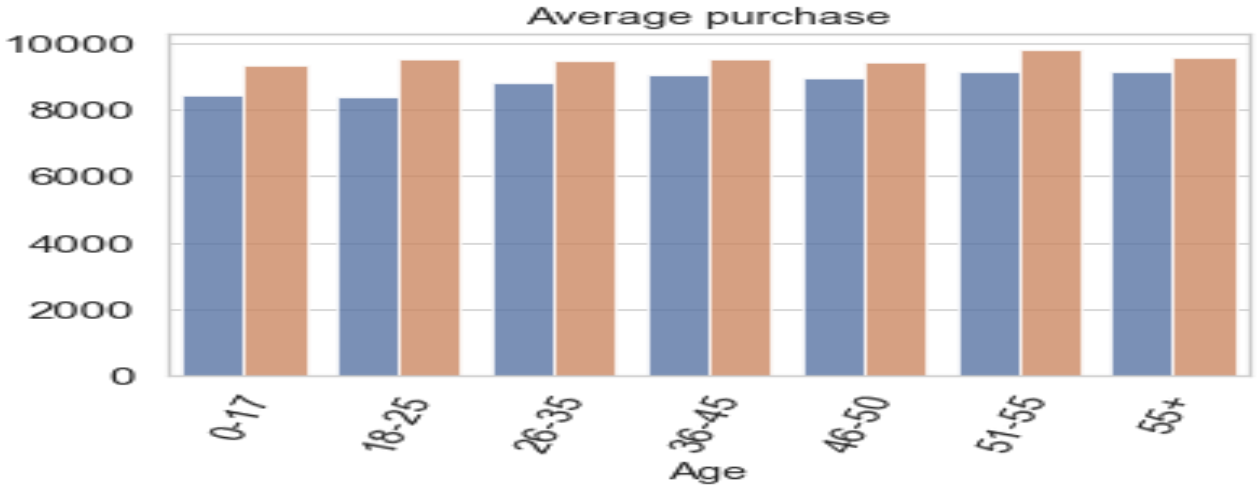
Fig:Scatterplot of Average Purchase in different Age Group



Also as per above visualization , the age range between 26-35 (Green color) and age range between 36-45 (red color) dominates the results of visualization.

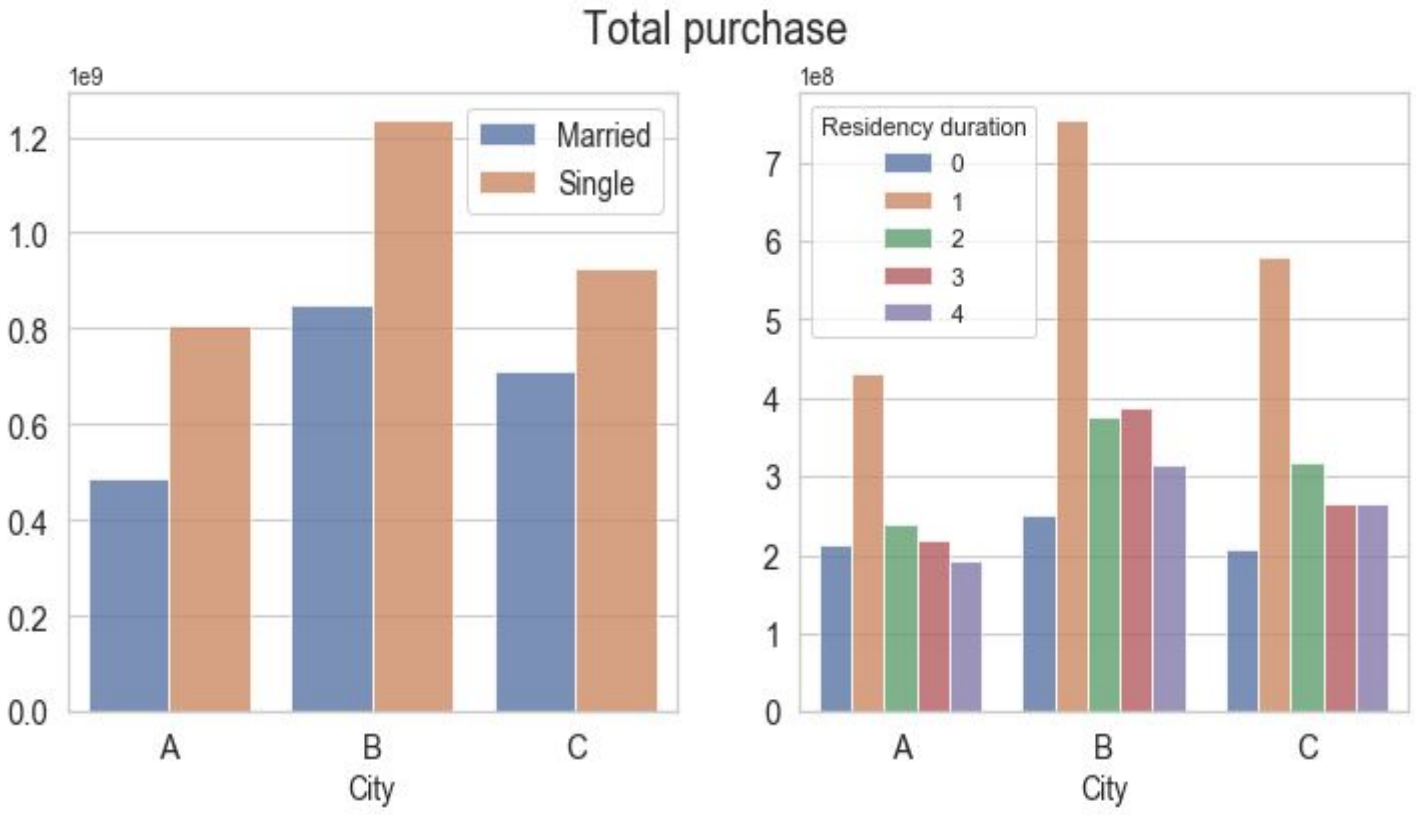
Also observe the dominance of Male in the age range of 26-35 followed by age range 36-45 And 18-25.

