Capstone Project
Seoul Bike Sharing Demand Prediction

Team

Rahul Kumar Soni, Lakdawala Ali Asgar,
Kanishka Raj, Sridhar Nagar

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.

Content

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 analysed and we tried to study various models so as to get the best performing model for our project.

Data Description

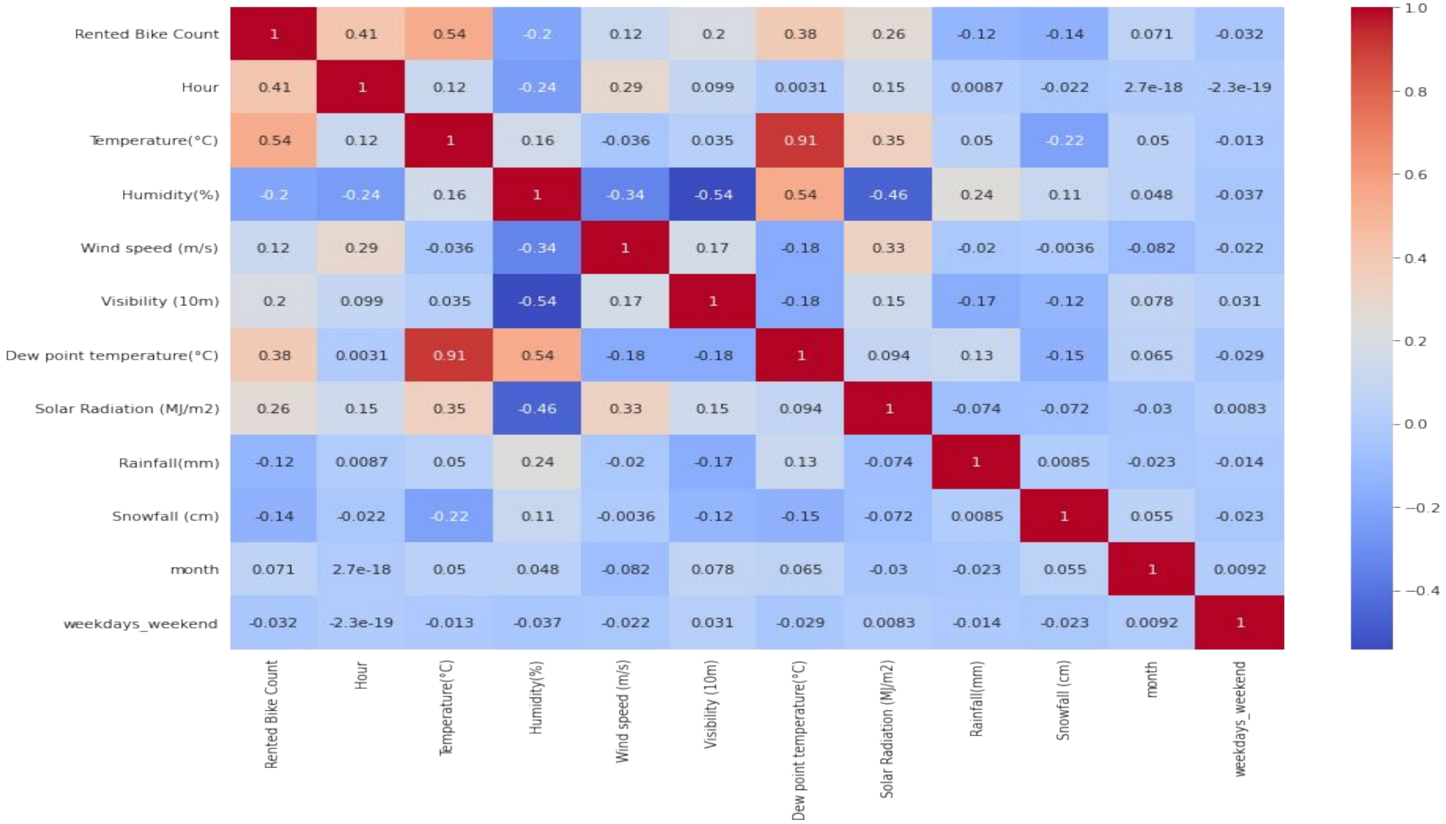
Dependent variable:

- Rented Bike count - Count of bikes rented at each hour

Independent variables:

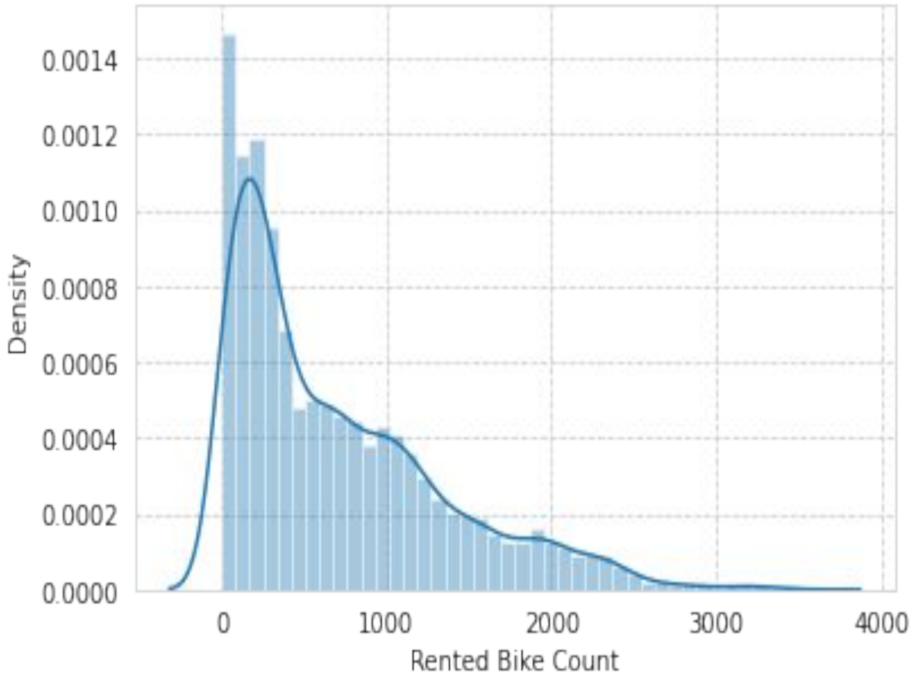
- Date : year-month-day
- Hour - Hour of the day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10 m
- Dew point temperature - Celsius
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

