Supervised Machine Learning Capstone Project

Bike Sharing Demand Analysis

By
DEEPAK

INTRODUCTION

The increased usage of private vehicles in metropolitan areas has resulted in significant rise in fuel consumptions that have adverse effect on the climate. It has led people in today’s society to accept problems like road traffic as the norm. Therefore the government and organizations started adopting measures to facilitate sustainable development to address the issue.

In that context, the Bike Sharing initiative was launched to tackle the public mobility problem. It provided the people with an alternative to using a sustainable mode of transport for a small distance at a minimal cost. And gave people the freedom to utilize the service by themselves. In a bike-share system, a user could lend a bike from any bike stations and return it to a bike station near the destination and since it involves the activity of pedaling the bike it has beneficial health effects. And the city-wide installation of bike stations improved the accessibility of areas by bikes.

PROBLEM STATEMENT

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

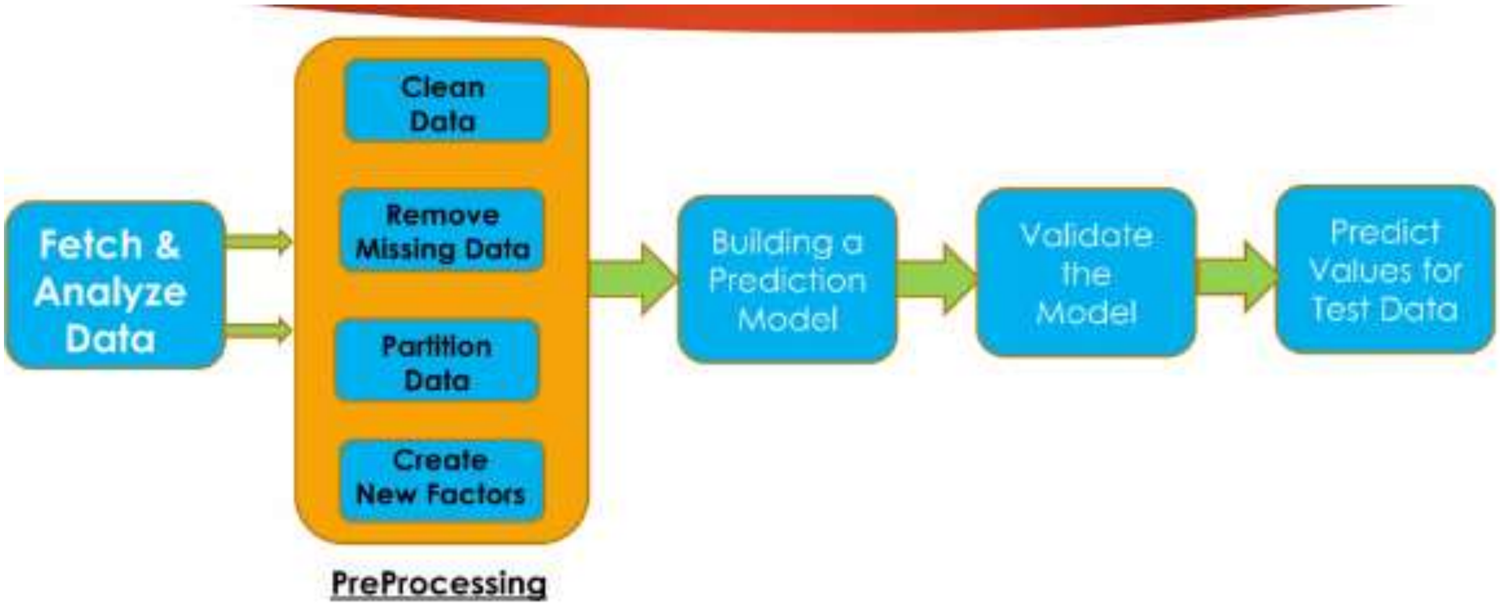
Overview of our Dataset

The file we have used is ‘SeoulData.csv’ in which there are 14 columns and 8760 rows. Data is totally cleaned having no null values and no duplicate data found in it. Among 14 columns, we have 3 categorical columns- Functional Day , Holiday, Seasons and other 10 columns are numerical in nature. Here our target feature is ‘Rented bike count’

Column in our dataset

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind speed - m/s
- Visibility - 10m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day – No Func(Non Functional Hours), Func(Functional hours)

Proposed Methodology



Data Pipeline

- Exploratory Data Analysis (EDA): In this part we have done some EDA on the features to see the trend.

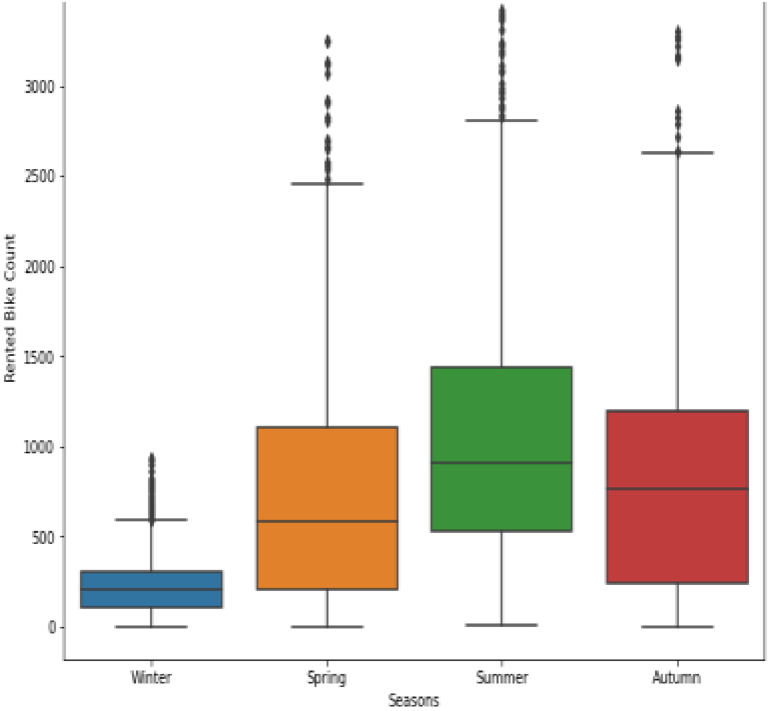
- Data Processing: In this part we went through each attributes and encoded the categorical features.