However it is interesting to see that average purchase amount between various age group is same,slightly more for Men than women as per below figure.



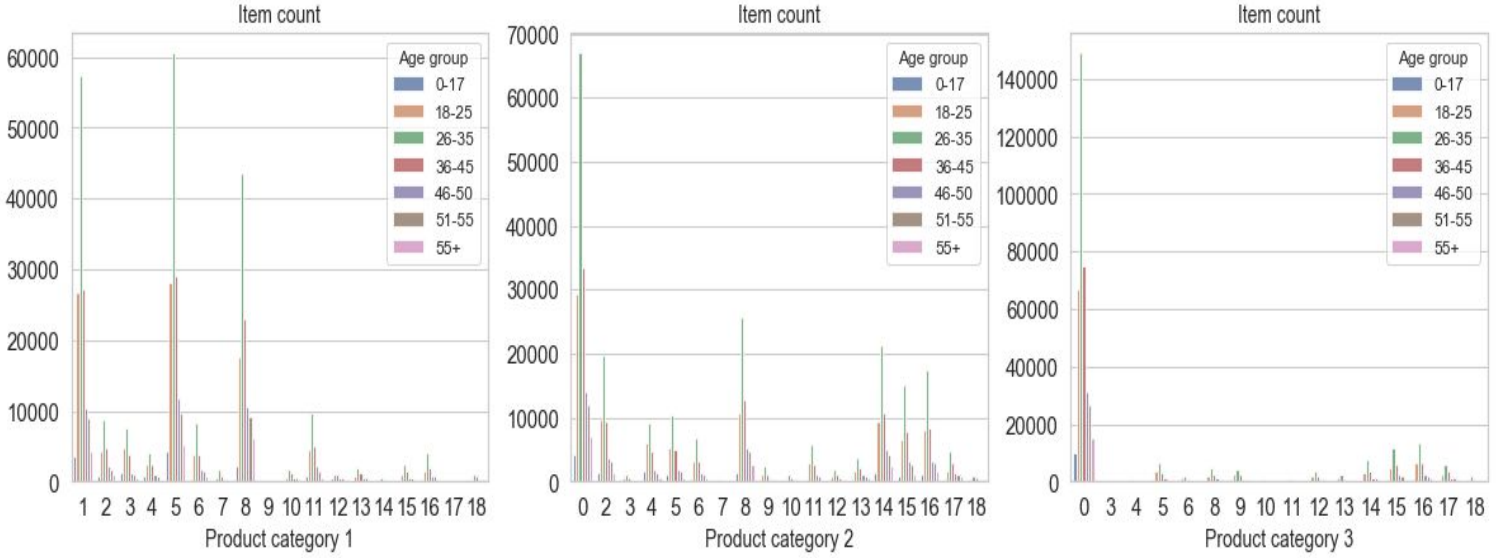
CityWise shoppers Distribution:

It appears that city B dominates in number of Customers. Also the Occupation 0 and 4 dominates as shown below.



