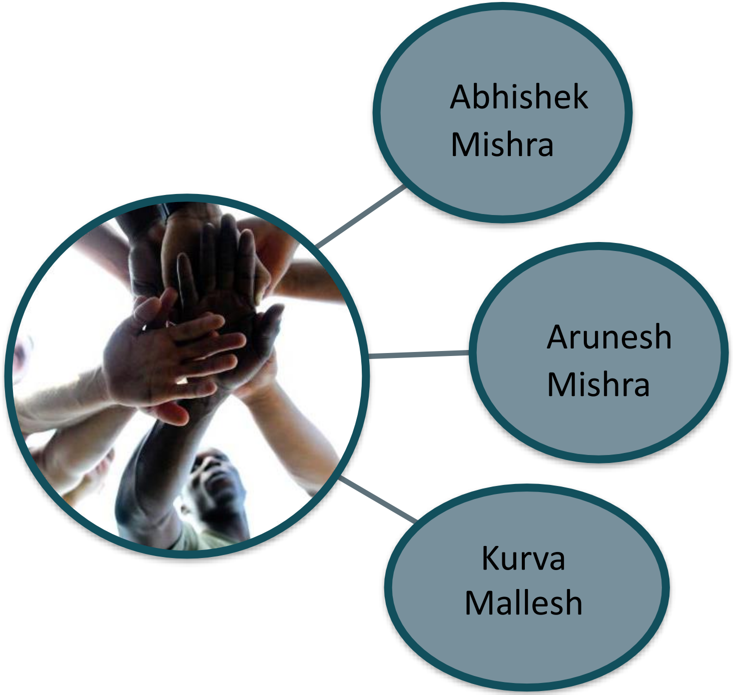
Capstone Project Regression

Seoul Bike Sharing

TEAM MEMBERS



Abhishek Mishra

Arunesh Mishra

Kurva Mallesh

WHAT IS Seoul Bike Sharing

SEOUL -- Seoul's public bike rental service called "Ttareungi," also known as Seoul Bike, has become one of the capital city's most popular public transport systems by being used more than 100 million times in about six and a half years since the service was launched in December 2015. The public bicycle rental system is favored by Seoulites who wish to travel short distances of a few kilometers instead of using crowded buses or subway trains.

HOW IT WORKS

Seoul Bike has rental stations around the entire city of Seoul. This bike rental system is perfect for riders looking for short-term, hour long trips from point A to point B in Seoul. For this one the bike rentals drop off and pickup location can be found almost anywhere using the map on the website or app.

Objective of The Project

The main objective of this project is to find the business insider from the data. Also create the model which can predict Seoul bike sharing demand.

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.

Business Problems

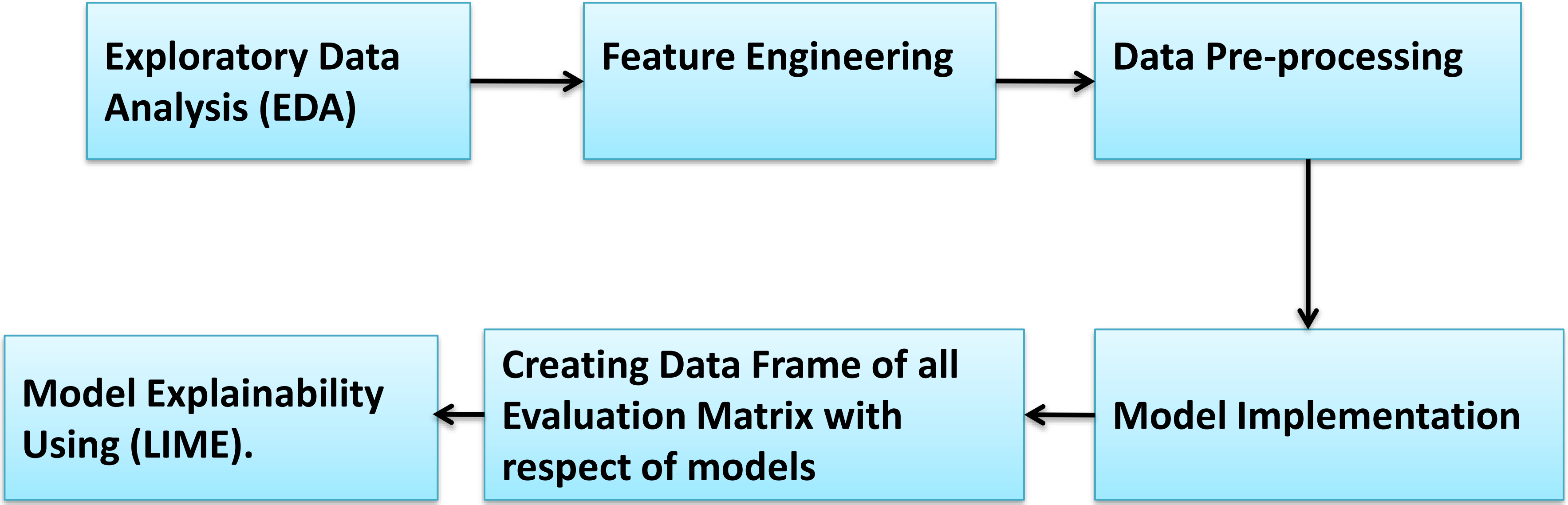
What can we learn about seasons and rented bike count?

What can we learn about holiday seasons and rented bike count?

What can we learn about seasons humidity and bike count?

What can we learn about temperature, rented bike count for different seasons?

Road map of the project



Exploratory Data Analysis (EDA)

Feature Engineering

Data Pre-processing

Model Implementation

Creating Data Frame of all Evaluation Matrix with respect of models

Model Explainability Using (LIME).

Exploratory Data Analysis (EDA)

- ✓ This data does not contain null values. Also we have four object columns and one date column.
- ✓ Seasons making huge impact on bike rentals where in summer there was very high count and very low in winter.
- ✓ During Holiday people are more like to book the bike in Autumn season.
- ✓ Temperature making huge impact on Rented Bike Count. Because the count is very low at temperature less than -10 (°C).
- ✓ At 8 am and 6pm there are very high bookings.
- ✓ Rainfall also making huge impact on Rented bike count there are very less Rented bike count on greater then 5(mm) Rainfall.

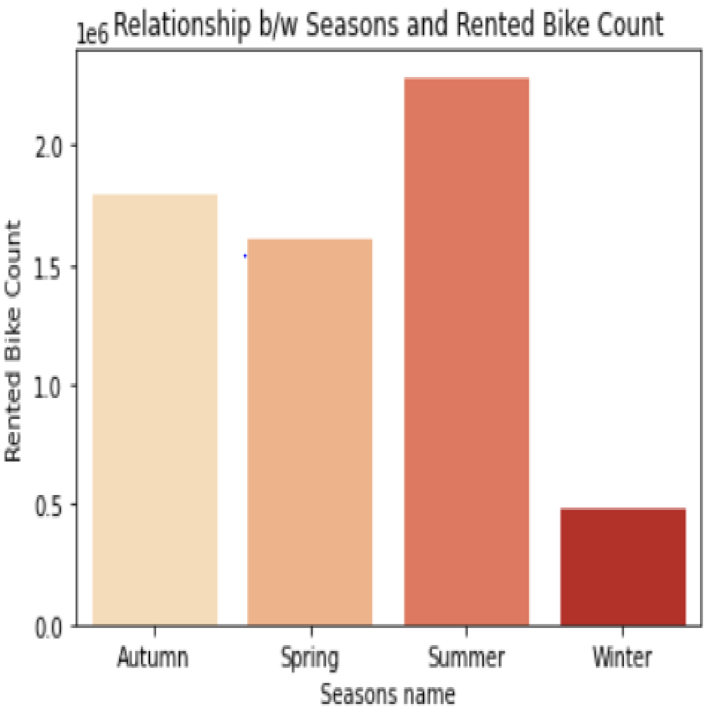
Feature Engineering

- ✓ Encoding on date column, Split month from the date.
- ✓ Creating new feature week day and weekend.
- ✓ Do some feature engineering on hour column and make them as evening, morning, noon, night so the model can easily understand.
- ✓ Used correlation plot and variance inflation factor(VIF) for features selections.
- ✓ In data pre-processing we used power transformer, minimum, maximum scaler for standardize the data.