- Model Creation: Finally in this part we created the various models. These various models are being analyzed and we tried to study various models so as to get the best performing model for our project.

EDA key insights

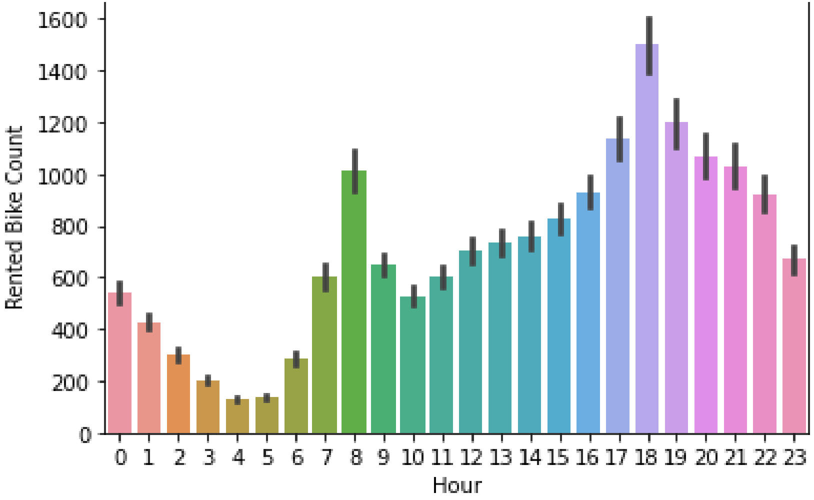
Observation-1

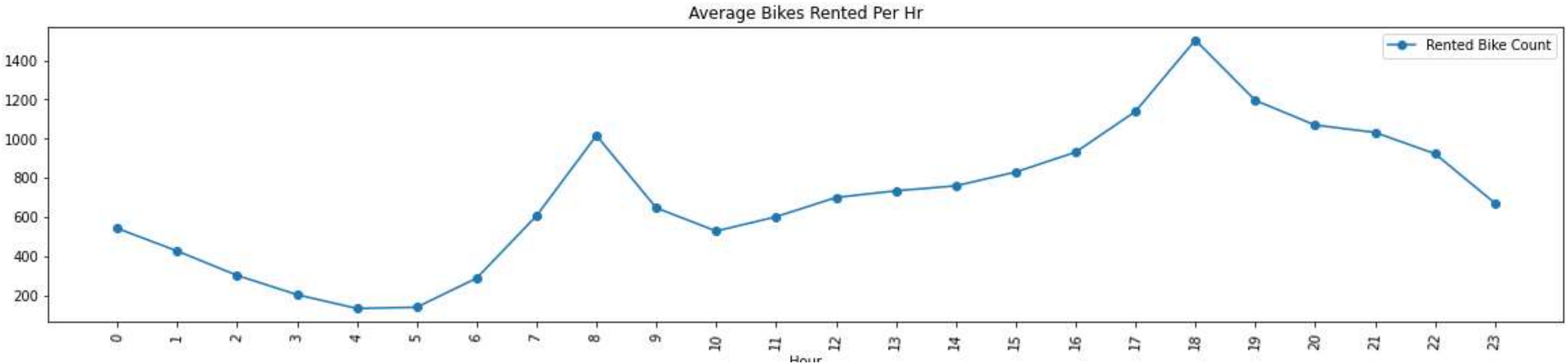
From the boxplot, it is seen that a very high demand of rented bike is seen in summer season and very low demand seen in winter season. We can arrange the season also as per the rented bike demand in decreasing order: Summer>Autumn>Spring>Winter



Observation-2

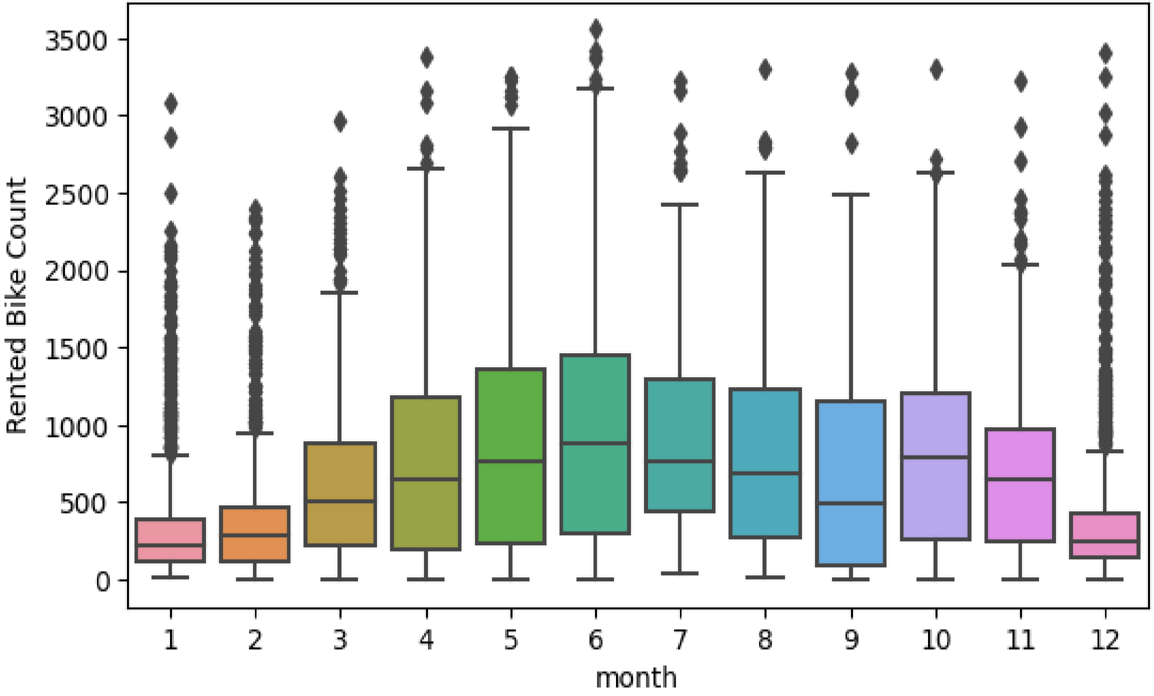
There is a surge in high demand in the morning 7-9 AM and in evening 6-8 PM as the people might be going to their work in the morning and returning from their work in the evening.