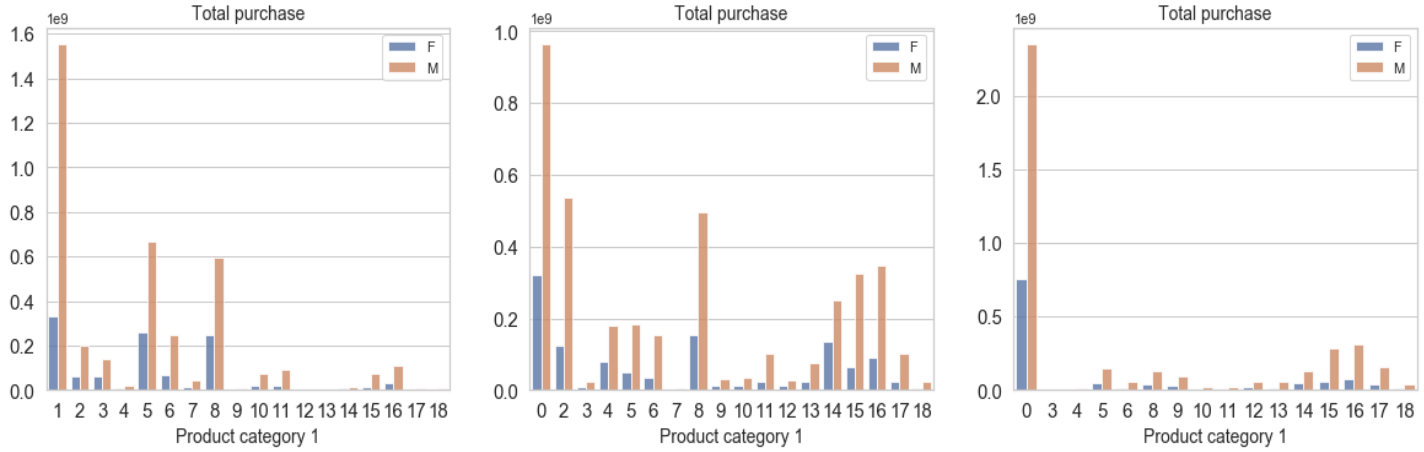
From above plot it is observed that Single males shopped more than Married and people residing at location for about 1 year shopped more compared to other.

Demand for Various items can be visualised as below:



Observe that Item Category 1,5,8 are frequently purchased in Category 1 and 0,2,8,14 ,15,16 are frequently purchased in category 2.

We can see that the maximum purchase amount was collected from item sell of Product Category 1.



Inferential Statistics

The inferential statistics starts with the exploration of normality of various variables which seems to be important such as Purchase amount,Product_Category_1,Product_Category_2,Product_Category_2 and age,gender etc .

And various distributions amongst them are statistically explored applying suitable inferential statistics.

Please refer graphs as mentioned in Exploratory data analysis to support the outcome inferential statistics hypothesis.

Problem I:

H0-The Null Hypothesis :Mean Purchase amount for male and female is same

H1-The Alternative Hypothesis:Mean Purchase amount is different in male and female.

After applying Mann-Whitney U test for comparing two samples of non-normal Distribution,it was observed that both genders have different average Purchase amount.

Problem II:

It would be helpful to understand about the demand for various products to manage the supply and demand.

Three sample hypothesis test is performed to find the difference in distribution.

H0- All the products have same order quantity

H1 - All the products have different order quantity

After applying Kruskal test it was observed that three items have different distributions in terms of order quantity.