Model Implementation

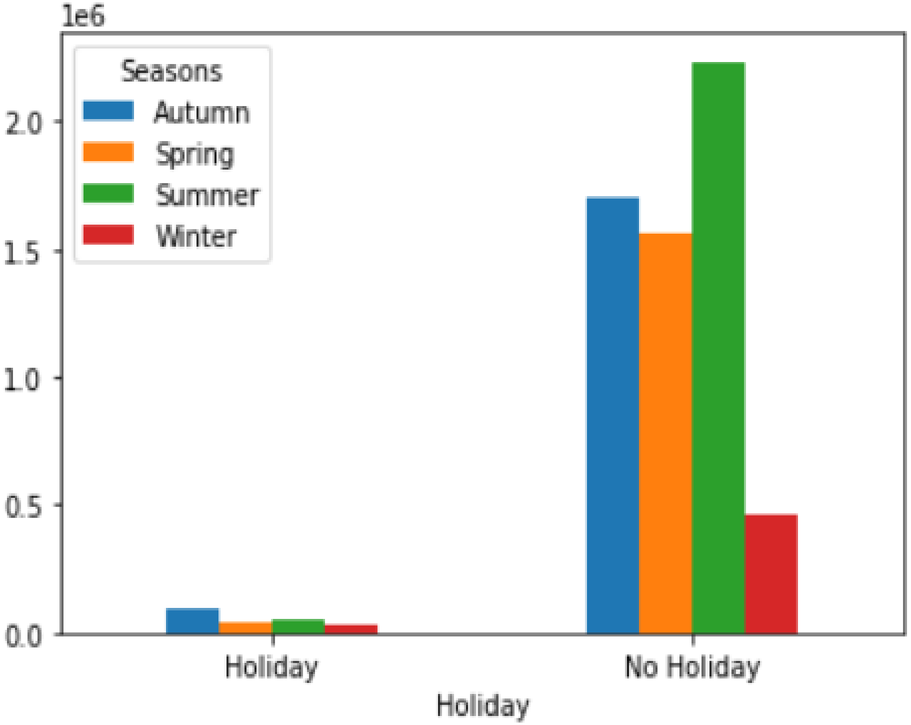
- Used Linear Regression, Lasso with (hyperparameter tuning), Ridge with (hyperparameter tuning).
- Used Polynomial with Linear Regression.
- Also used some Tree Base Models like Decision Tree, Random Forest Regressor.
- We used some Boosting Regression like Gradient Boosting Regression, Adaboost Boost Regressor and XGBoost Regression.
- Also use KNN just for quarticity.

Rented Bike Count vs Seasons plot



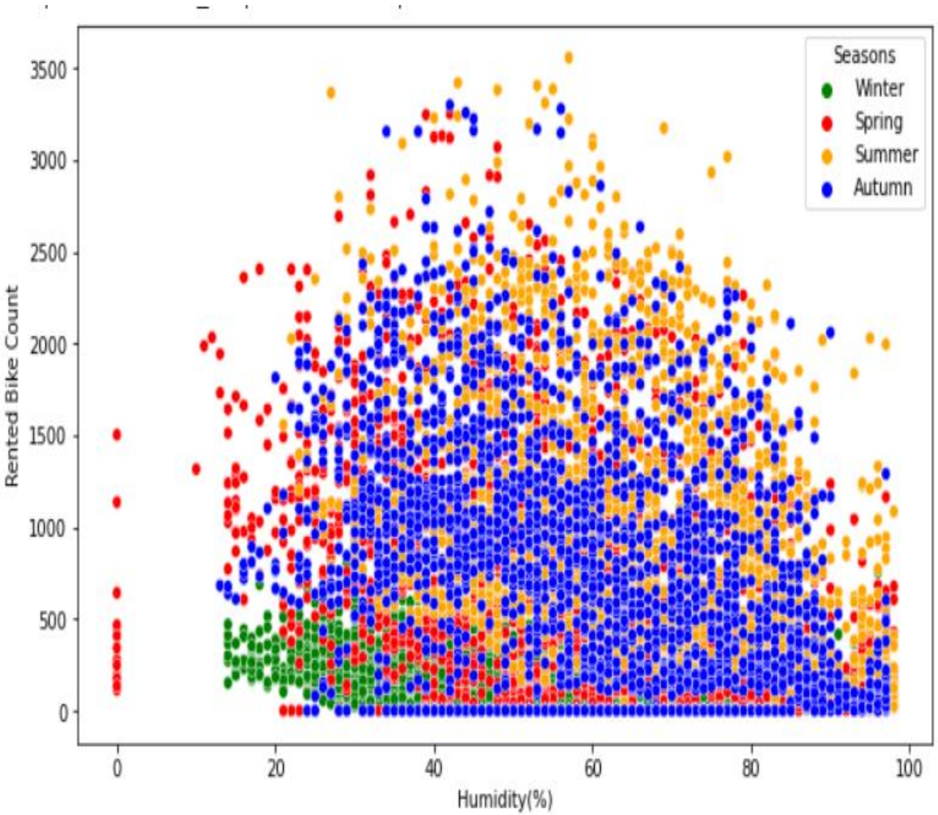
1. According to graph we can say that in "Summer" there are 2283234 rented bike bookings happened
2. We can see that in "Autumn“ season there are 1790002 rented bike bookings happened.
3. And we can say that in "Spring" there are 1611909 rented bike bookings happened.
4. According to graph we can say that in "Winter" there are 487169 rented bike bookings happened

Rented Bike Count vs Holiday plot



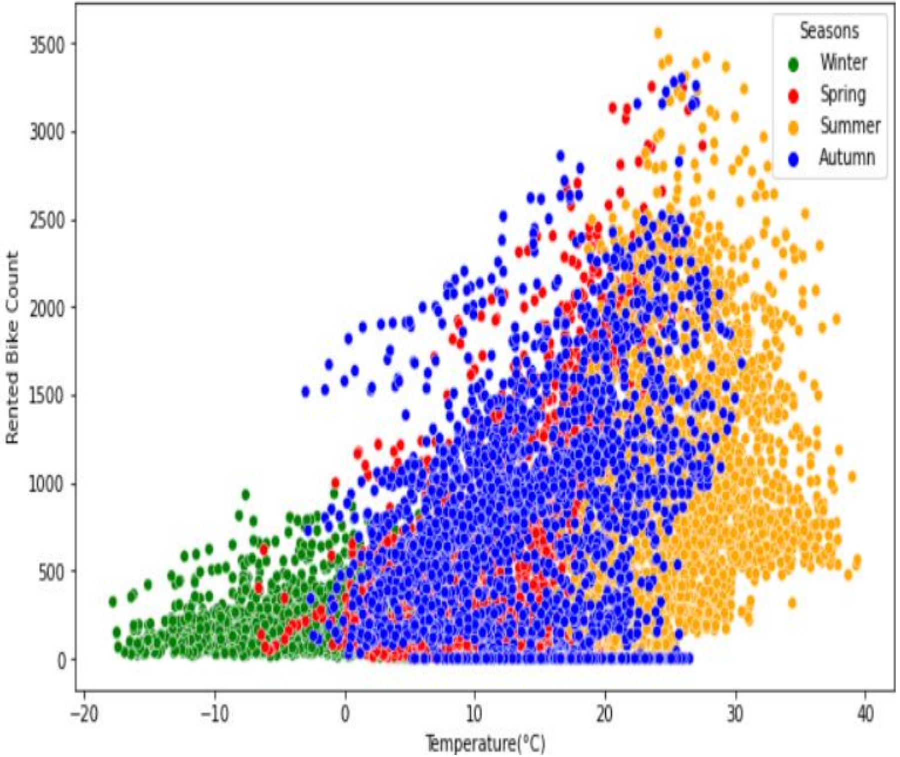
1. According to the graph we can say in the Non Holiday time there are 5956419 bikes rented count
2. In Holiday only 215895 counts happened.
3. Again we can understand that Seasons is impacting on Holiday.
4. Normally we know there are high bike bookings in Summer.
5. But during holiday time in Autumn season there was very high rented bike count.

Rented Bike Count vs Humidity plot



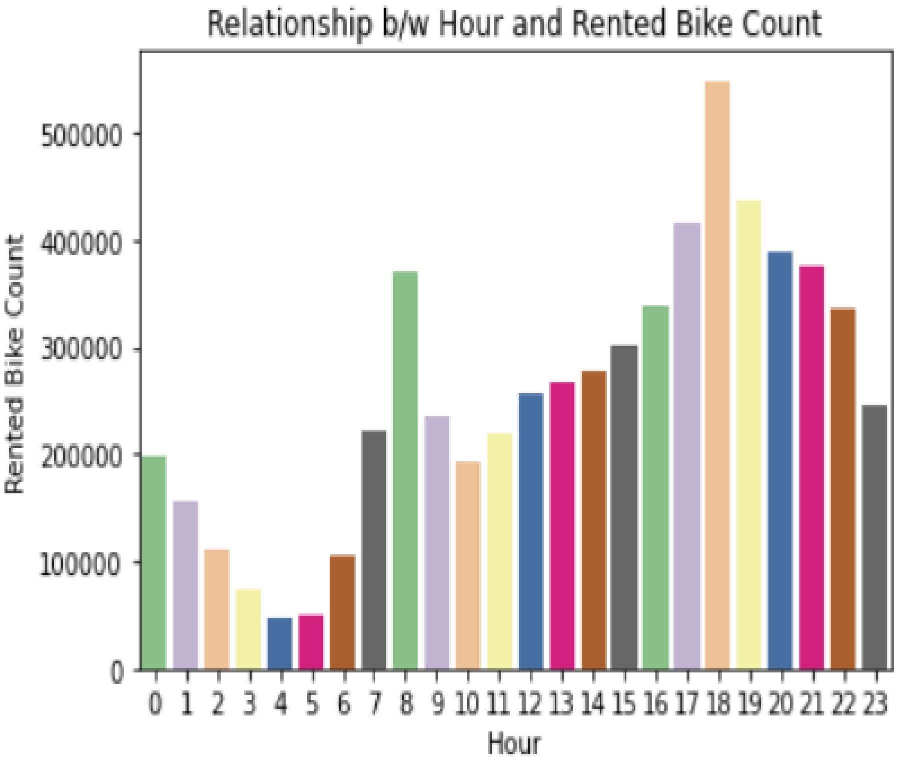
1. According to bar plot Humidity is not making huge impact on Rented Bike Count.
2. We found there are some rows that have Humidity as 0, it is not possible so we need to check this.