EDA - Feature Correlation

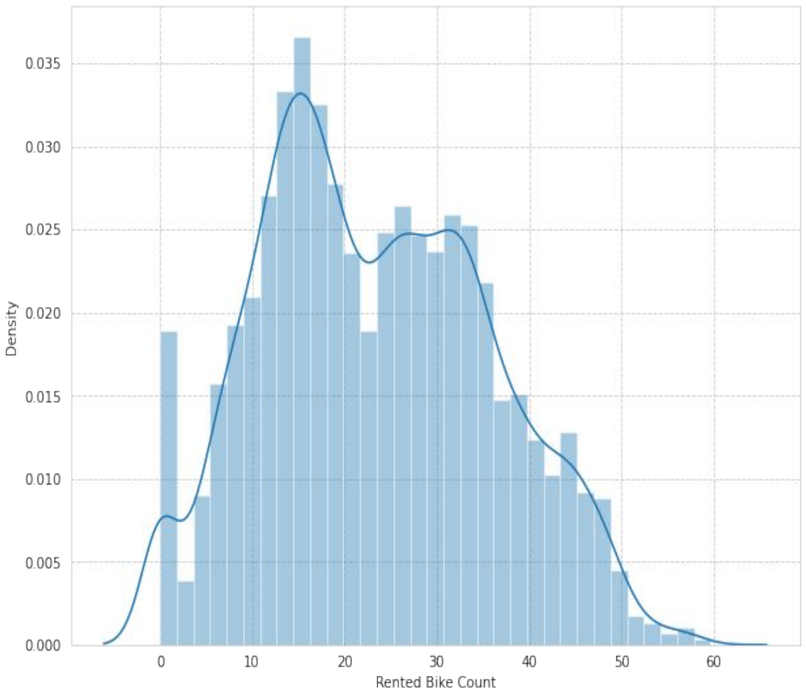


Correlation Graph

EDA (contd...)