Problem III:

H0 - All the cities have same average Purchase Amount

H1 - All the cities have different average Purchase Amount

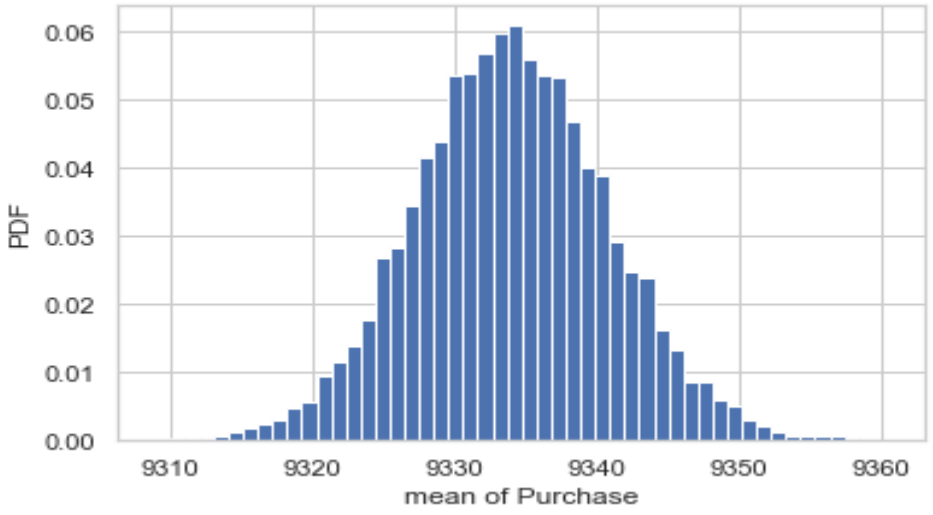
After applying Kruskal test for comparing means of more than 2 variables it was observed that the average purchase amount for all the cities is different.

Confidence interval for non-normal Distribution:

After bootstrapping confidence interval for purchase amount observed to be

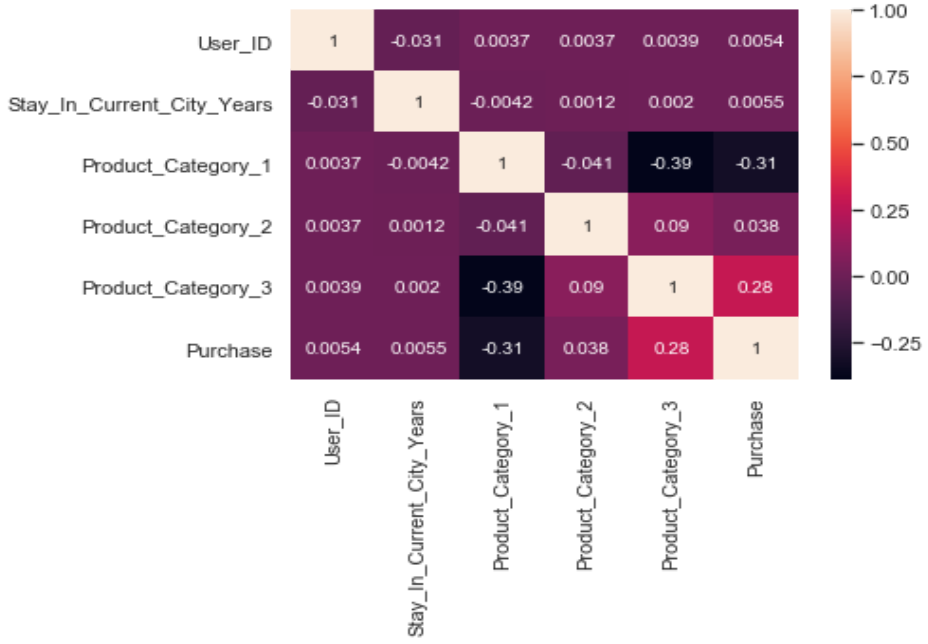
[9320.7775574 , 9347.65727631]

After applying bootstrapping the confidence interval for Purchase amount is as per below graph.



As we can observe , there is a moderate positive correlation observed between Product_Category 3 and Purchase and moderate negative correlation between Product_Category 1 and Purchase.

Also Product category 1 and Product Category 3 are moderately negatively correlated.



It is easy to calculate and interpret when the variables have a well understood Gaussian.

However after detailed investigation using Levene's test of homogeneity of variances it was found

that our variables violated homogeneity of variances.Given that, the appropriate correlation test to use would be a non-parametric test such as the Spearman rank correlation or Kendall Tau correlation test.