Rented Bike Count vs Temperature Plot



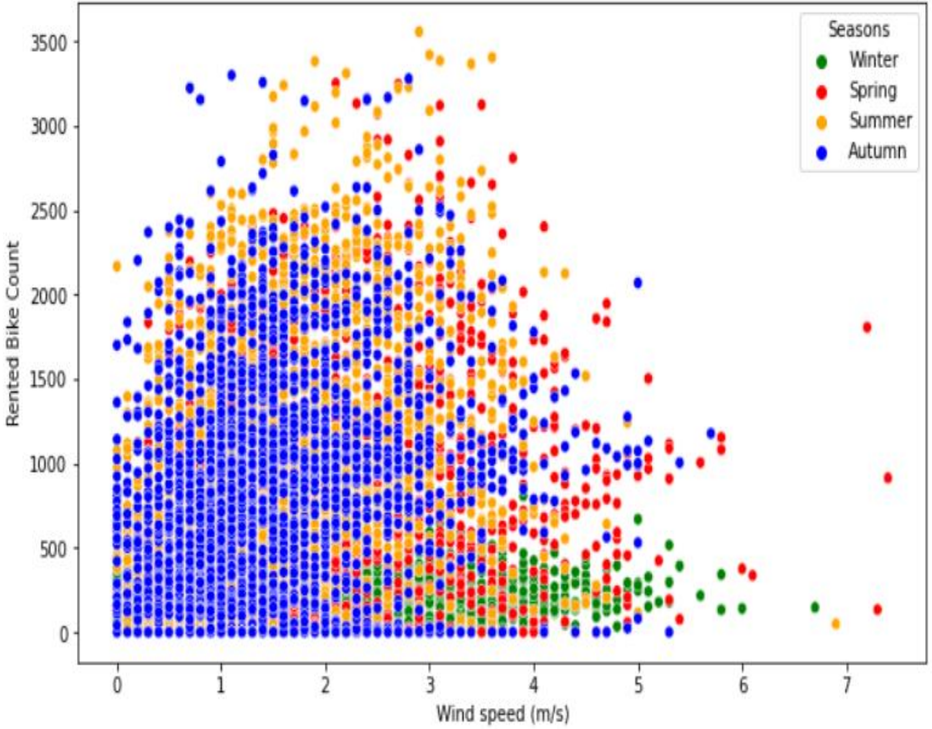
1. According to scatter plot Rented Bike Count is very low in less then -10 Temperature(°C)
2. In Winter **Seasons** Temperature goes min -17.8°C
3. This type of condition comes in winter season. So we can suggest the company in the winter season to check the Temperature(°C) for each day and plan accordingly.
4. The temperature(°C)less then -10 have impact on Rented Bike Count

Rented Bike Count vs Hour plot



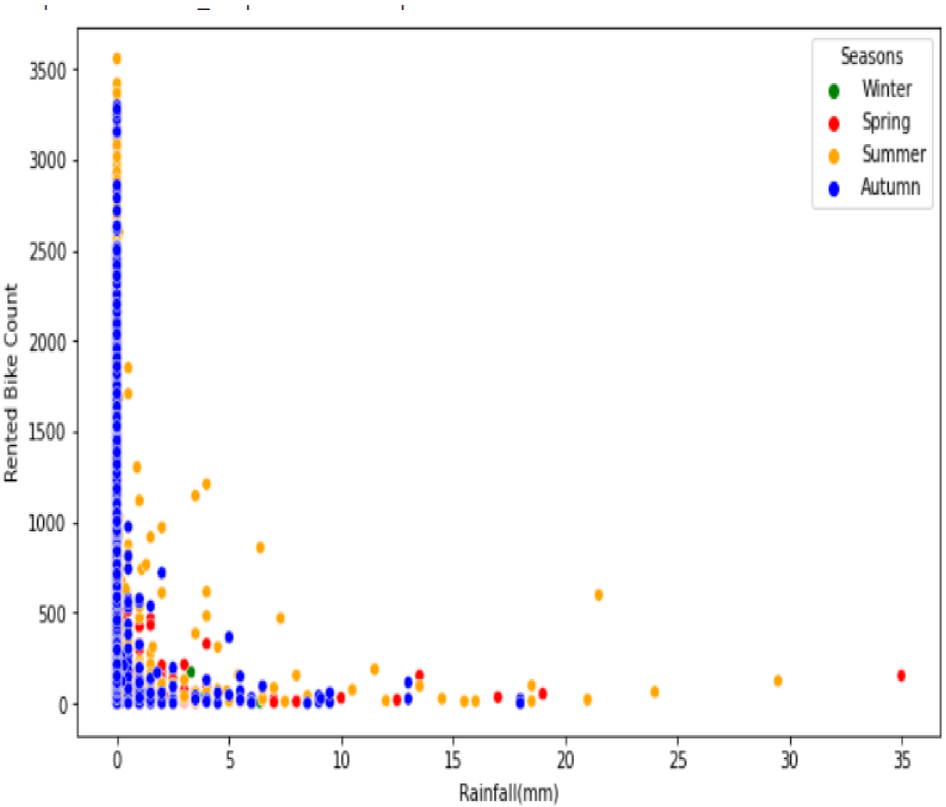
1. According to the graph we can observe that at 8am there are good amount of bookings.
2. At **6pm** there are **very high** bookings.
3. We can suggest the company to give high preference to 8pm
4. Also we can see there are very less bookings b/w 3pm - 6pm

Rented Bike Count vs Wind speed plot



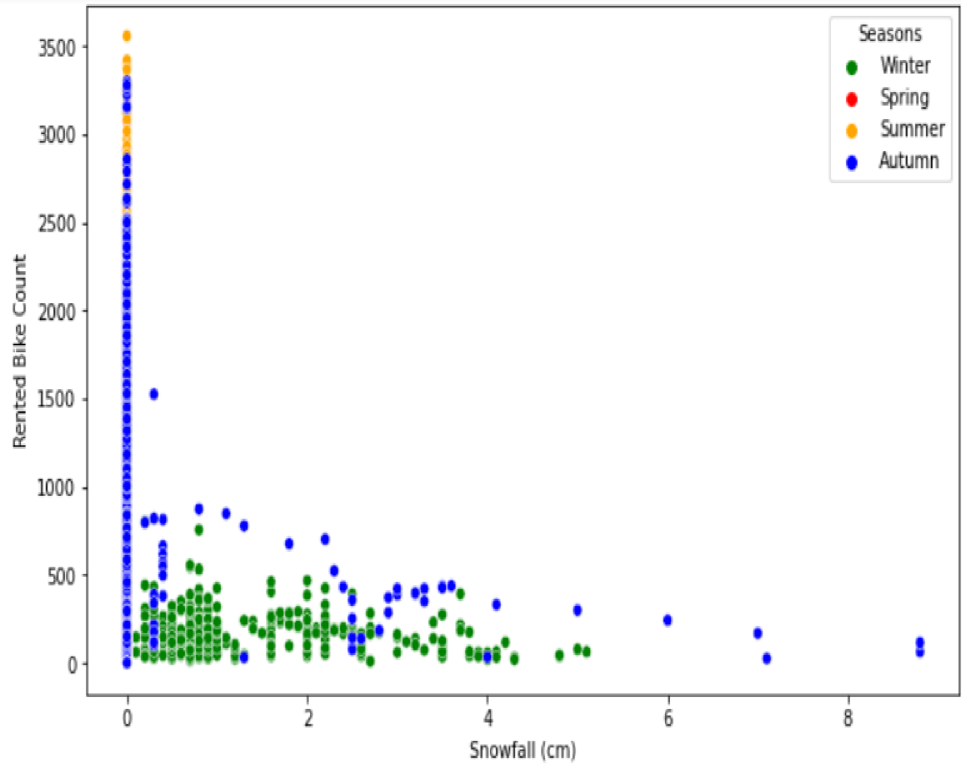
1. We can see there are not much bike rented count after Wind speed 5(m/s).
2. According to the graph in Spring Seasons Wind speed goes till 8(m/s).
3. So we can say this type of condition meets one or two times in a year.