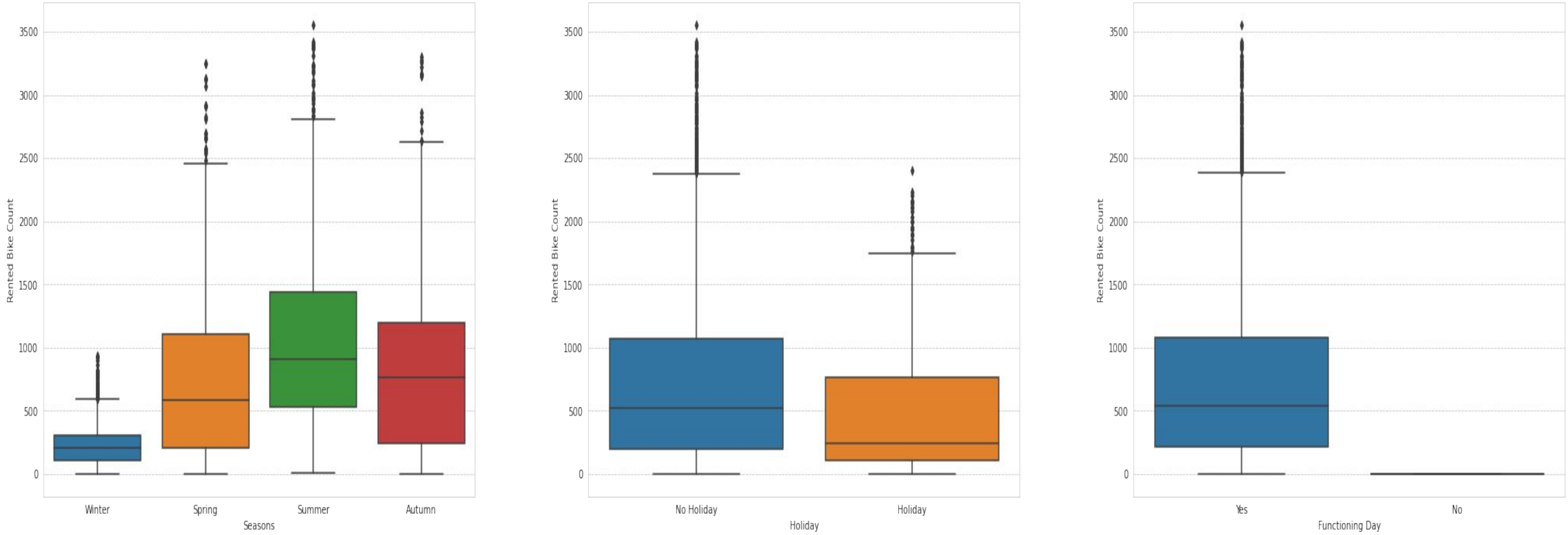


Distribution of rented bike count



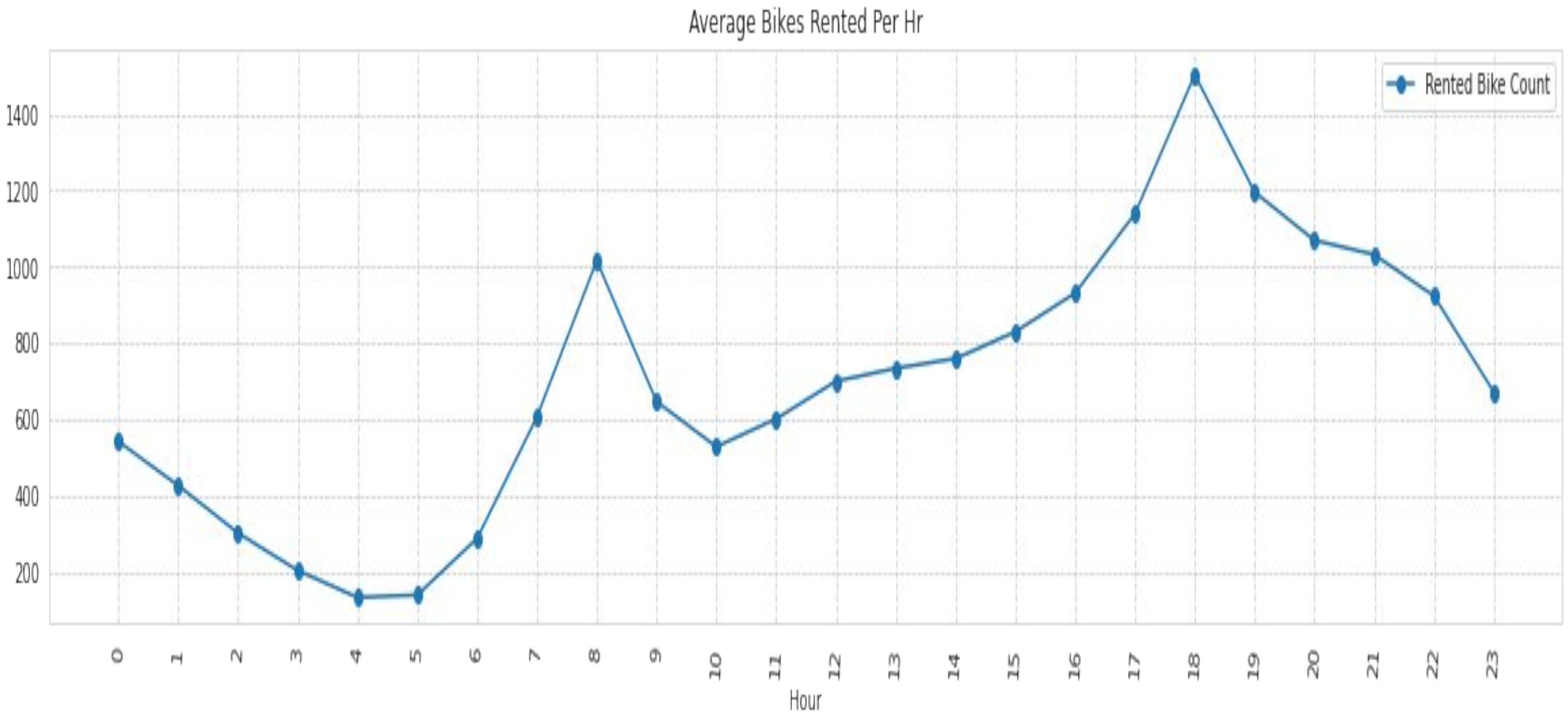
Square root transformation of rented bike count

EDA (contd...)

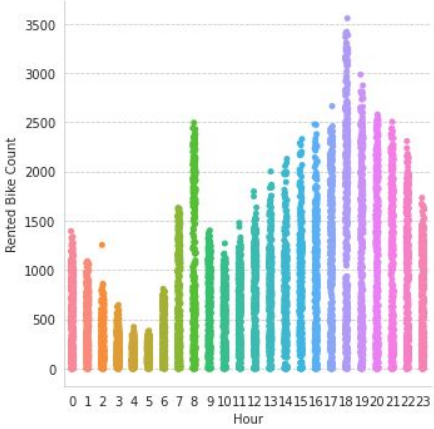


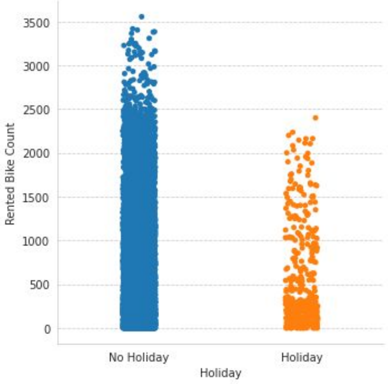
- Less demand on winter seasons
- Slightly Higher demand during Non holidays
- Almost no demand on non functioning day

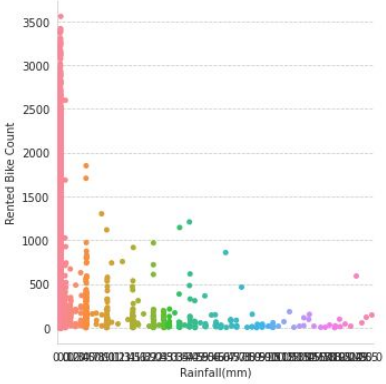
EDA (contd...)

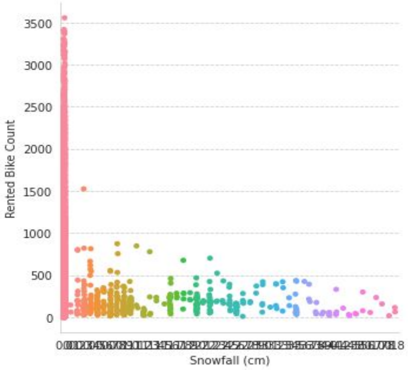

Average Bikes Rented Per Hr
Rented Bike Count
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour

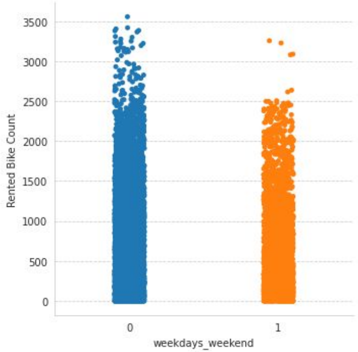
EDA (contd…)