After applying Kendall Tau correlation it was found that there was moderate correlation between

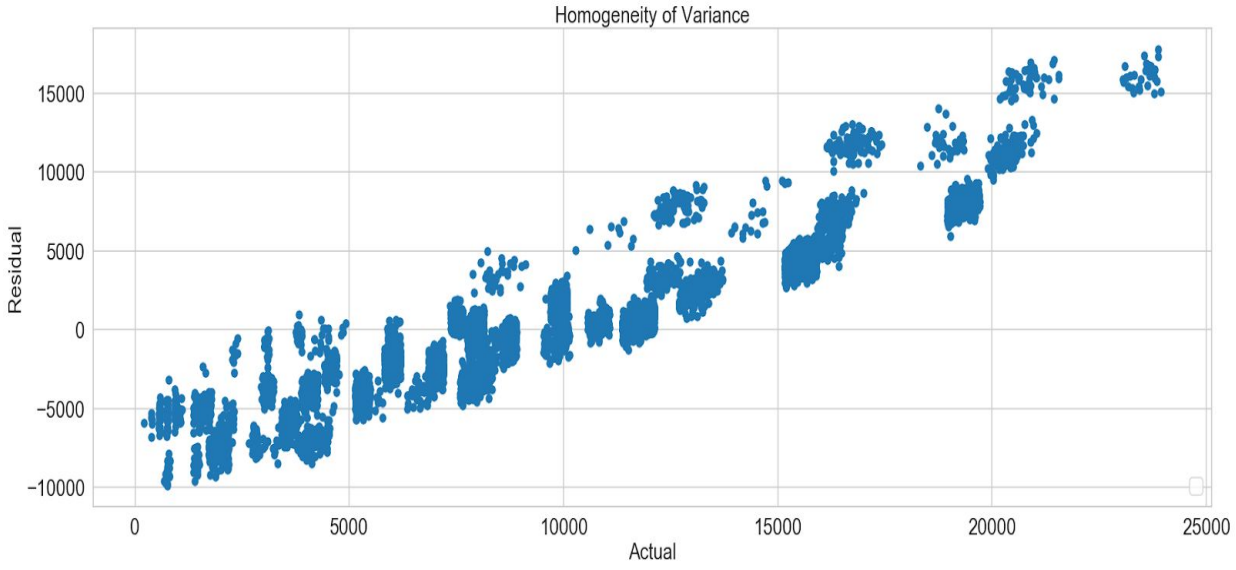
Product_Category_1 and Purchase and weak correlation between Product_Category_2 and Purchase

In depth Analysis of Machine Learning models:

This is Supervised Model and we need to perform regression Analysis to predict Purchase Amount by Customer.

Let's start with basic form of Regression i.e.Linear Regression

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.



The same can be observed as below :

Cross Validation is performed with 10 folds and it is observed that r2 is very low (around 0.11) and same is verified using statsmodel as below figure.