Rented Bike Count vs Rain fall plot



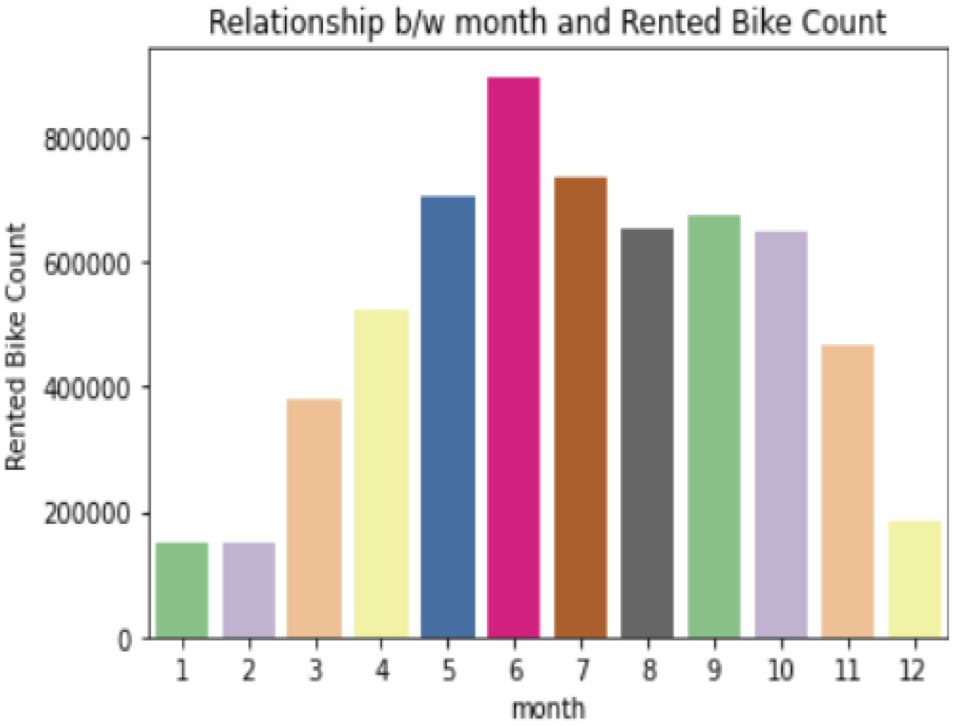
1. According to scatter plot we can say that there are very less Rented bike count when Rainfall greater than 5(mm).
2. Also we can see that there is very less possibility of more than 5(mm) Rain fall in the Winter season.

Rented Bike Count vs Snow fall plot



1. According to scatter plot we can see that Snowfall happens only in Winter and Autumn seasons.
2. we can see Snowfall directly impacting on Rented Bike Count.
3. And there is no Snowfall in Summer and Spring seasons.

Rented Bike Count vs Month plot



- According to bar plot we can say that Rented Bike Count is very high in the month of June.

EDA Conclusion

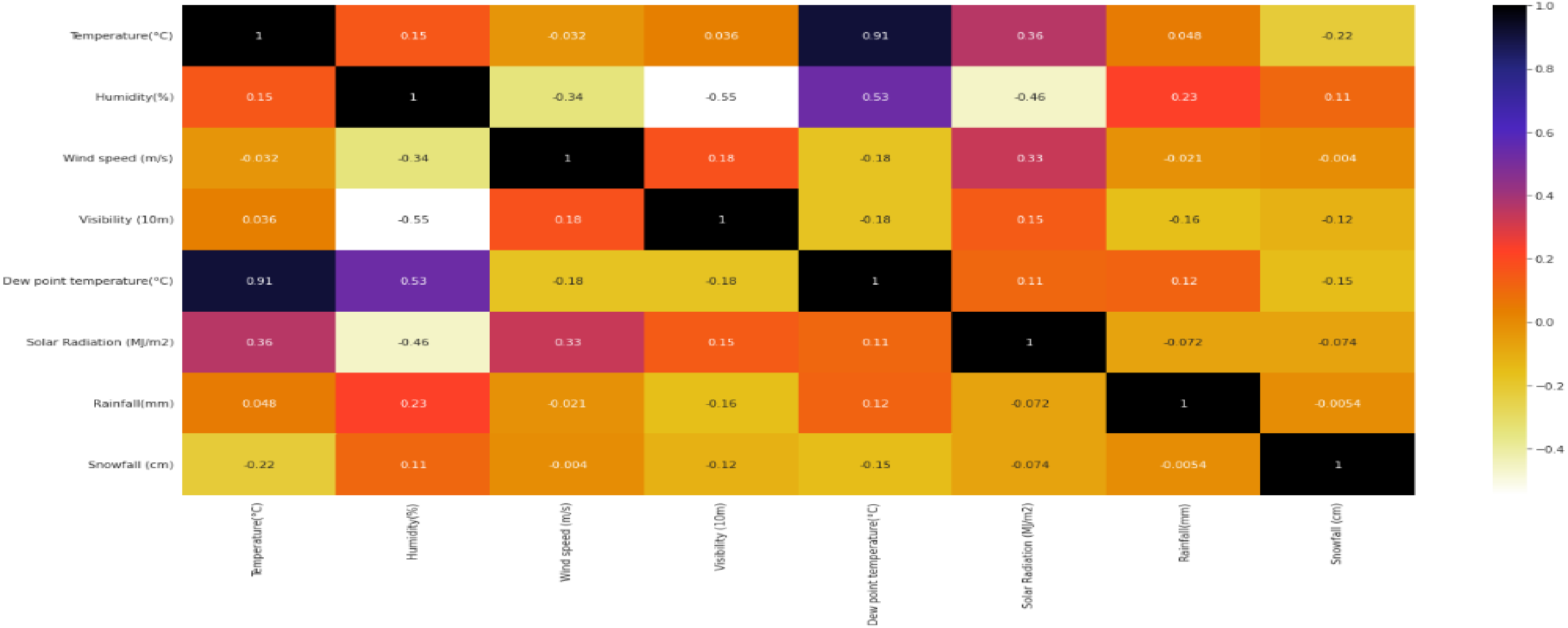
- ✓ Seasons are making huge impact on Rented Bike count. There was very high count in summer (2283234) and very low count in Winter (487169) season.
- ✓ We knew that there is high Rented Bike Count in Summer but during Holiday people like to book the bikes in Autumn season more. So we can say Autumn season is best for Holidays.
- ✓ We found there are some rows that have Humidity as 0, so it is not possible. We need to check this out.
- ✓ The Temperature less than -10 (°C) making huge impacting on Rented Bike Count
- ✓ We suggest the company to give high preference to 8pm. Because at that time Rented Bike Count is very high.
- ✓ We can say that there is not much bike rented count after Wind speed 5(m/s)
- ✓ We can say that there is very less Rented bike count when Rainfall is greater than 5(mm).

Feature Engineering

- Used Histogram plot, Pair plot to understand the data.
- Encoding on Date Column Creating new feature.
- Done some feature engineering on Hour column.

Features Selections

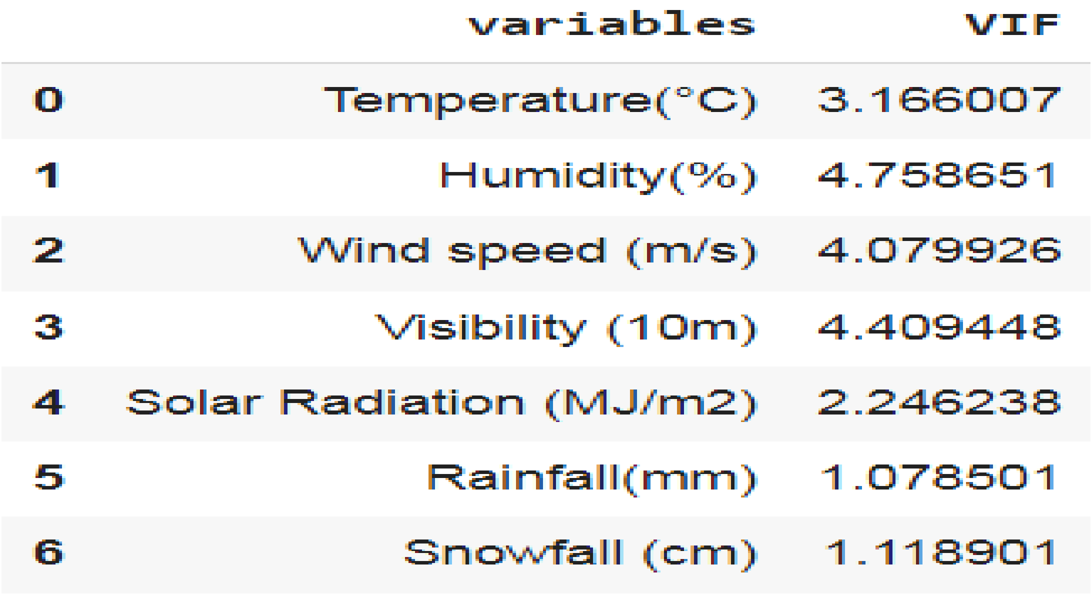
After using correlation plot we find "Dew point temperature(°C)" variable is highly correlated