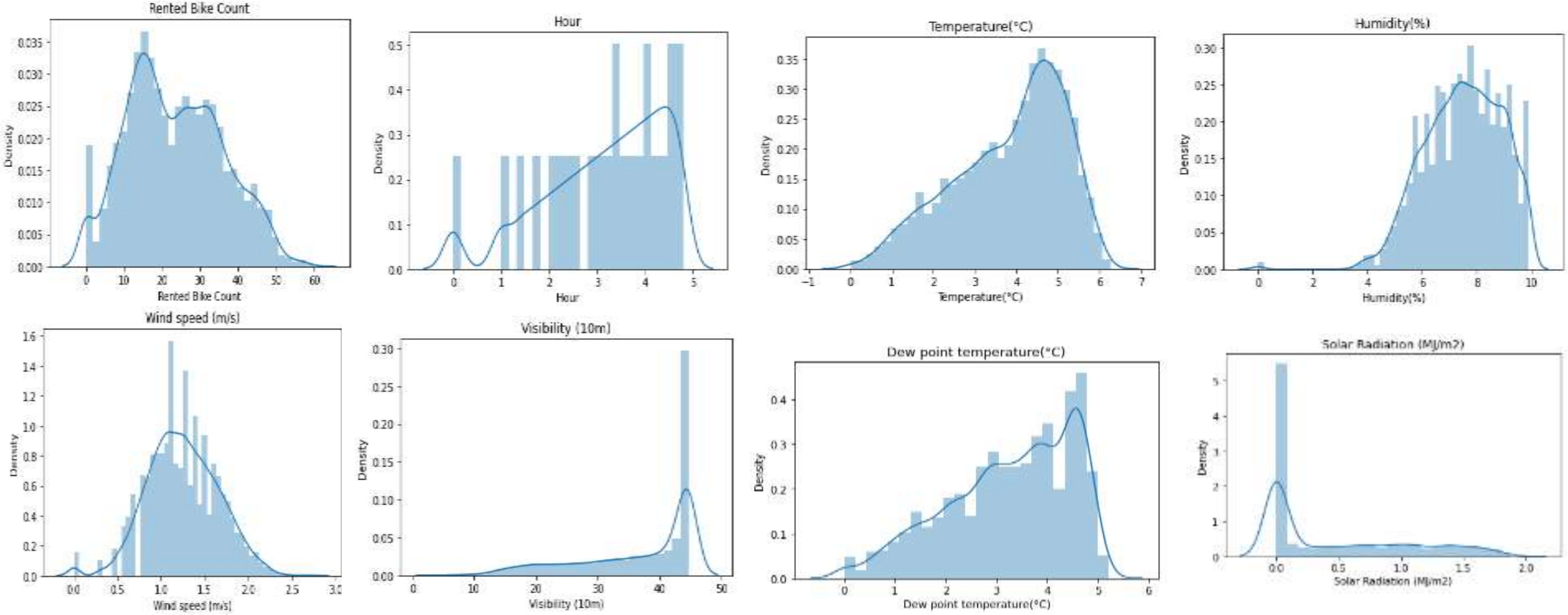


Observation-3

- From the boxplot graph, we can interpret that rented bike demand is low from December to February and then from march, the demand goes on increasing and maximum in June – July (summer season).