OLS Regression Results

Dep. Variable:	Purchase	R-squared:	0.107
Model:	OLS	Adj. R-squared:	0.107
Method:	Least Squares	F-statistic:	2086.
Date:	Tue, 12 Feb 2019	Prob (F-statistic):	0.00
Time:	03:06:05	Log-Likelihood:	-5.3089e+06
No. Observations:	537577	AIC:	1.062e+07
Df Residuals:	537545	BIC:	1.062e+07
Df Model:	31		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.052e+04	66.779	157.483	0.000	1.04e+04	1.06e+04
Gender[T.M]	470.0124	15.540	30.246	0.000	439.555	500.470
Age[T.18-25]	36.5646	64.635	0.566	0.572	-90.118	163.248
Age[T.26-35]	232.6377	64.430	3.611	0.000	106.357	358.918
Age[T.36-45]	344.4475	65.360	5.270	0.000	216.344	472.551
Age[T.46-50]	335.4300	68.249	4.915	0.000	201.664	469.196
Age[T.51-55]	680.8999	68.995	9.869	0.000	545.672	816.128
Age[T.55+]	503.2829	73.108	6.884	0.000	359.993	646.573
Occupation[T.1]	-68.9745	28.656	-2.407	0.016	-125.139	-12.810
Occupation[T.2]	15.9198	34.471	0.462	0.644	-51.643	83.483
Occupation[T.3]	199.2196	40.194	4.956	0.000	120.440	277.999
Occupation[T.4]	168.8160	27.402	6.161	0.000	115.109	222.524
Occupation[T.5]	54.0673	46.735	1.157	0.247	-37.531	145.666
Occupation[T.6]	193.0073	38.360	5.031	0.000	117.822	268.192
Occupation[T.7]	181.8037	26.913	6.755	0.000	129.056	234.552
Occupation[T.8]	-128.9752	122.277	-1.055	0.292	-368.635	110.685
Occupation[T.9]	-266.1307	63.791	-4.172	0.000	-391.158	-141.103
Occupation[T.10]	-224.6038	69.255	-3.243	0.001	-360.342	

```
-88.866
Occupation[T.11]       35.7040     47.921      0.745      0.456     -58.219     129.627
Occupation[T.12]      479.7159     32.596     14.717      0.000     415.829
543.603
Occupation[T.13]      -37.9645     61.662     -0.616      0.538    -158.821      82.892
Occupation[T.14]      313.5092     34.040      9.210      0.000     246.791     380.227
Occupation[T.15]      550.2019     46.982     11.711      0.000     458.118
642.286
Occupation[T.16]      150.0089     35.292      4.251      0.000      80.838     219.180
Occupation[T.17]      385.0792     30.125     12.783      0.000     326.035
444.123
Occupation[T.18]      -94.9646     61.222     -1.551      0.121    -214.959      25.029
Occupation[T.19]     -356.0993     54.900     -6.486      0.000    -463.702
-248.496
Occupation[T.20]     -129.2615     31.752     -4.071      0.000    -191.494
-67.029
City_Category[T.B]    163.8913     16.013     10.235      0.000     132.507
195.275
City_Category[T.C]    730.2774     17.375     42.031      0.000     696.223
764.332
Marital_Status[T.1]   -31.2820      6.981     -4.481      0.000     -44.965     -17.599
Product_Category_1   -413.2333      1.719   -240.368      0.000    -416.603
-409.864
Product_Category_2    -31.2820      6.981     -4.481      0.000     -44.965
-17.599
==============================================================================
Omnibus:                    56396.849   Durbin-Watson:                   1.713
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            76961.282
Skew:                           0.858   Prob(JB):                         0.00
Kurtosis:                       3.701   Cond. No.                     1.15e+16
==============================================================================
```

As our model fail to perform on Linear Regression Analysis,lets try various ensembling models for non-linear techniques as below:

As a part of this report ,the following tree based models are evaluated on the data.

1.Decision Tree:

Decision trees are built using recursive partitioning, which splits the data into subsets based on several dichotomous independent attributes. This recursive process may split the data multiple times until the splitting process terminates after a particular stopping criterion is reached. The best split is the one that maximizes a splitting criterion. For classification learning, the techniques used as the splitting criterion are entropy and information gain, the Gini index, and the gain ratio. For regression tasks, however, standard deviation reduction is used.

2.Random Forest:

A random forest is a supervised machine learning algorithm based on ensemble learning. It is used for both regression and classification problems. The general idea behind random forests is to build multiple decision trees and aggregate them to get an accurate result. A decision tree is a deterministic algorithm, which means if the same data is given to it, the same tree will be produced each time. They have a tendency to overfit, because they build the best tree possible with the given data, but may fail to generalize when unseen data is provided. All the decision trees that make up a random forest are different because we build each tree on a different random subset of our data. A random forest tends to be more accurate than a single decision tree because it minimizes overfitting.

3.Gradientboosting Regressor

Gradient boosting is a machine learning technique that works on the principle of boosting, where weak learners iteratively shift their focus toward error observations that were difficult to predict in previous iterations and create an ensemble of weak learners, typically decision trees.

Gradient boosting trains models in a sequential manner, and involves the following steps:

1. Fitting a model to the data
2. Fitting a model to the residuals
3. Creating a new model

4.XGBoost

XGBoost stands for extreme gradient boosting. It is a variant of the gradient boosting machine that aims to improve performance and speed. The XGBoost library in Python implements the gradient boosting decision tree algorithm. The name gradient boosting comes from its us of the gradient descent algorithm to minimize loss when adding new models. XGBoost can handle both regression and classification tasks.

XGBoost is the algorithm of choice among those participating in Kaggle competitions because of its performance and speed of execution in difficult machinelearning problems.