	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)
Temperature(°C)	1	0.15	-0.032	0.036	0.91	0.36	0.048	-0.22
Humidity(%)	0.15	1	-0.34	-0.55	0.53	-0.46	0.23	0.11
Wind speed (m/s)	-0.032	-0.34	1	0.18	-0.18	0.33	-0.021	-0.004
Visibility (10m)	0.036	-0.55	0.18	1	-0.18	0.15	-0.16	-0.12
Dew point temperature(°C)	0.91	0.53	-0.18	-0.18	1	0.11	0.12	-0.15
Solar Radiation (MJ/m2)	0.36	-0.46	0.33	0.15	0.11	1	-0.072	-0.074
Rainfall(mm)	0.048	0.23	-0.021	-0.16	0.12	-0.072	1	-0.0054
Snowfall (cm)	-0.22	0.11	-0.004	-0.12	-0.15	-0.074	-0.0054	1

Features Selections

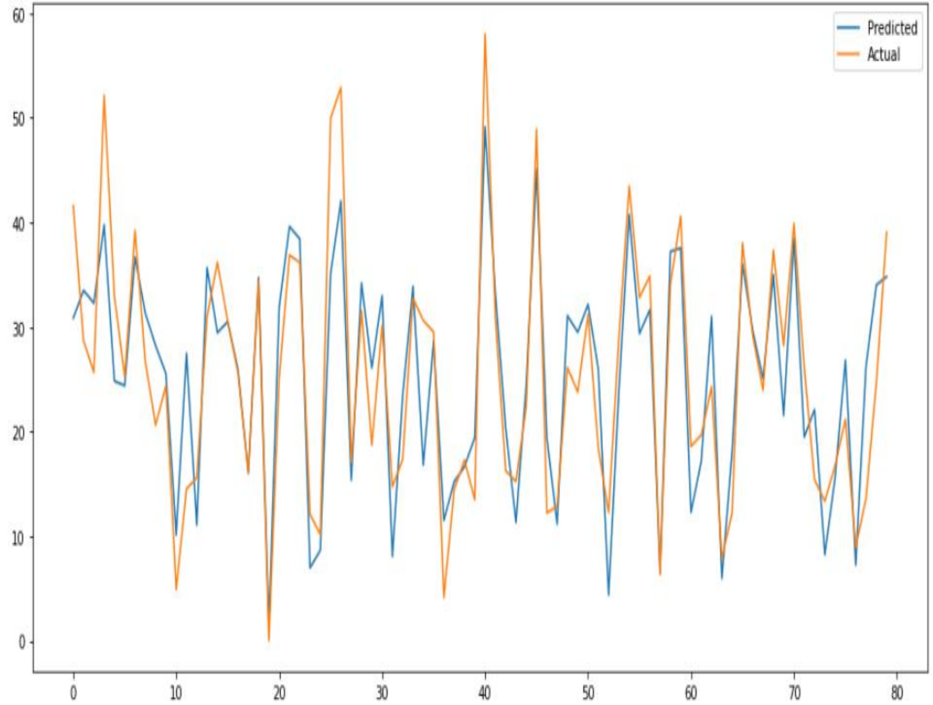
So according to VIF there is 'Dew point temperature(°C)' is multicollinearity.

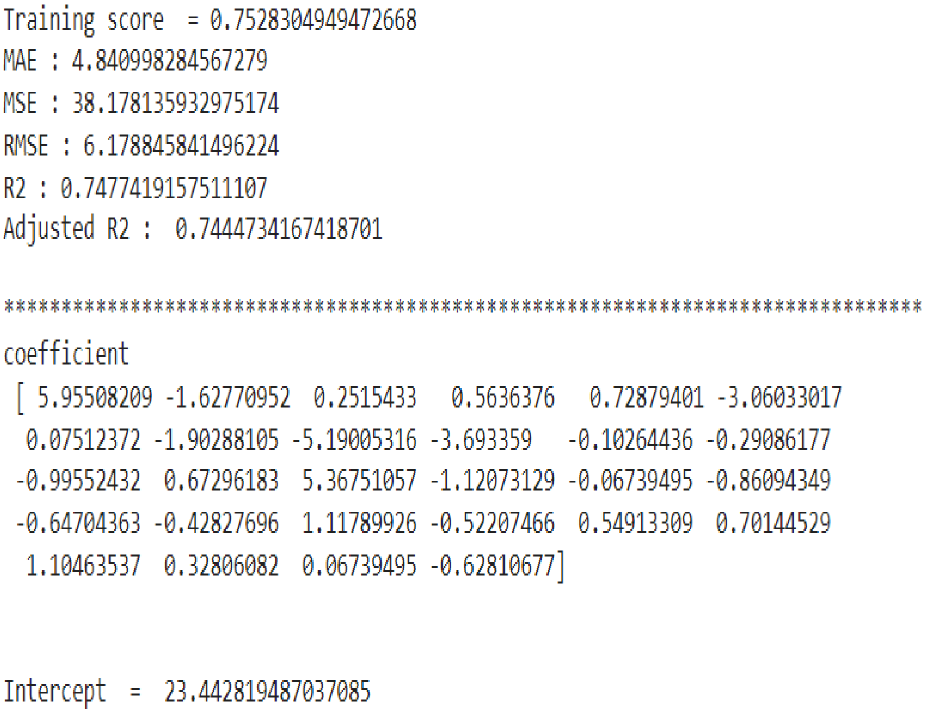


	variables	VIF
0	Temperature(°C)	3.166007
1	Humidity(%)	4.758651
2	Wind speed (m/s)	4.079926
3	Visibility (10m)	4.409448
4	Solar Radiation (MJ/m2)	2.246238
5	Rainfall(mm)	1.078501
6	Snowfall (cm)	1.118901

Linear Regression

```
******************* ploting the graph of Actual and predicted only with 80 observation *******************
```

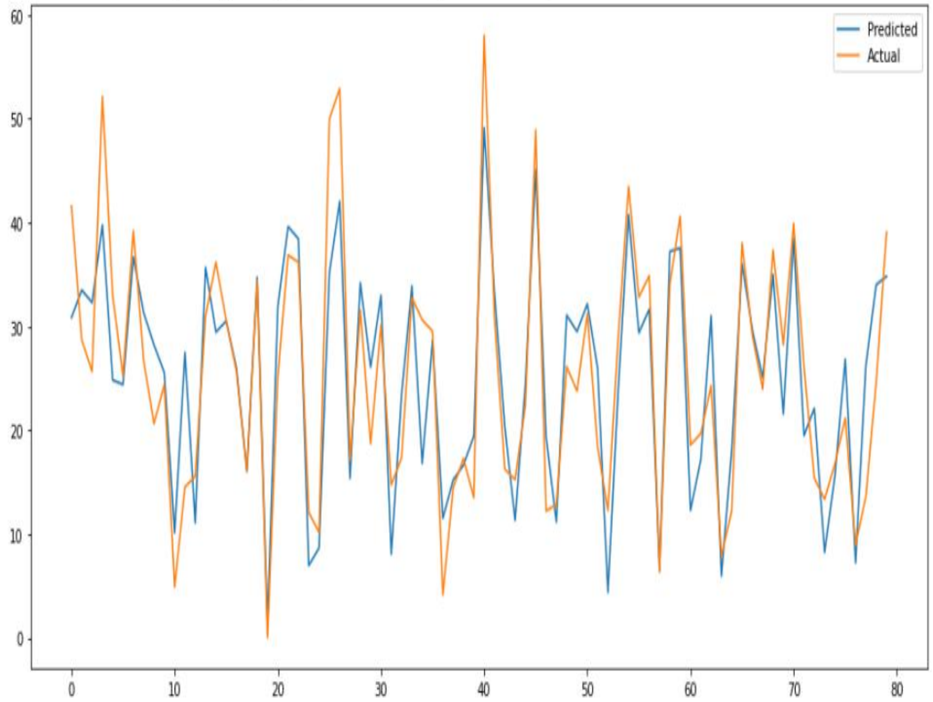


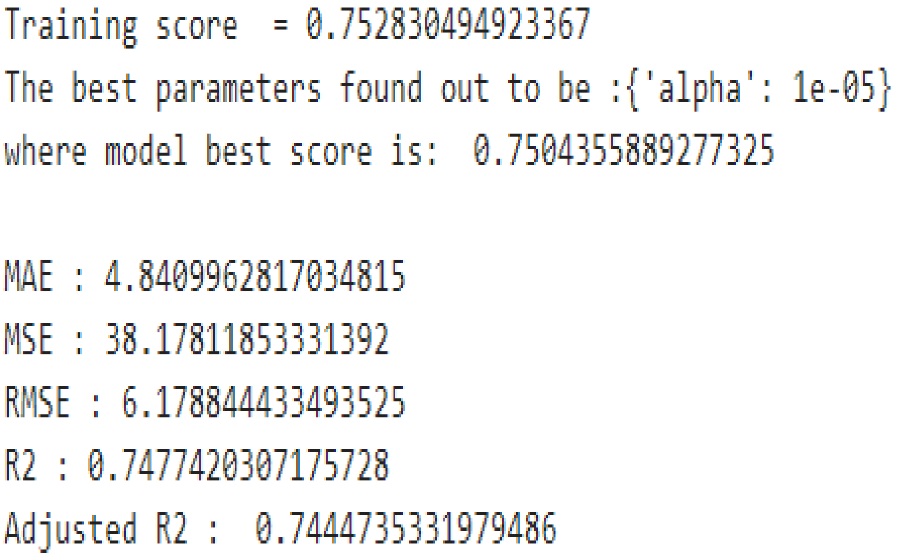


```
Training score  = 0.7528304949472668
MAE : 4.840998284567279
MSE : 38.178135932975174
RMSE : 6.178845841496224
R2 : 0.7477419157511107
Adjusted R2 :  0.7444734167418701

*************************************************************************************

coefficient
 [ 5.95508209 -1.62770952  0.2515433   0.5636376   0.72879401 -3.06033017
  0.07512372 -1.90288105 -5.19005316 -3.693359   -0.10264436 -0.29086177
 -0.99552432  0.67296183  5.36751057 -1.12073129 -0.06739495 -0.86094349
 -0.64704363 -0.42827696  1.11789926 -0.52207466  0.54913309  0.70144529
  1.10463537  0.32806082  0.06739495 -0.62810677]


Intercept  =  23.442819487037085
```