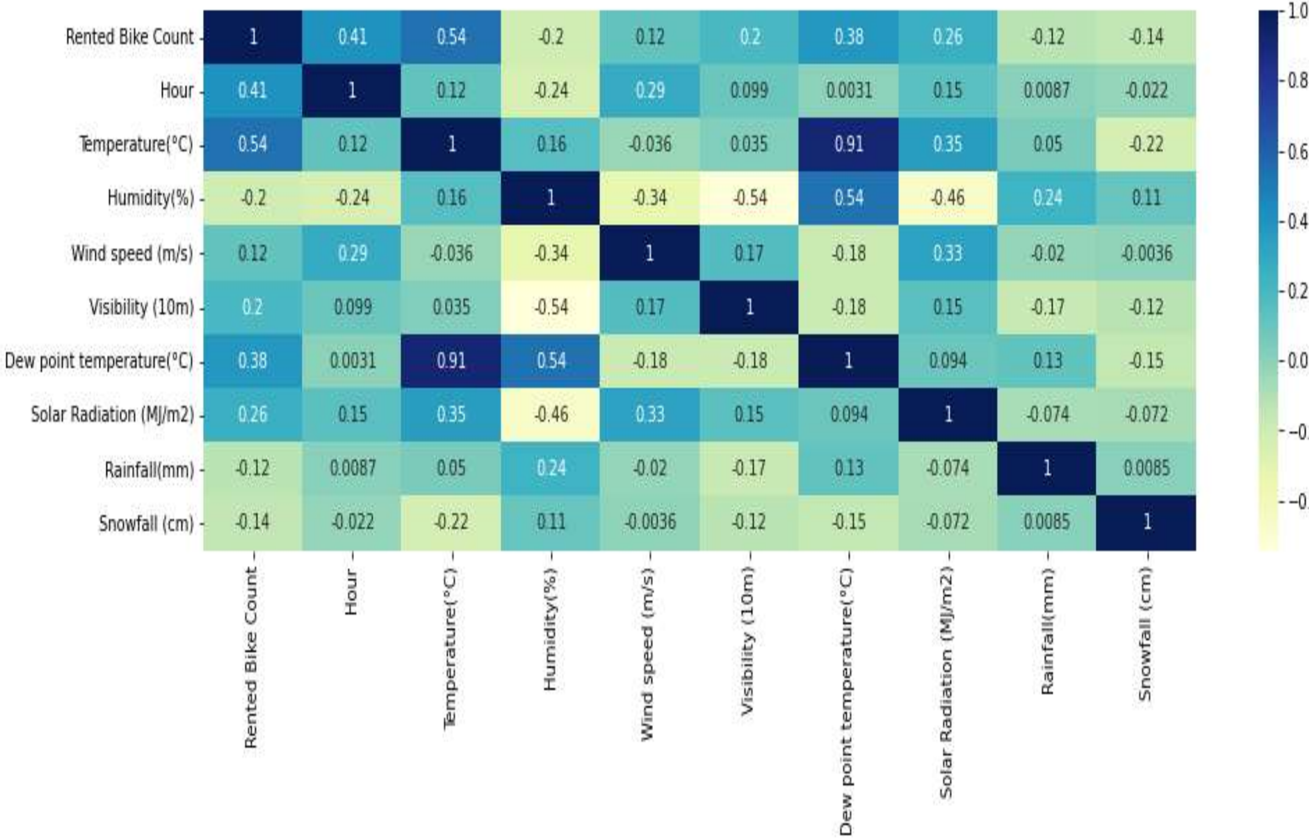


Observation-4



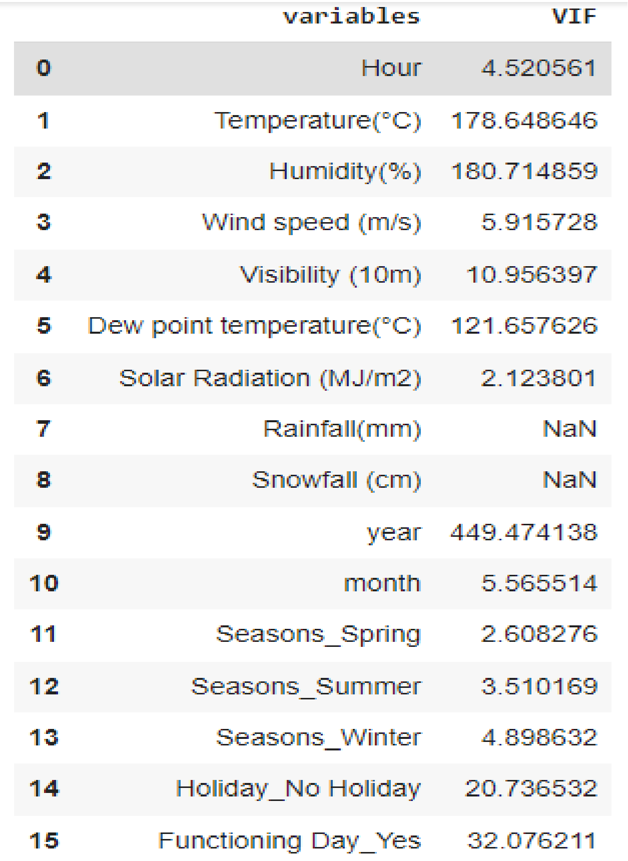
Correlation using heatmap

From the heatmap, i have observed that 'rented bike count' or target feature is showing a high correlation with 'temperature' and 'hour' and low correlation with other features. Moreover, dew point is also having high correlation with 'temperature' and 'humidity'.

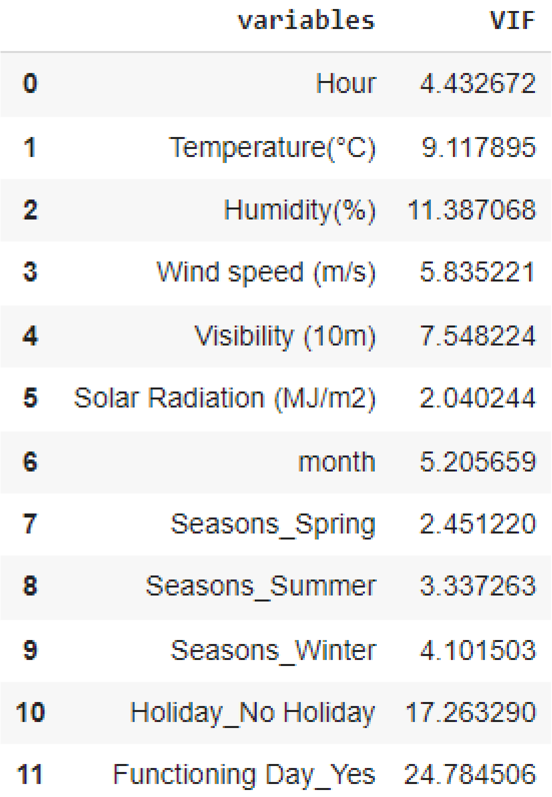


Multicollinearity

The left part is the VIF values for our dataset columns and since the vif value is very high for many columns, so there are some operations performed to compress the vif value as low as possible. I have targeted vif value less than 10 and right part showing the same.