Some of the important parameters that are used in XGBoost are as follows:

- n_estimators/ntrees: This specifies the number of trees to build. The default value is 50.
- max_depth: This specifies the maximum tree depth. The default value is 6. Higher values will make the model more complex and may lead to overfitting. Setting this value to 0 specifies no limit.
- min_rows: This specifies the minimum number of observations for a leaf. The default value is 1.
- learn_rate: This specifies the learning rate by which to shrink the feature weights. Shrinking feature weights after each boosting step makes the boosting process more conservative and prevents overfitting. The range is 0.0 to 1.0. The default value is 0.3.
- sample_rate: This specifies the row sampling ratio of the training instance (the *x axis*). For example, setting this value to 0.5 tells XGBoost to randomly collect half of the data instances to grow trees. The default value is 1 and the range is 0.0 to 1.0. Higher values may improve training accuracy.
- col_sample_rate: This specifies the column sampling rate (the *y axis*) for each split in each level. The default value is 1.0 and the range is

5.LightGBM

LightGBM is an open source software for the gradient boosting framework that was developed by Microsoft. It uses the tree-based algorithm differently to other **Gradient Boosting Machines** (**GBMs**)

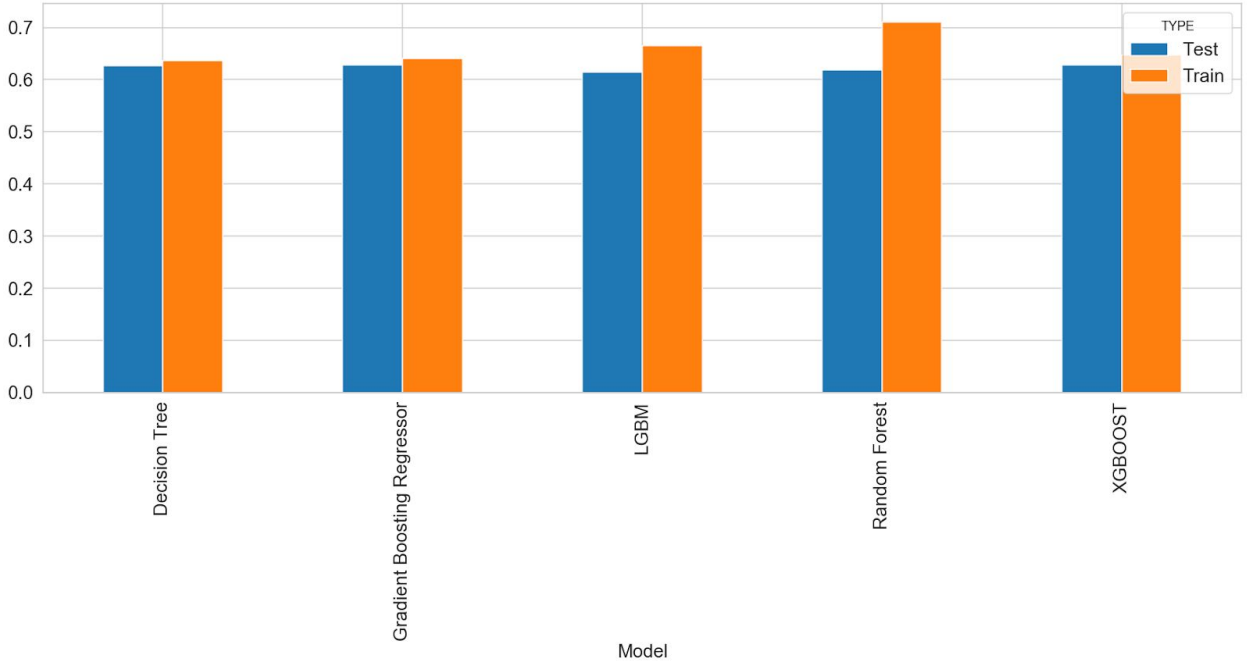
Feature Selection:

1.Removal of constant value features
There are no constant value features observed.

2.Cardinality
As we envisaged during data analysis stage ,there is no cardinality present in the data.

Due to limiting computing capacity ,the machine learning models are built on only 5%
Of the data which accounts to be
Train Data size-(18815, 39)
Test Data size-((8064, 39)

The Gridsearch was performed in order to get maximum accuracy and cross validation parameter is limited to 2 owing to limiting computing capacity.