Lasso (hyper parameter tuning)

```
******************** ploting the graph of Actual and predicted only with 80 observation ********************
```

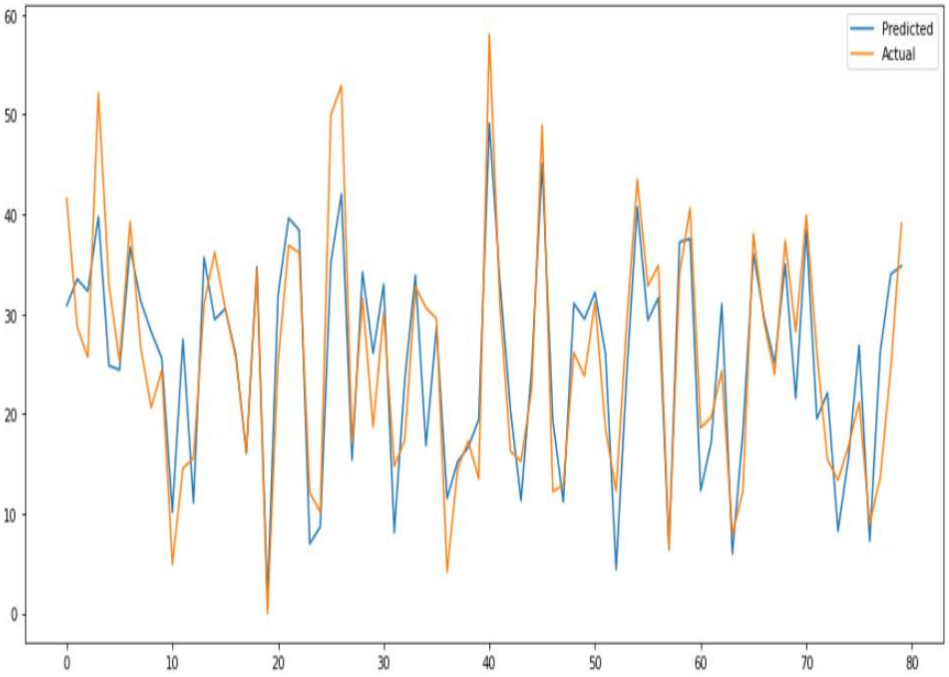


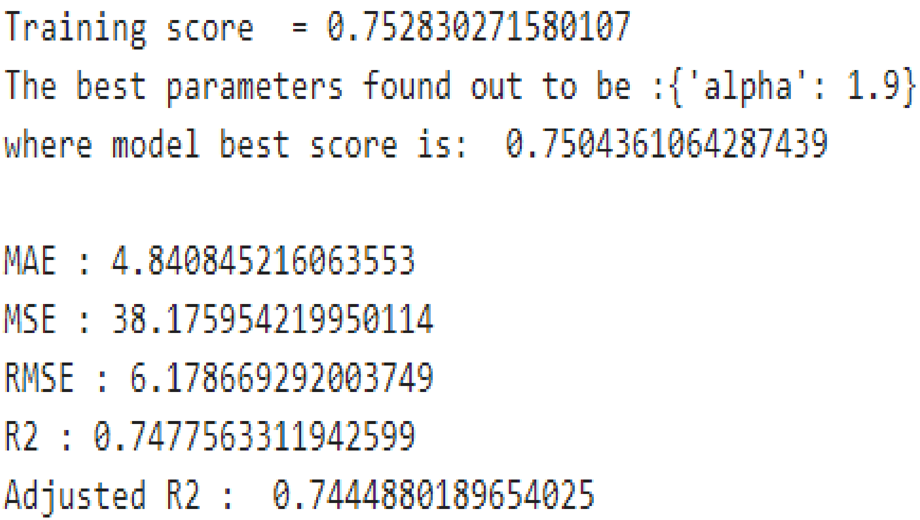


```
Training score  = 0.752830494923367
The best parameters found out to be :{'alpha': 1e-05}
where model best score is:  0.7504355889277325


MAE : 4.8409962817034815
MSE : 38.17811853331392
RMSE : 6.178844433493525
R2 : 0.7477420307175728
Adjusted R2 :  0.7444735331979486
```

Ridge (hyper parameter tuning)

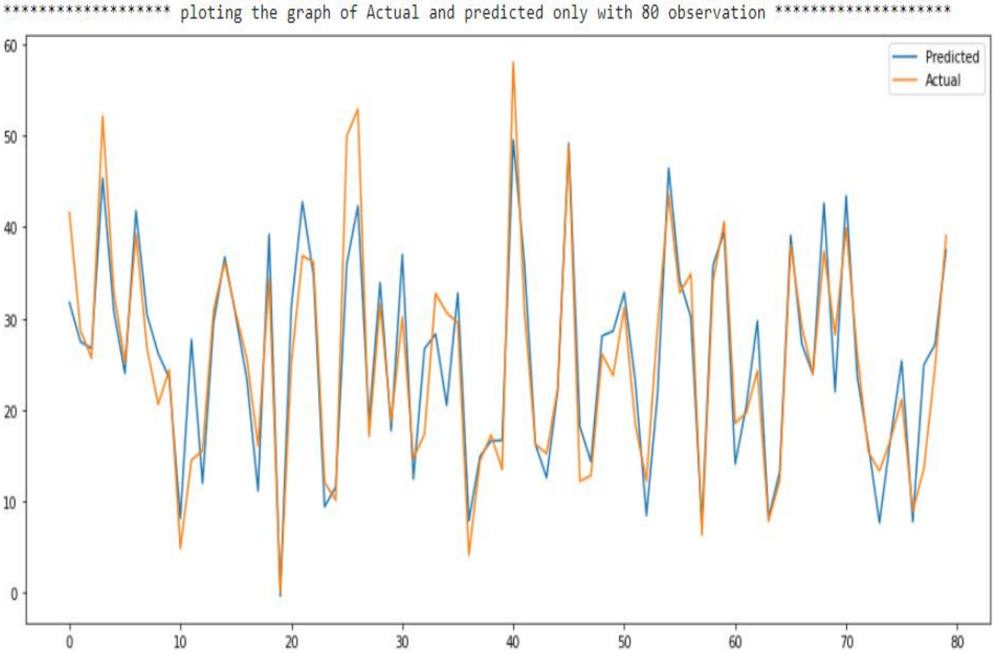
```
****************** ploting the graph of Actual and predicted only with 80 observation ********************
```





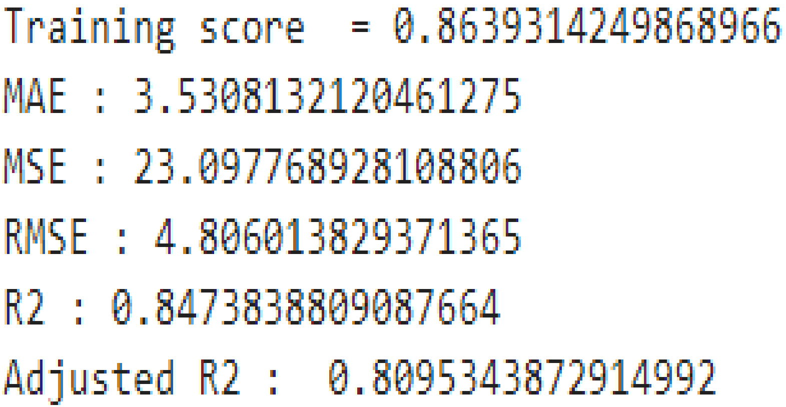
```
Training score  = 0.752830271580107
The best parameters found out to be :{'alpha': 1.9}
where model best score is:  0.750436106428743 9

MAE : 4.840845216063553
MSE : 38.175954219950114
RMSE : 6.178669292003749
R2 : 0.747756311942599
Adjusted R2 :  0.744488018965402 5
```

Linear Regression(Polynomial Features)



```
******************* ploting the graph of Actual and predicted only with 80 observation ********************
```