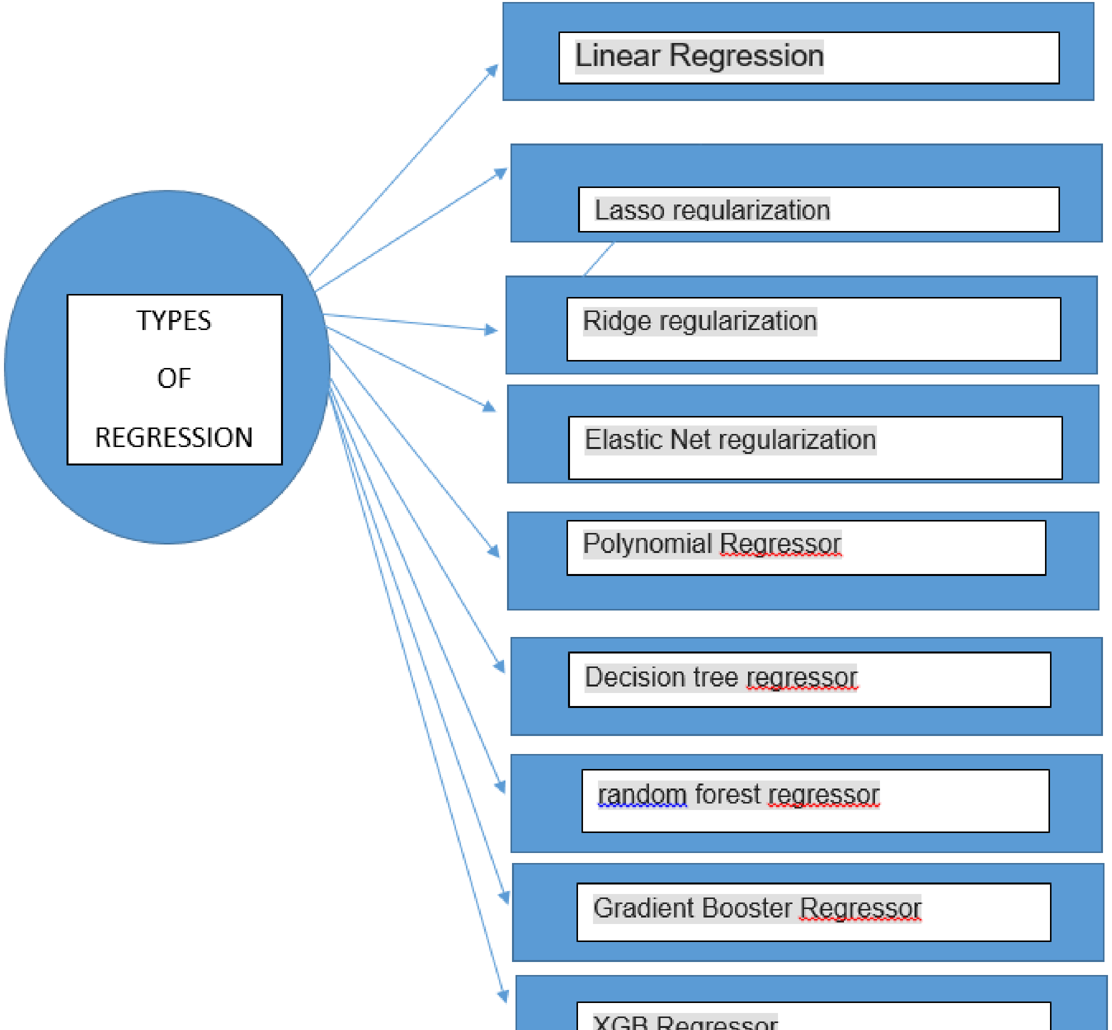


	variables	VIF
0	Hour	4.520561
1	Temperature(°C)	178.648646
2	Humidity(%)	180.714859
3	Wind speed (m/s)	5.915728
4	Visibility (10m)	10.956397
5	Dew point temperature(°C)	121.657626
6	Solar Radiation (MJ/m2)	2.123801
7	Rainfall(mm)	NaN
8	Snowfall (cm)	NaN
9	year	449.474138
10	month	5.565514
11	Seasons_Spring	2.608276
12	Seasons_Summer	3.510169
13	Seasons_Winter	4.898632
14	Holiday_No Holiday	20.736532
15	Functioning Day_Yes	32.076211



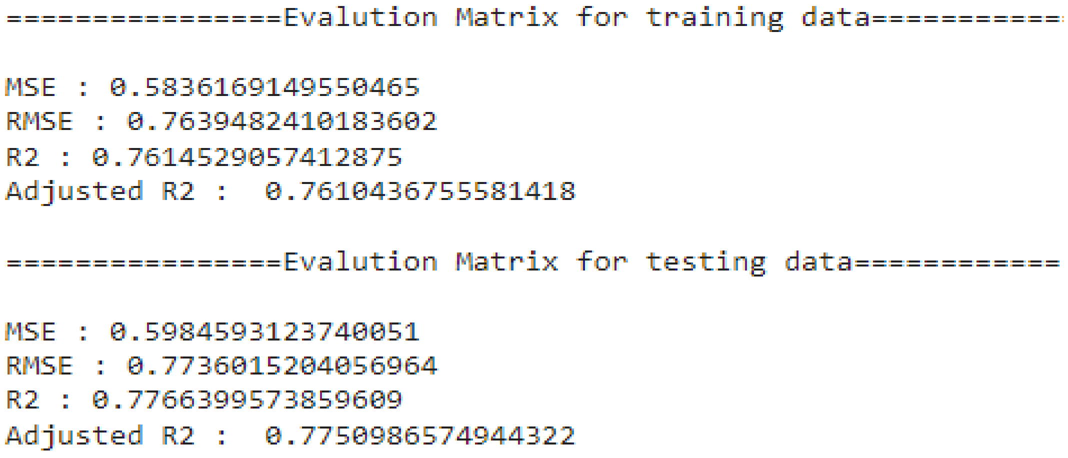
	variables	VIF
0	Hour	4.432672
1	Temperature(°C)	9.117895
2	Humidity(%)	11.387068
3	Wind speed (m/s)	5.835221
4	Visibility (10m)	7.548224
5	Solar Radiation (MJ/m2)	2.040244
6	month	5.205659
7	Seasons_Spring	2.451220
8	Seasons_Summer	3.337263
9	Seasons_Winter	4.101503
10	Holiday_No Holiday	17.263290
11	Functioning Day_Yes	24.784506

ML Models used

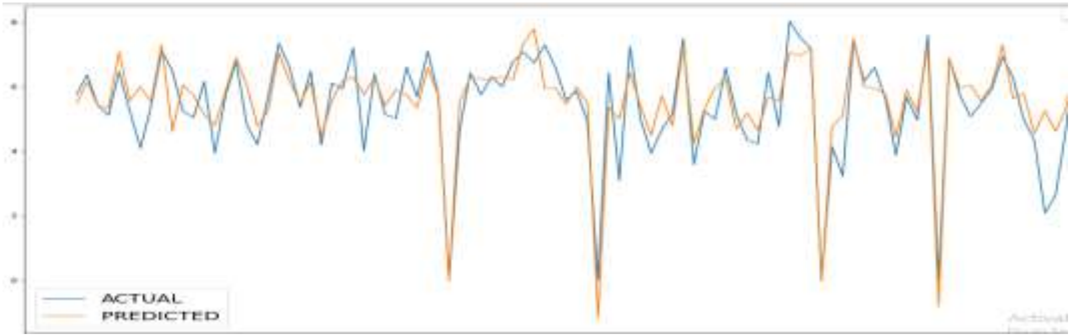


Model 1: Linear Regression

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. LR makes prediction for continuous as well as numeric variables.



```
================Evalution Matrix for training data===========

MSE : 0.583616914955046 5
RMSE : 0.763948241018360 2
R2 : 0.761452905741287 5
Adjusted R2 :  0.761043675558141 8

================Evalution Matrix for testing data============

MSE : 0.598459312374005 1
RMSE : 0.773601520405696 4
R2 : 0.776639957385960 9
Adjusted R2 :  0.775098657494432 2
```