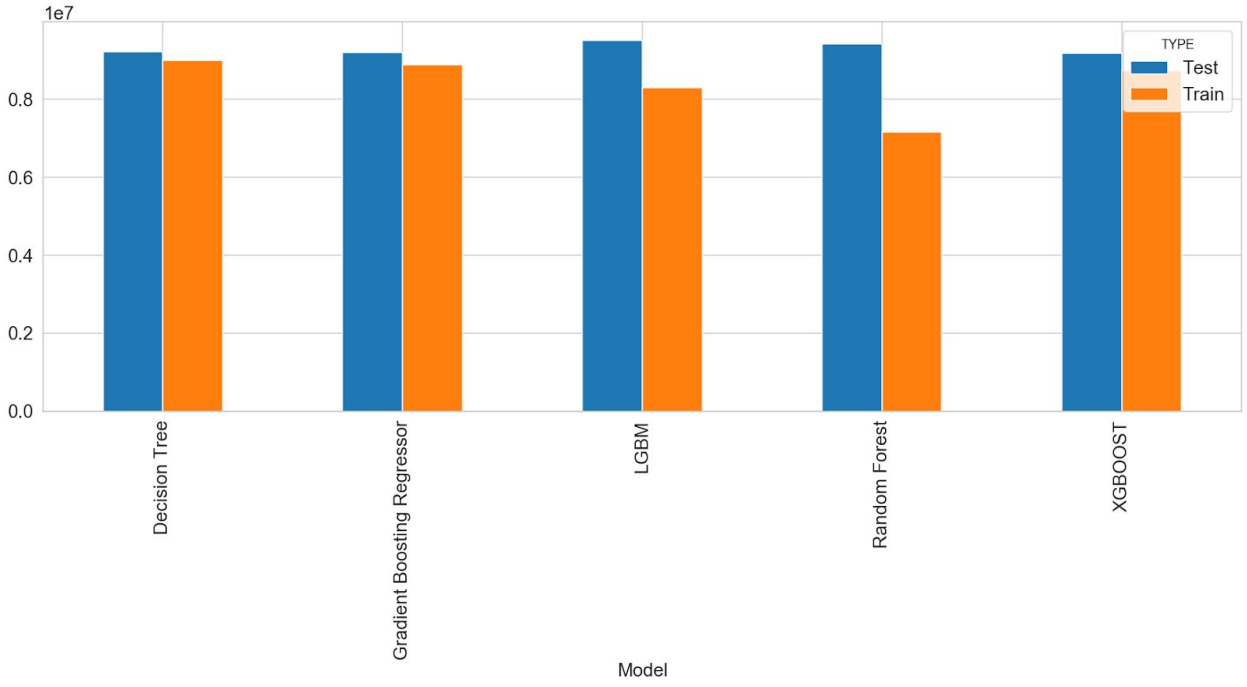


Based on above plot ,we can say that Gradient Boosting Regressor performed well on Training test ,however r2 value for training is higher than test data.Model has overfitting.
Decision Tree,LGBM and XGBoost have performed equally but the r2 value 0.63

Random forest performed well as it has high r2 but overfitting.An overfit model may look impressive on the training set, but will be useless in a real application.

Therefore, the standard procedure for hyperparameter optimization accounts for overfitting through cross validation.

Also we can observe the Mean squared Errors graphically as below which are inline with R sqared value as plotted above.



As above results are based on sample data.But with the help of good computing resources and advanced technologies for handling large amount of data it may possible to achieve good prediction accuracy without overfitting.

Simply speaking, ensemble machine learning refers to a technique that integrates output from multiple learners and is applied to a dataset to make a prediction. These multiple learners are usually referred to as base learners. When multiple base models are used to extract predictions that are combined into one single prediction, that prediction is likely to provide better accuracy than individual base learners.

To perform well, the ensemble models require a sufficient amount of data. Ensemble techniques prove to be more useful when you have large and non-linear datasets.

Irrespective of how well you fine-tune your models, there's always the risk of high bias or high variance. Even the best model can fail if the bias and variance aren't taken into account while training the model. Both bias and variance represent a kind of error in the predictions. In fact, the total error is comprised of bias-related error, variance-related error, and unavoidable noise-related error (or irreducible error). The noise-related error is mainly due to noise in the training data and can't be removed. However, the errors due to bias and variance can be reduced.

Hyperparameter tuning relies more on experimental results than theory, and thus the best method to determine the optimal settings is to try many different combinations evaluate the performance of each model.

Metrics Evaluation:

As we know the most commonly used evaluation metrics for regression analysis as below:

1. Mean Absolute Error.
2. Mean Squared Error.
3. R^2.

In this project we tried to analyse metric using Mean Squared Error and R2 value but the MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the *variance* in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude
Both the MAE and RMSE can range from 0 to ∞. They are negatively-oriented scores: Lower values are better.

Conclusion and Future Scope:

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.