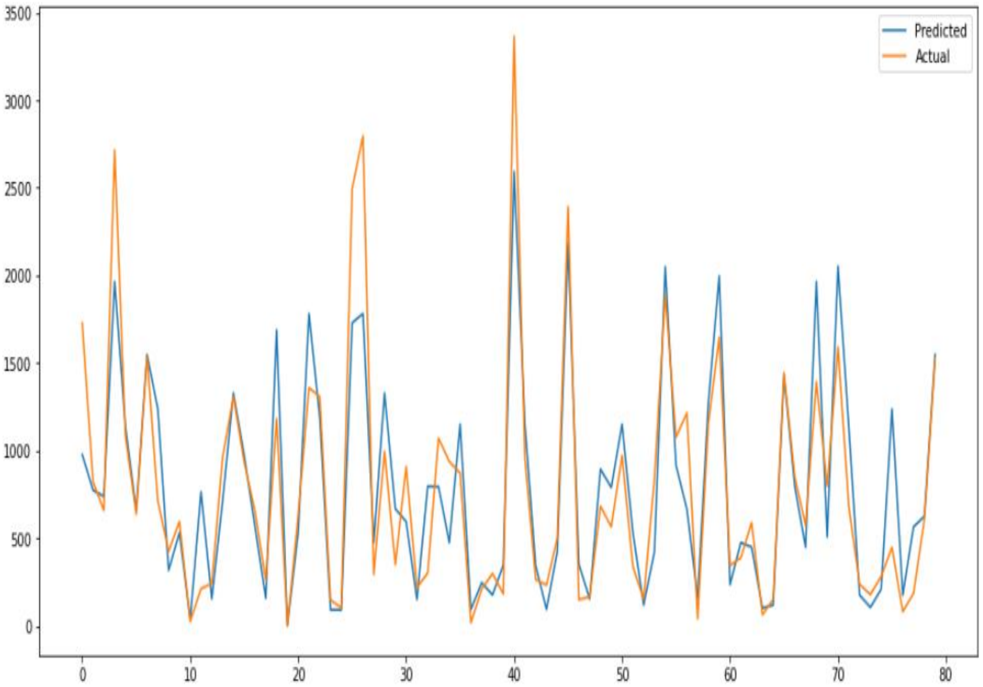
Using Polynomial Features with degree 2



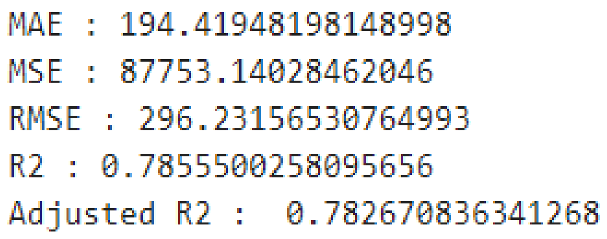
```
Training score  = 0.863931424986896б
MAE : 3.530813212046127５
MSE : 23.09776892810880б
RMSE : 4.806013829371365
R2 : 0.8473838809087664
Adjusted R2 :  0.809534387291499２
```

Decision Tree

```
****************** ploting the graph of Actual and predicted only with 80 observation *******************
```



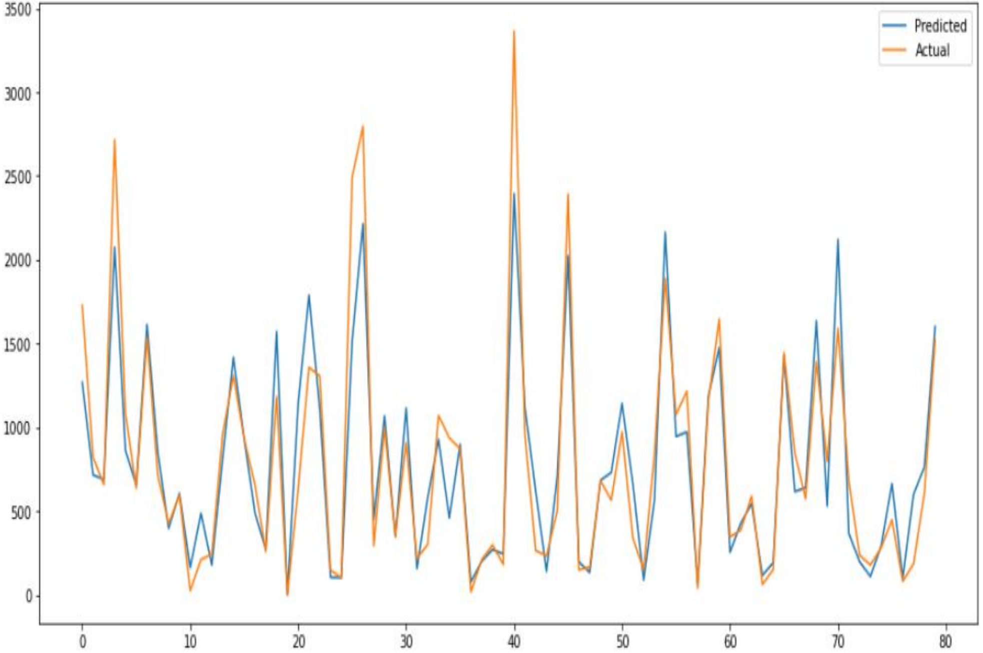
Training score = 0.8337445528699623 The best parameters found out to be :{'criterion': 'mse', 'max_depth': 15, 'max_features': 24, 'min_samples_split': 50, 'splitter': 'random'} where model best score is: 0.7931489412927507



```
MAE : 194.41948198148998
MSE : 87753.14028462046
RMSE : 296.23156530764993
R2 : 0.785500258095656
Adjusted R2 :  0.782670836341268
```

Random Forest Regressor

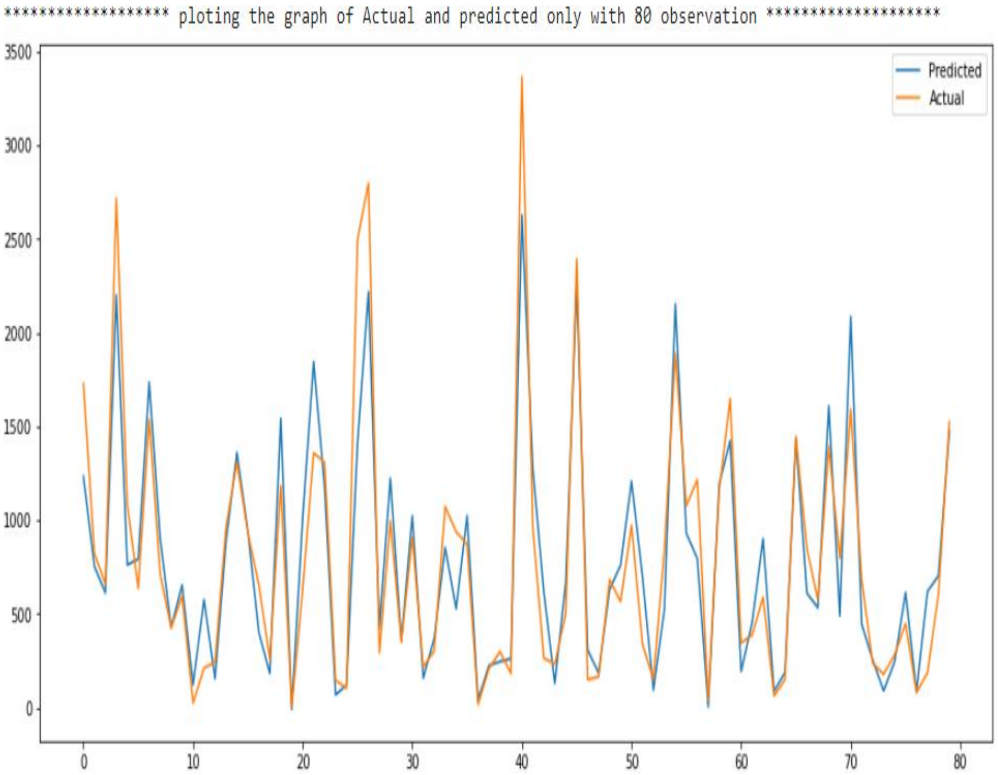
```
******************* ploting the graph of Actual and predicted only with 80 observation ********************
```



Training score = 0.9511223280260815 The best parameters found out to be:
{'max_depth': 20, 'max_features': 24, 'min_samples_split': 10, 'n_estimators': 100}
 where model best score is: 0.8577875917155999

```
MAE : 154.6444976595282
MSE : 57299.259482874615
RMSE : 239.37263728938322
R2 : 0.8599728206035834
Adjusted R2 :  0.8580928260653907
```