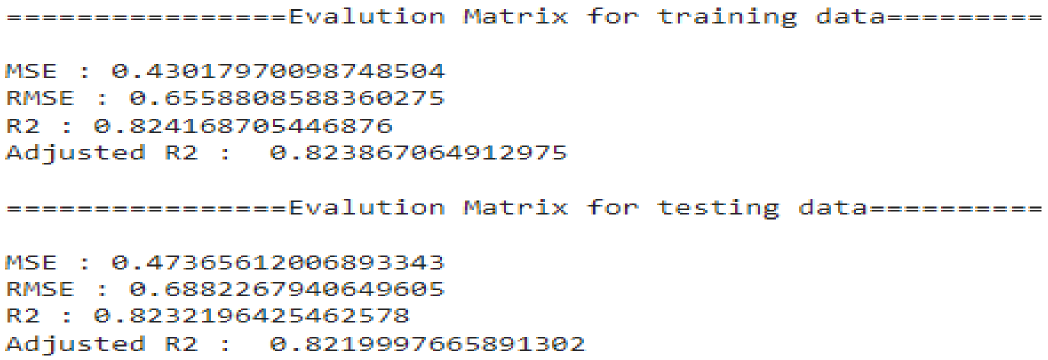


Model 2: Polynomial Regression

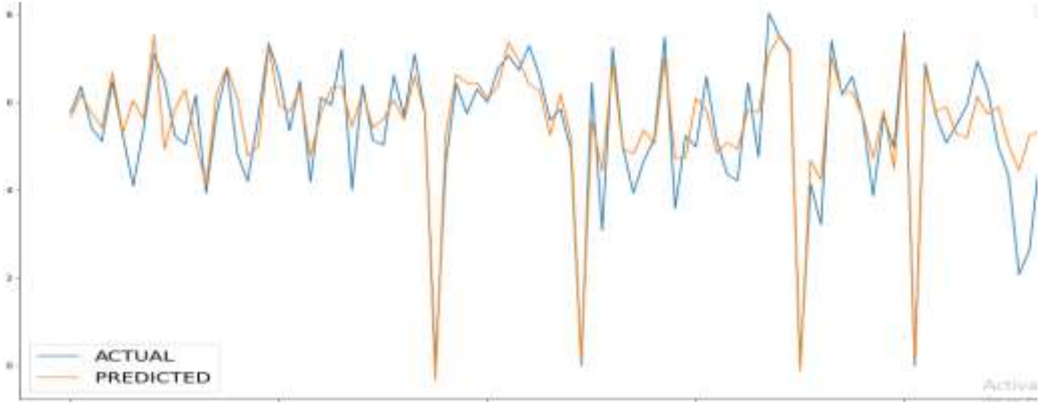
Polynomial Regression is a regression algorithm that models the relationship between a dependent and independent variable as nth degree polynomial. The following equation defines a Polynomial Regression equation:

$y=a0+a1X+a2X2+\cdots+anXn$

When data points are arrange in non linear fashion, we need the Polynomial Regression model.

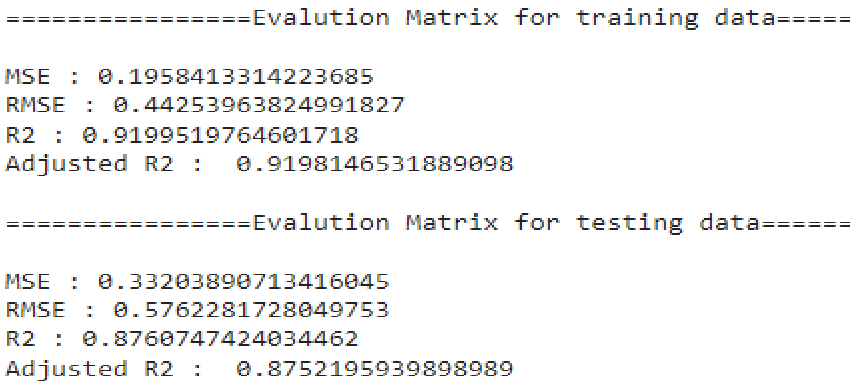


```
================Evalution Matrix for training data=========

MSE : 0.43017970098748504
RMSE : 0.6558808588360275
R2 : 0.824168705446876
Adjusted R2 :  0.823867064912975

================Evalution Matrix for testing data==========

MSE : 0.47365612006893343
RMSE : 0.6882267940649605
R2 : 0.8232196425462578
Adjusted R2 :  0.8219997665891302
```