Rented Bike Count
0
500
1000
1500
2000
2500
3000
3500
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour

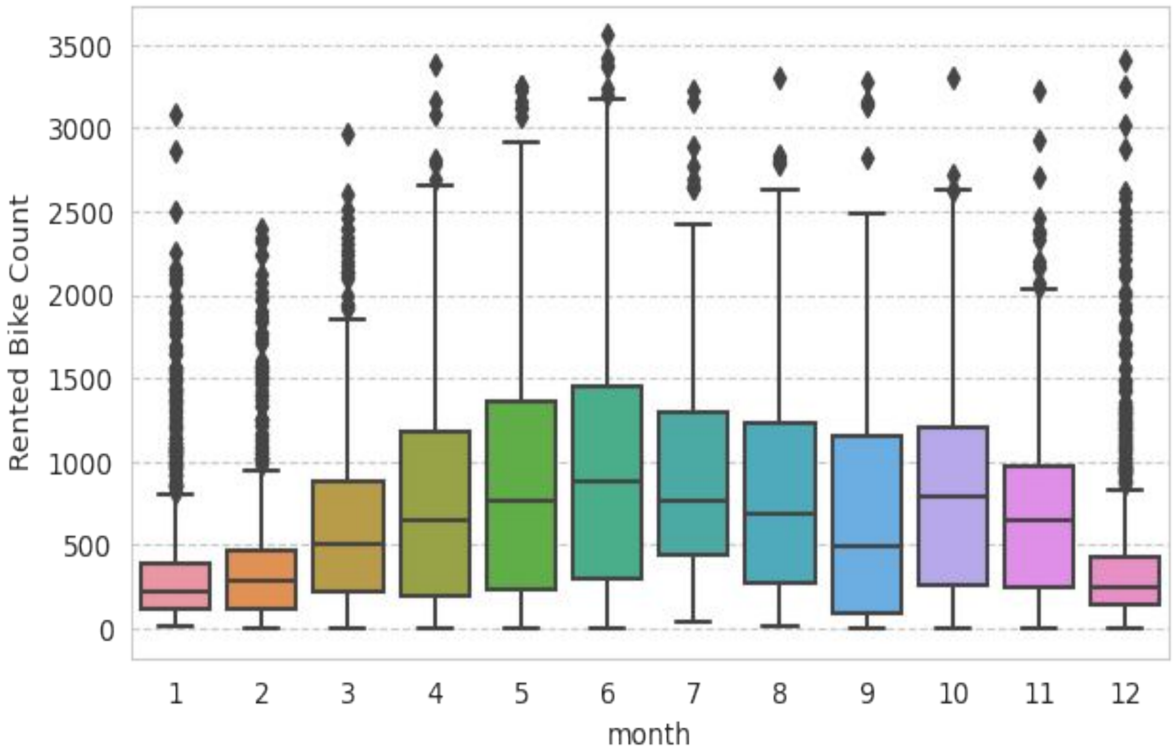

Rented Bike Count
0
500
1000
1500
2000
2500
3000
3500
No Holiday
Holiday
Holiday


Rented Bike Count
0
500
1000
1500
2000
2500
3000
3500
Rainfall(mm)


Rented Bike Count
0
500
1000
1500
2000
2500
3000
3500
Snowfall (cm)


Rented Bike Count
0
500
1000
1500
2000
2500
3000
3500
0
1
weekdays_weekend

EDA (contd...)



- We can see that there less demand of Rented bike in the month of December, January, February i.e. during winter seasons

- Also demand of bike is maximum during May, June, July i.e Summer seasons

Model’s Performed

- Linear Regression with regularizations
- Polynomial Regression
- K nearest neighbours
- Decision tree
- Random forest
- Gradient Boost
- eXtreme Gradient Boost
- lightGBM
- CatBoost

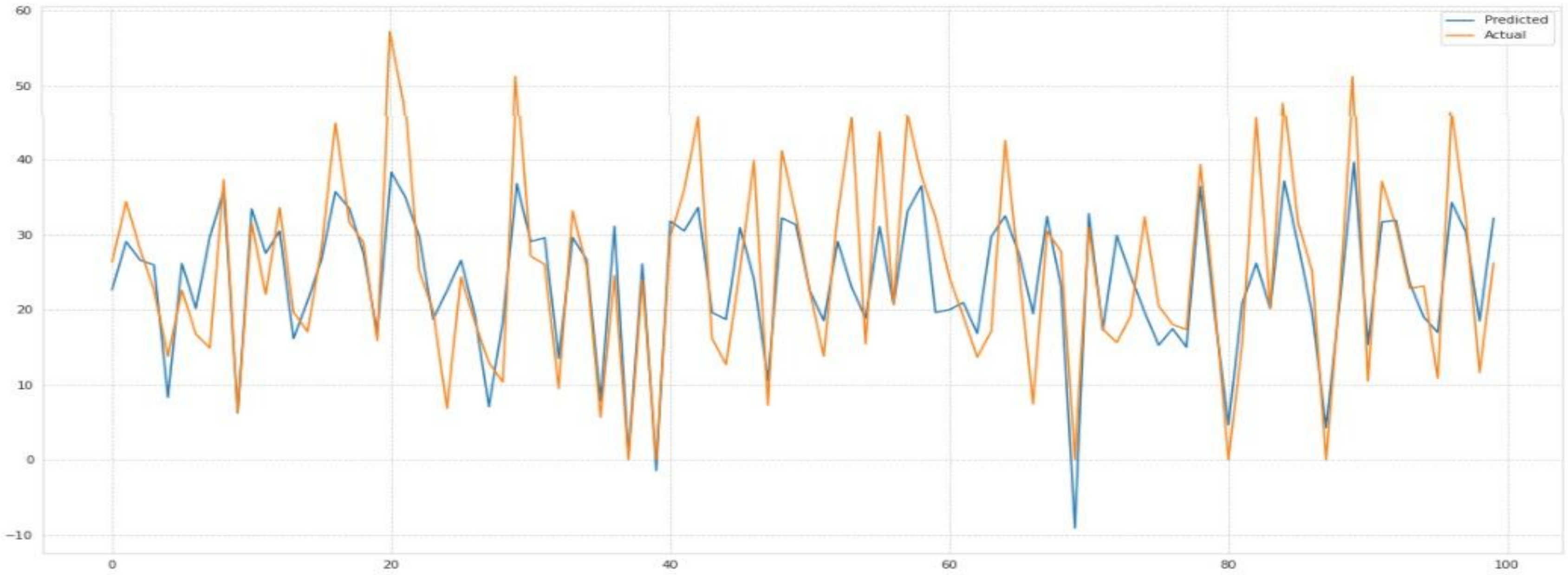
Linear Regression

```
================Evalution Matrix=========================

MSE : 175590.55287332062
RMSE : 419.035264474627
R2 : 0.5729108337712393
Adjusted R2 :  0.5697661367350404

================Evalution Matrix=========================
```

```
=========================================================  Evalution Graph  ===================================================
```