Gradient Boosting Regressor



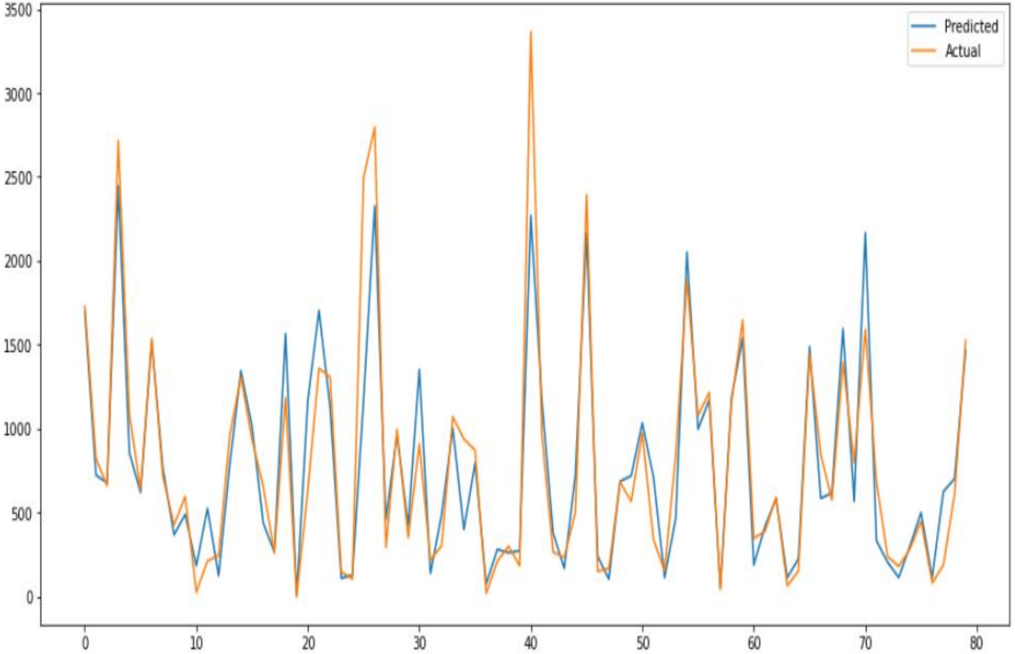
```
****************** ploting the graph of Actual and predicted only with 80 observation ********************
```

Fitting 5 folds for each of 80 candidates, totaling 400 fits
Training score = 0.9600157356513978 The best parameters found out to be :{'learning_rate': 0.1, 'max_depth': 6, 'estimators': 250}

```
where model best score is:  0.8589398718814742

MAE : 163.202950996003
MSE : 61375.71470538196
RMSE : 247.74122528433162
R2 : 0.850108327542623
Adjusted R2 :  0.8479970893051296
```

Adaboost Boost Regressor

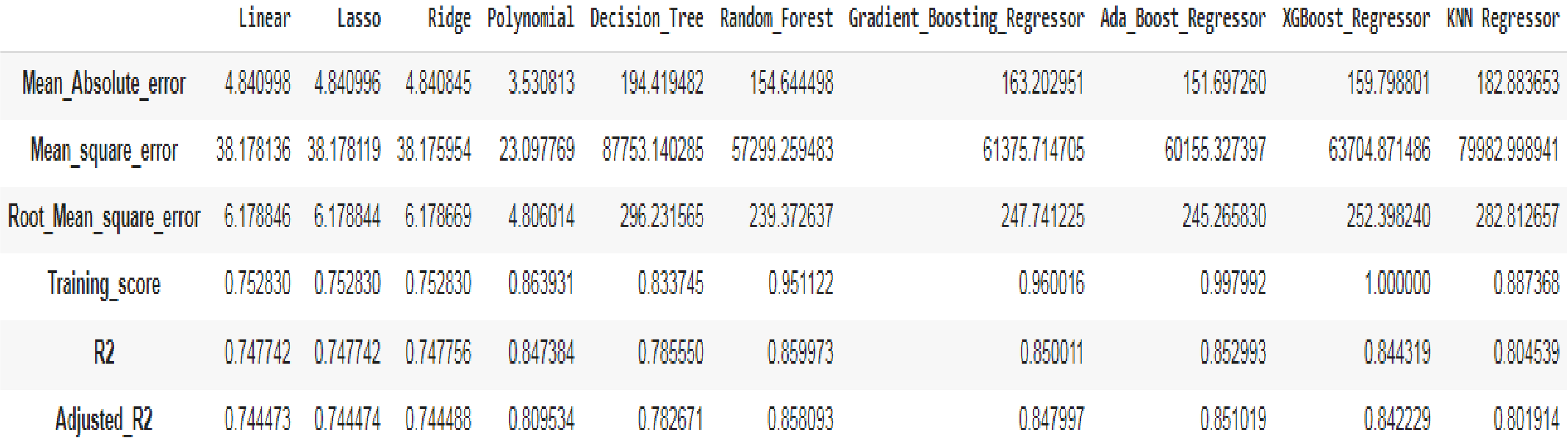
```
******************* ploting the graph of Actual and predicted only with 80 observation ********************
```



Training score = 0.9979923433449015 The best parameters found out to be :{'base_estimator': Decision Tree Regressor(), 'learning_rate': 1.5, 'n_estimators': 150}

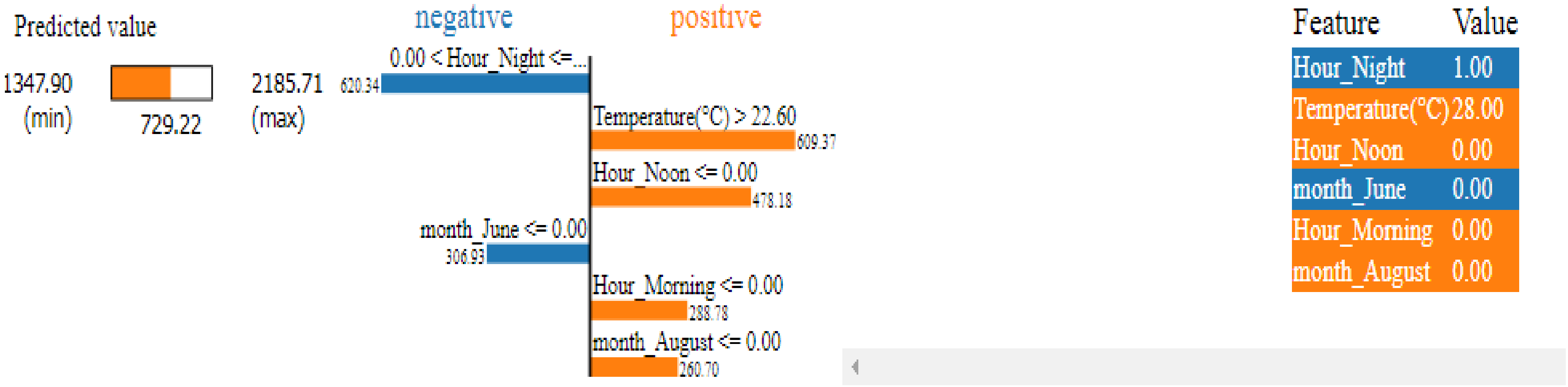
```
where model best score is:  0.8586673277122084

MAE : 151.6972602739726
MSE : 60155.32739726028
RMSE : 245.26583006456542
R2 : 0.8529931992642262
Adjusted R2 :  0.8510194968469403
```

Data Frame of all Evaluation Matrix with respect to each models



	Linear	Lasso	Ridge	Polynomial	Decision_Tree	Random_Forest	Gradient_Boosting_Regressor	Ada_Boost_Regressor	XGBoost_Regressor	KNN Regressor
Mean_Absolute_error	4.840998	4.840996	4.840845	3.530813	194.419482	154.644498	163.202951	151.697260	159.798801	182.883653
Mean_square_error	38.178136	38.178119	38.175954	23.097769	87753.140285	57299.259483	61375.714705	60155.327397	63704.871486	79982.998941
Root_Mean_square_error	6.178846	6.178844	6.178669	4.806014	296.231565	239.372637	247.741225	245.265830	252.398240	282.812657
Training_score	0.752830	0.752830	0.752830	0.863931	0.833745	0.951122	0.960016	0.997992	1.000000	0.887368
R2	0.747742	0.747742	0.747756	0.847384	0.785550	0.859973	0.850011	0.852993	0.844319	0.804539
Adjusted_R2	0.744473	0.744474	0.744488	0.809534	0.782671	0.858093	0.847997	0.851019	0.842229	0.801914

Model Explainability Using (LIME).



Feature	Value
Hour_Night	1.00
Temperature(°C)	28.00
Hour_Noon	0.00
month_June	0.00
Hour_Morning	0.00
month_August	0.00

Conclusion from Model Training

About models:

- ✓ Over fitting can be avoided using Power Transformer for scaling and transform it after train test split.
- ✓ Need to create a function to print the scores and can use be used in the model.
- ✓ Initially, we used few Linear Regression and then used Lasso and Ridge with the help of Hyper parameter Tuning.
- ✓ Used Polynomial Features with Linear Regression but did not get good scores.
- ✓ Later, we used Tree Base Models because we know multicollinearty not effect tree base models. So used Random Forest Regressor and Decision Tree.

- ✓ Then decided to use some boosting regression for better predictions and used Gradient Boosting Regressor, Adaboost Boost Regressor, XGBoost Regression.
- ✓ Achieved the best RMSE score with Linear, Lasso, Ridge, Polynomial. We got best Training score with Random Forest. We achieved best R2, Adjusted_R2 with Decision Tree, Random Forest, Bagging.

Thank You