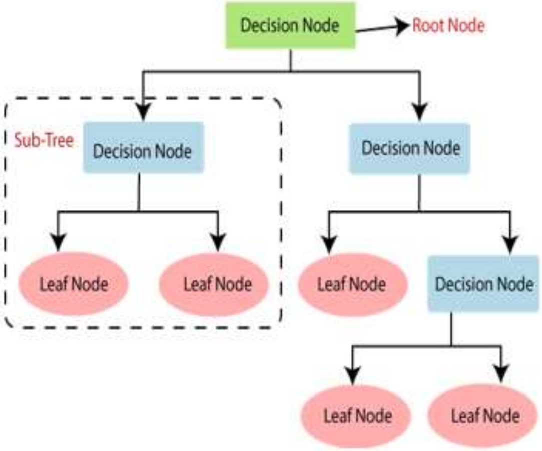


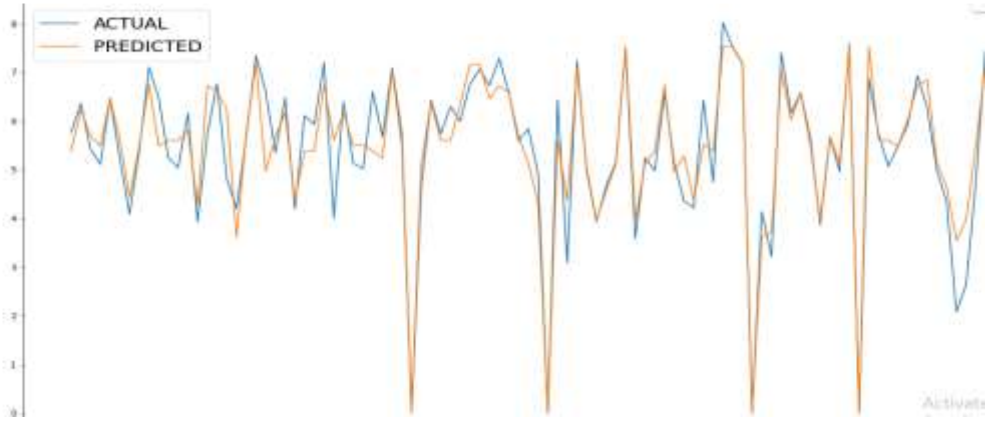
Model 3: Decision Tree

Decision Tree is a supervised learning method used in data mining for classification and regression methods. It is a tree that helps us in decision-making purposes. It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed. The final tree is a tree with the decision nodes and leaf nodes.



```
================Evalution Matrix for training data=====

MSE : 0.1958413314223685
RMSE : 0.44253963824991827
R2 : 0.9199519764601718
Adjusted R2 :  0.9198146531889098

================Evalution Matrix for testing data======

MSE : 0.33203890713416045
RMSE : 0.5762281728049753
R2 : 0.8760747424034462
Adjusted R2 :  0.8752195939898989
```

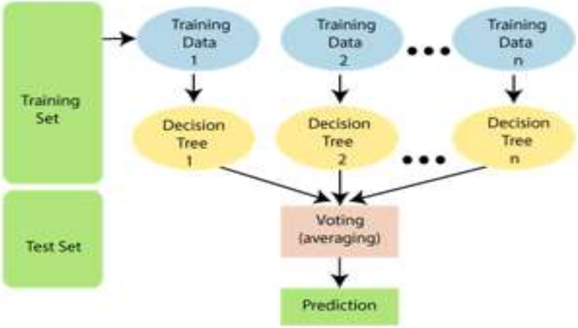


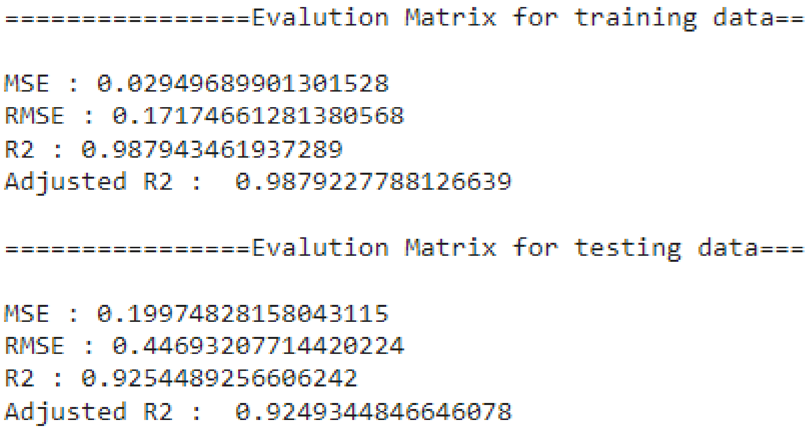


Model 4: Random Forest Regressor

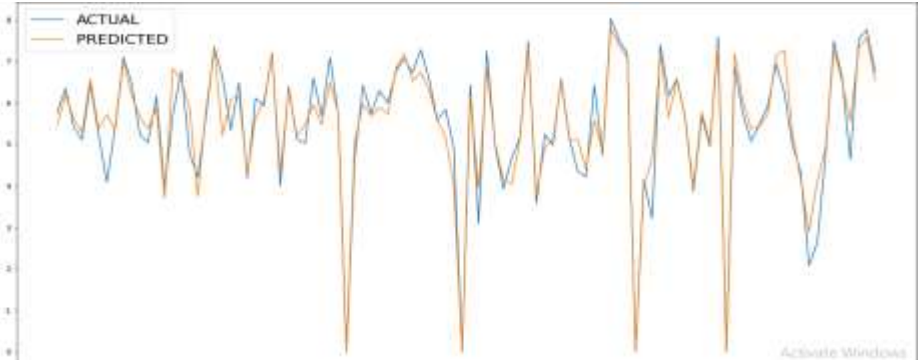
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



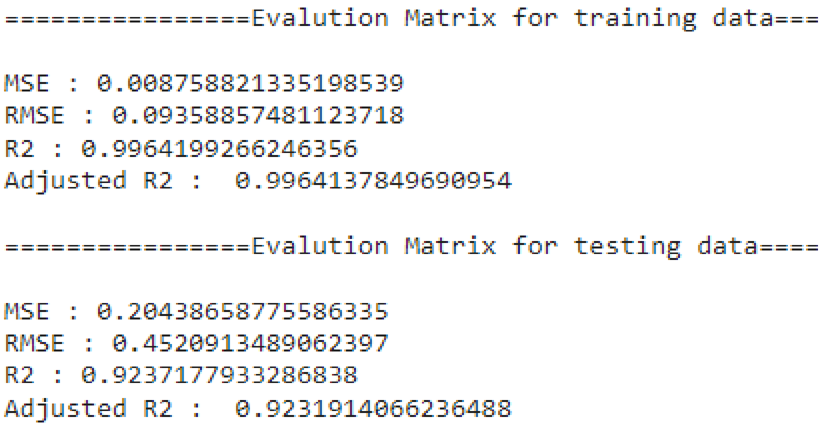


```
================Evalution Matrix for training data==

MSE : 0.02949689901301528
RMSE : 0.17174661281380568
R2 : 0.987943461937289
Adjusted R2 :  0.9879227788126639

================Evalution Matrix for testing data===

MSE : 0.19974828158043115
RMSE : 0.44693207714420224
R2 : 0.9254489256606242
Adjusted R2 :  0.9249344846646078
```