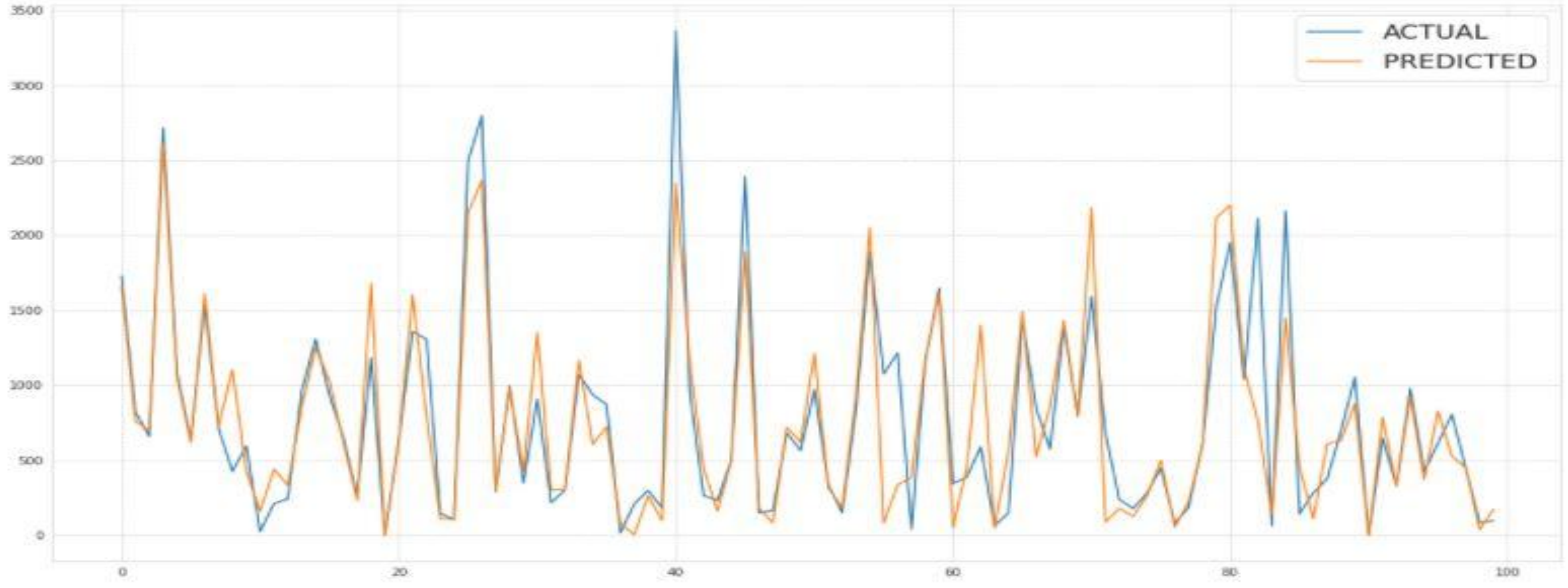
Decision Tree Regression

```
================Evalution Matrix=========================

MSE : 88288.61232876712
RMSE : 297.13399726178613
R2 : 0.7842414462456377
Adjusted R2 :  0.7826527960569264

================Evalution Matrix=========================
```