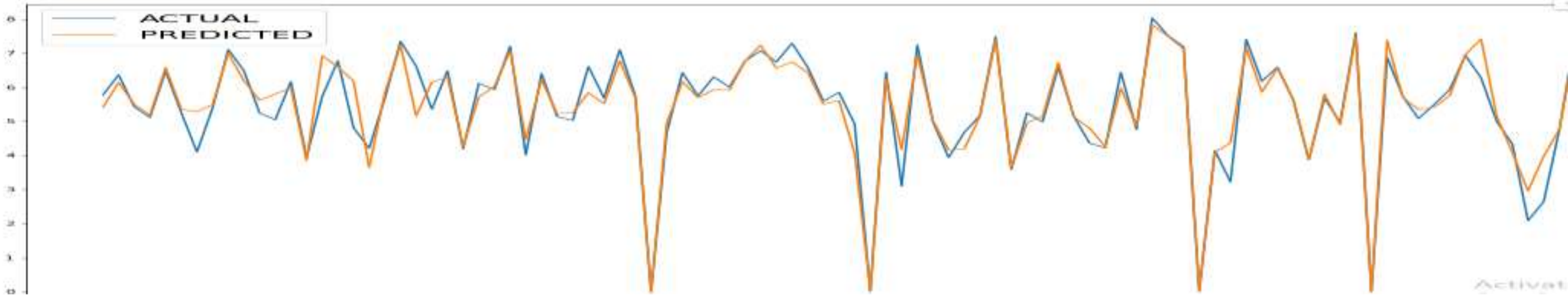


Model 5: Gradient Booster

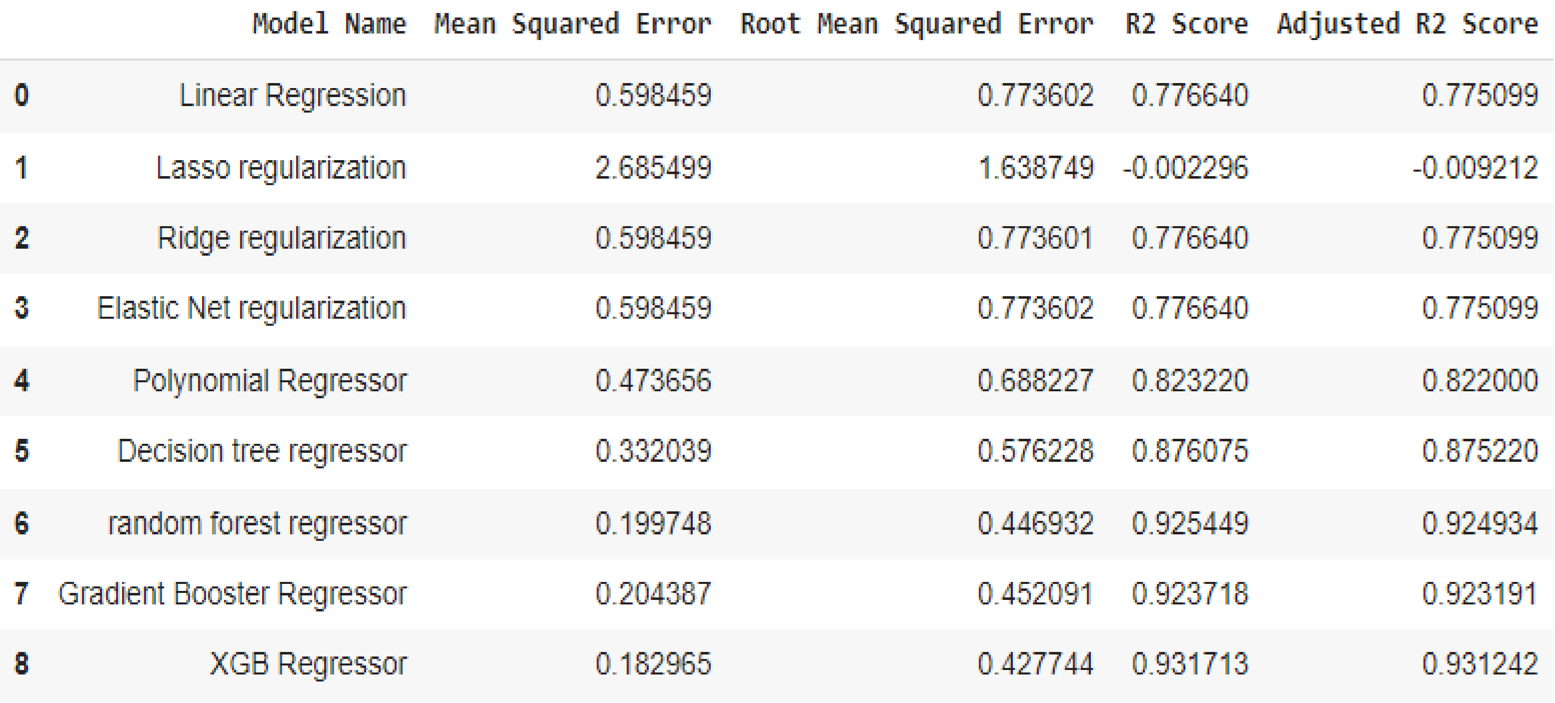
Gradient boosting is one of the most popular machine learning algorithms for tabular datasets. It is powerful enough to find any nonlinear relationship between your model target and features and has great usability that can deal with missing values, outliers, and high cardinality categorical values on your features without any special treatment.



```
================Evalution Matrix for training data===

MSE : 0.008758821335198539
RMSE : 0.09358857481123718
R2 : 0.9964199266246356
Adjusted R2 :  0.9964137849690954

================Evalution Matrix for testing data====

MSE : 0.20438658775586335
RMSE : 0.4520913489062397
R2 : 0.9237177933286838
Adjusted R2 :  0.9231914066236488
```



Model Summary



	Model Name	Mean Squared Error	Root Mean Squared Error	R2 Score	Adjusted R2 Score
0	Linear Regression	0.598459	0.773602	0.776640	0.775099
1	Lasso regularization	2.685499	1.638749	-0.002296	-0.009212
2	Ridge regularization	0.598459	0.773601	0.776640	0.775099
3	Elastic Net regularization	0.598459	0.773602	0.776640	0.775099
4	Polynomial Regressor	0.473656	0.688227	0.823220	0.822000
5	Decision tree regressor	0.332039	0.576228	0.876075	0.875220
6	random forest regressor	0.199748	0.446932	0.925449	0.924934
7	Gradient Booster Regressor	0.204387	0.452091	0.923718	0.923191
8	XGB Regressor	0.182965	0.427744	0.931713	0.931242

Final result

After fitting so many regressor models into the dataset,, if I consider low RMSE and high R2 score, i got the best results from XGB Regressor and then from random forest and Gradient Booster regressor.
Therefore, We can use either XGB regressor or random forest model for the bike rental stations.

THANK YOU