== Evalution Graph ==



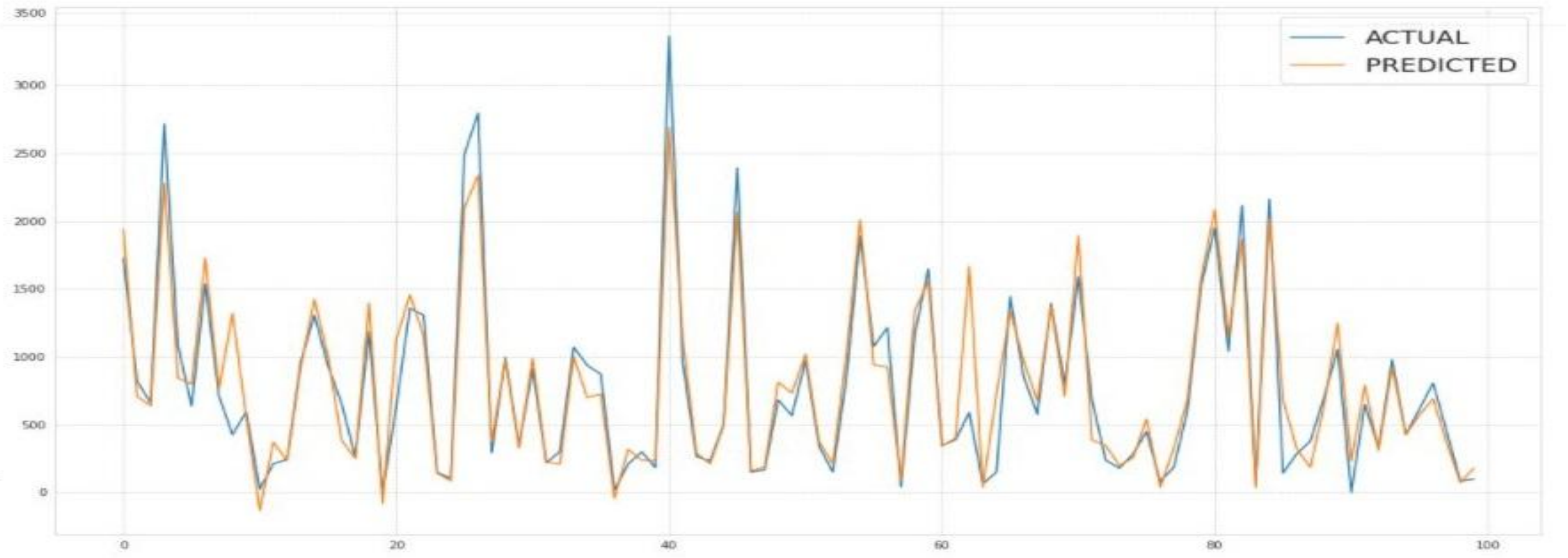
CatBoost

```
================Evalution Matrix=========================

MSE : 36706.5353729677
RMSE : 191.58949703198164
R2 : 0.910297049908164
Adjusted R2 :  0.9096365587892181

================Evalution Matrix=========================
```

=== Evalution Graph ===



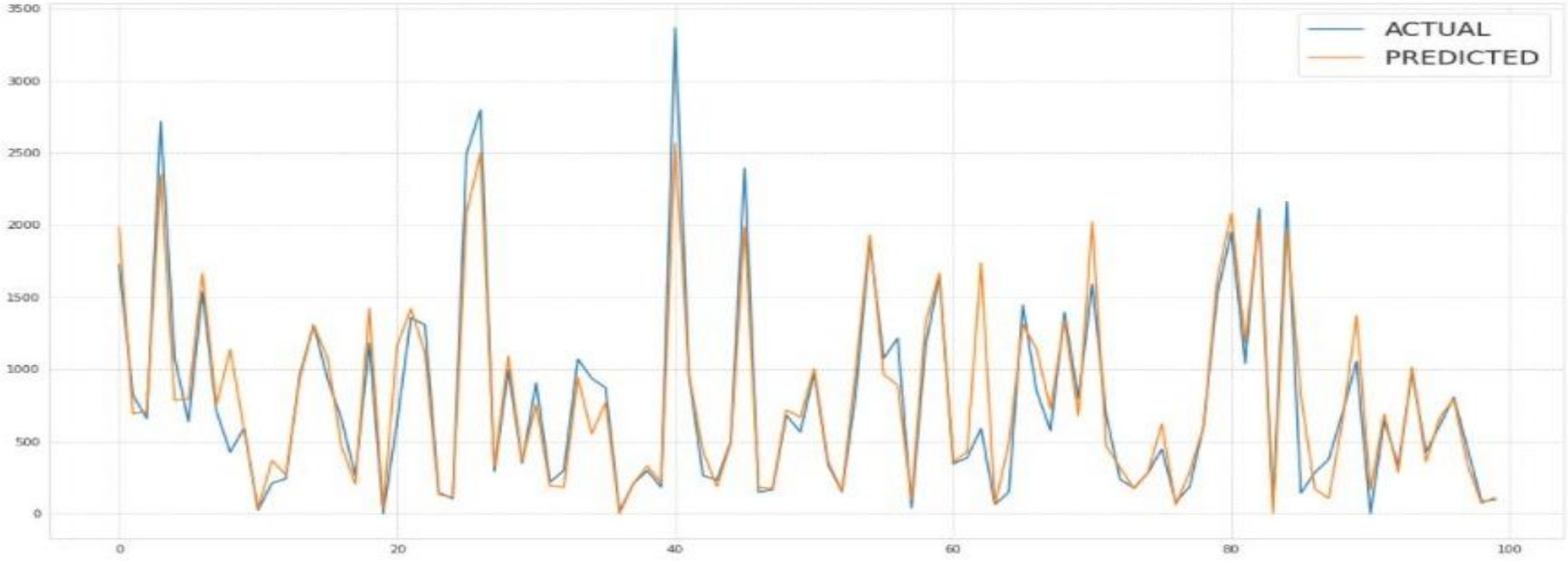
lightGBM

```
================Evalution Matrix=========================

MSE : 35410.75375394222
RMSE : 188.17745283094416
R2 : 0.9134636640470446
Adjusted R2 :  0.9128264890009115

================Evalution Matrix=========================
```

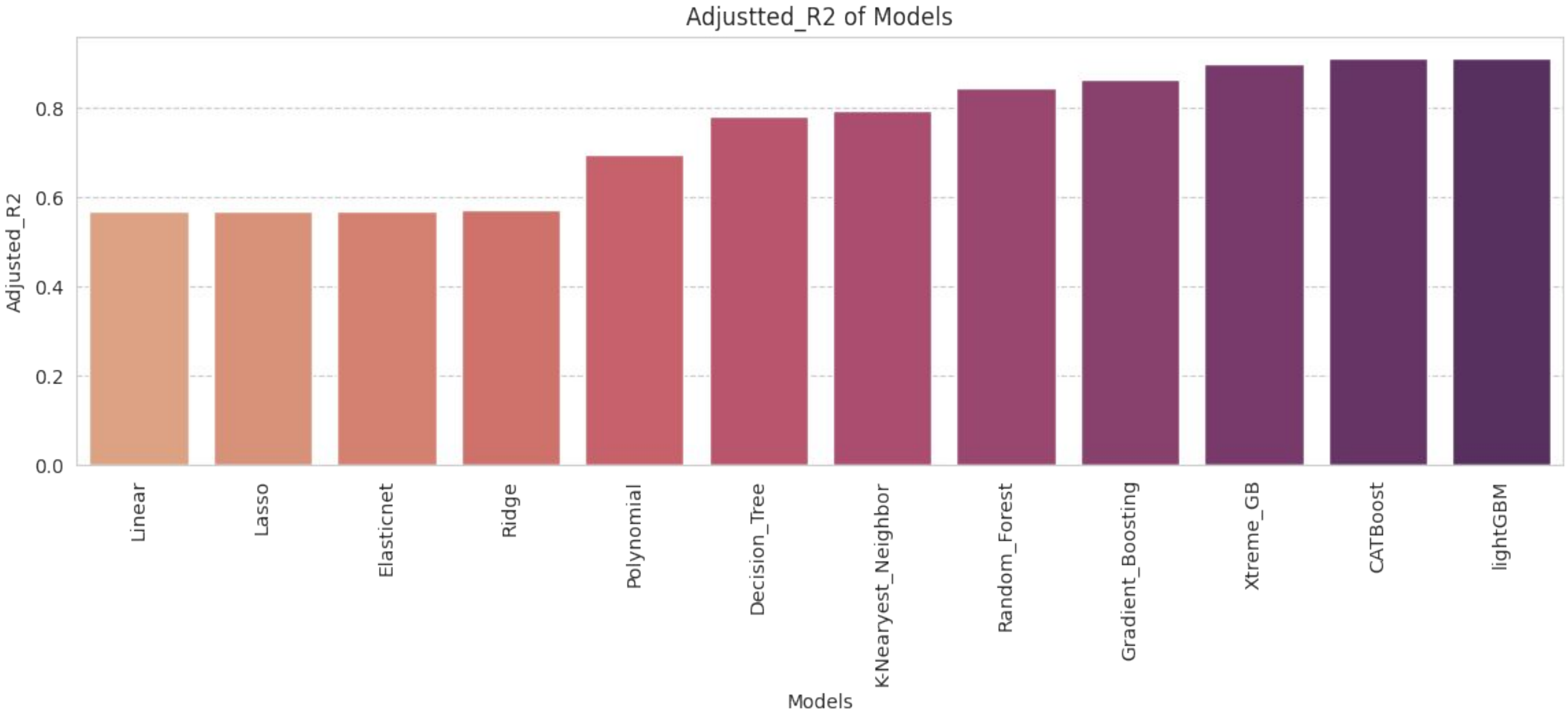
```
=========================================================    Evalution Graph    ===================================================
```



Model’s Evaluation Matrices

	Models	Mean_square_error	Root_Mean_square_error	R2	Adjusted_R2
0	Linear	175590.552873	419.035264	0.572911	0.569766
1	Lasso	175560.907118	418.999889	0.572983	0.569839
2	Ridge	175248.935066	418.627442	0.573742	0.570603
3	Elasticnet	175346.867499	418.744394	0.573504	0.570363
4	Polynomial	123952.860328	352.069397	0.698509	0.696289
5	K-Nearyest_Neighbor	83411.759209	288.810940	0.796159	0.794659
6	Decision_Tree	88506.087215	297.499726	0.783710	0.782117
7	Random_Forest	62790.180423	250.579689	0.846554	0.845424
8	Gradient_Boosting	55090.172685	234.712958	0.865371	0.864380
9	Xtreme_GB	40812.801816	202.021785	0.900262	0.899528
10	CATBoost	36339.421527	190.629015	0.911194	0.910540
11	lightGBM	35410.753754	188.177453	0.913464	0.912826

Adjusted R2 of Model’s Performed



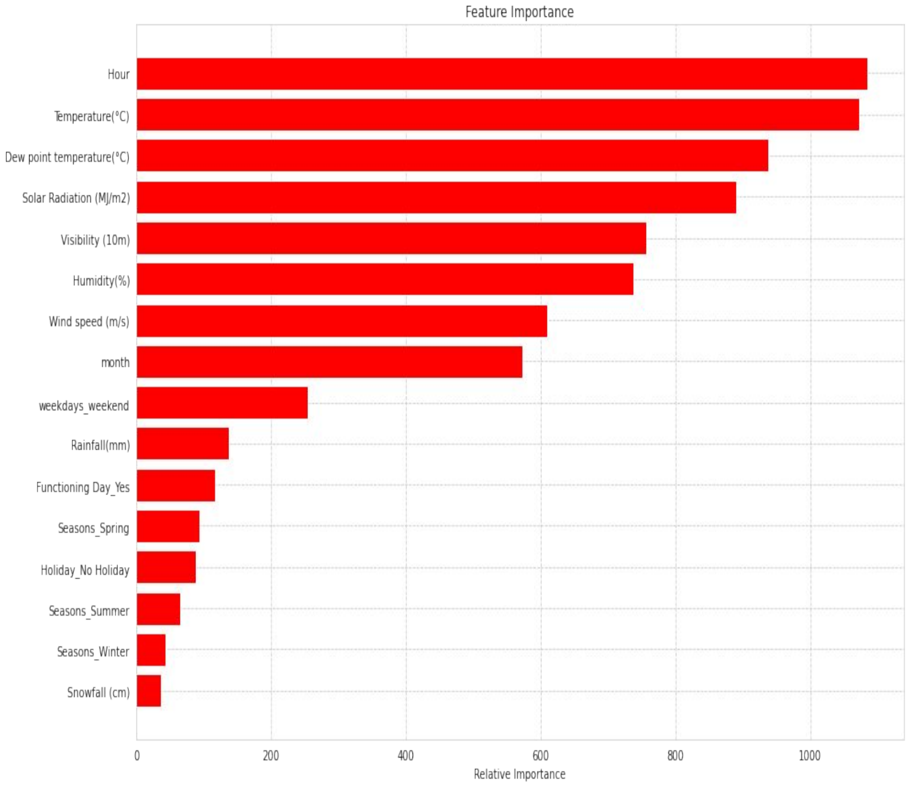
Model Validation & Selection(continued)

- **Observation 1:** As seen in the Model Evaluation Matrices table, Linear Regression, KNN is not giving great results.

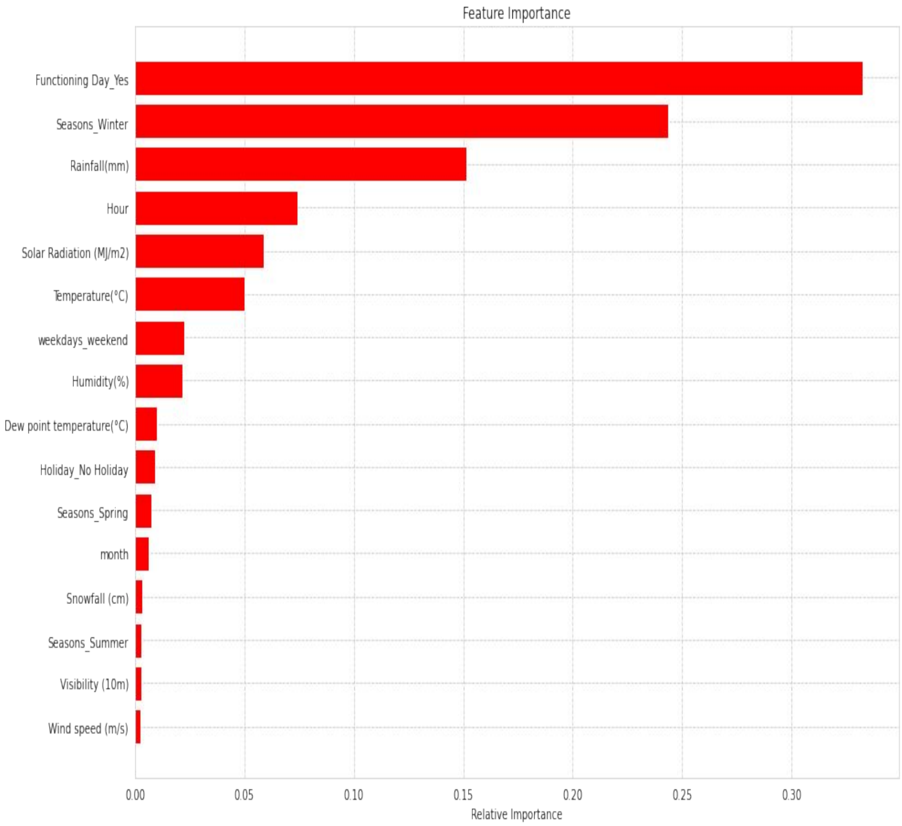
- **Observation 2:** Random forest & GBR have performed equally good in terms of adjusted r2.

- **Observation 3:** We are getting the best results from lightGBM and CatBoost.

Feature Importance

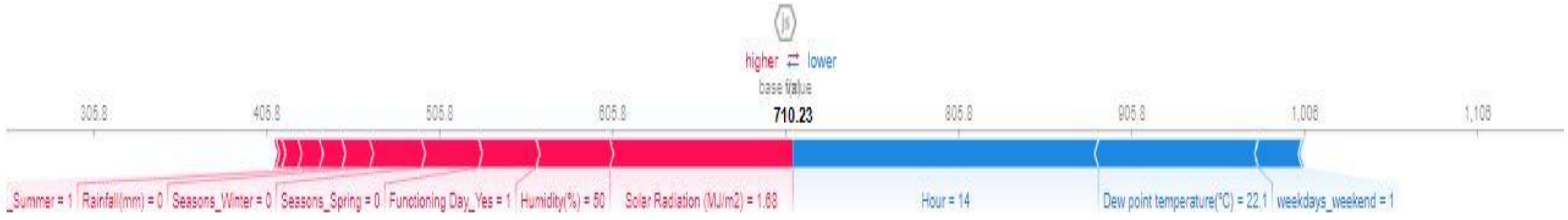


lightGBM



CatBoost

Model Explainability - SHAP



lightGBM

CatBoost

ELI5 for LGBR model

y (score **710.232**) top features

Contribution?	Feature	Value
+705.796	<BIAS>	1.000
+191.367	Solar Radiation (MJ/m2)	1.680
+74.526	Temperature(°C)	34.000
+38.931	Functioning Day_Yes	1.000
+37.236	Humidity(%)	50.000
+27.307	weekdays_weekend	1.000
+12.110	Seasons_Spring	0.000
+10.303	Seasons_Summer	1.000
+10.040	Wind speed (m/s)	1.200
+7.519	Rainfall(mm)	0.000
+3.179	month	7.000
+2.952	Holiday_No Holiday	1.000
+2.872	Visibility (10m)	1744.000
+2.695	Seasons_Winter	0.000
+1.149	Snowfall (cm)	0.000
-119.583	Dew point temperature(°C)	22.100
-298.167	Hour	14.000

Challenges

- A huge amount of data needed to be dealt while doing the project which is quite an important task and also even small inferences need to be kept in mind.
- Required lot of graph to analyze
- Carefully handled feature selection part as it affects the R2 score.
- As dataset was quite big enough which led more computation time.

Conclusion

- No overfitting is seen.
- It is quite evident from the results that lightGBM and Catboost is the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (mse,rmse) shows lower and (R2, Adjusted_R2) show a higher value for the lightGBM and Catboost models.
- So, we can deploy lightGBM or catboost model for the above problem

THANK YOU