Capstone Project

Seoul Bike Sharing Demand Prediction

By

NIDHI PANDEY

AlmaBetter

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
- Data Description
- Exploratory Data Analysis
- Models performed
- Model Validation & Selection
- Evaluation Matrix of All the models
- Model Explainability - SHAP
- Challenges
- Conclusion

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

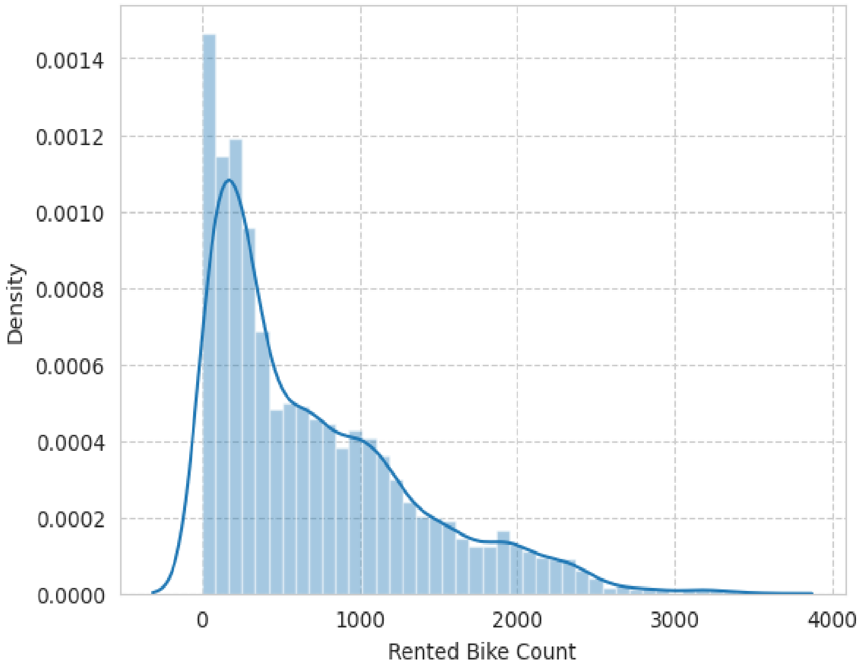
Dependent variable:

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

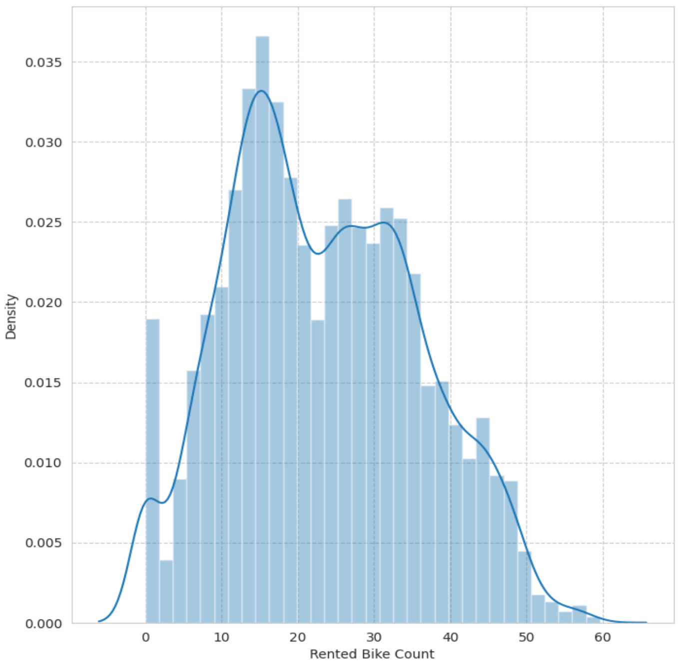
Independent variables:

- 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)

Exploratory Data Analysis

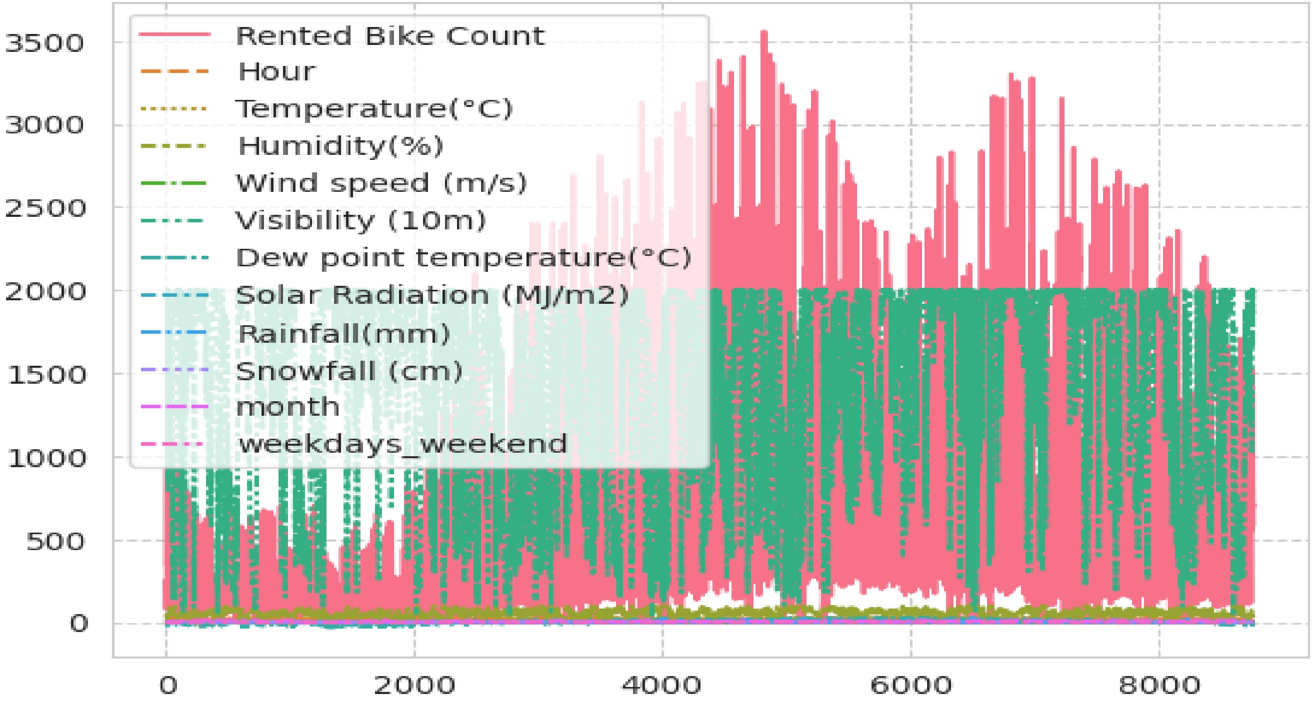


Distribution of rented bike count



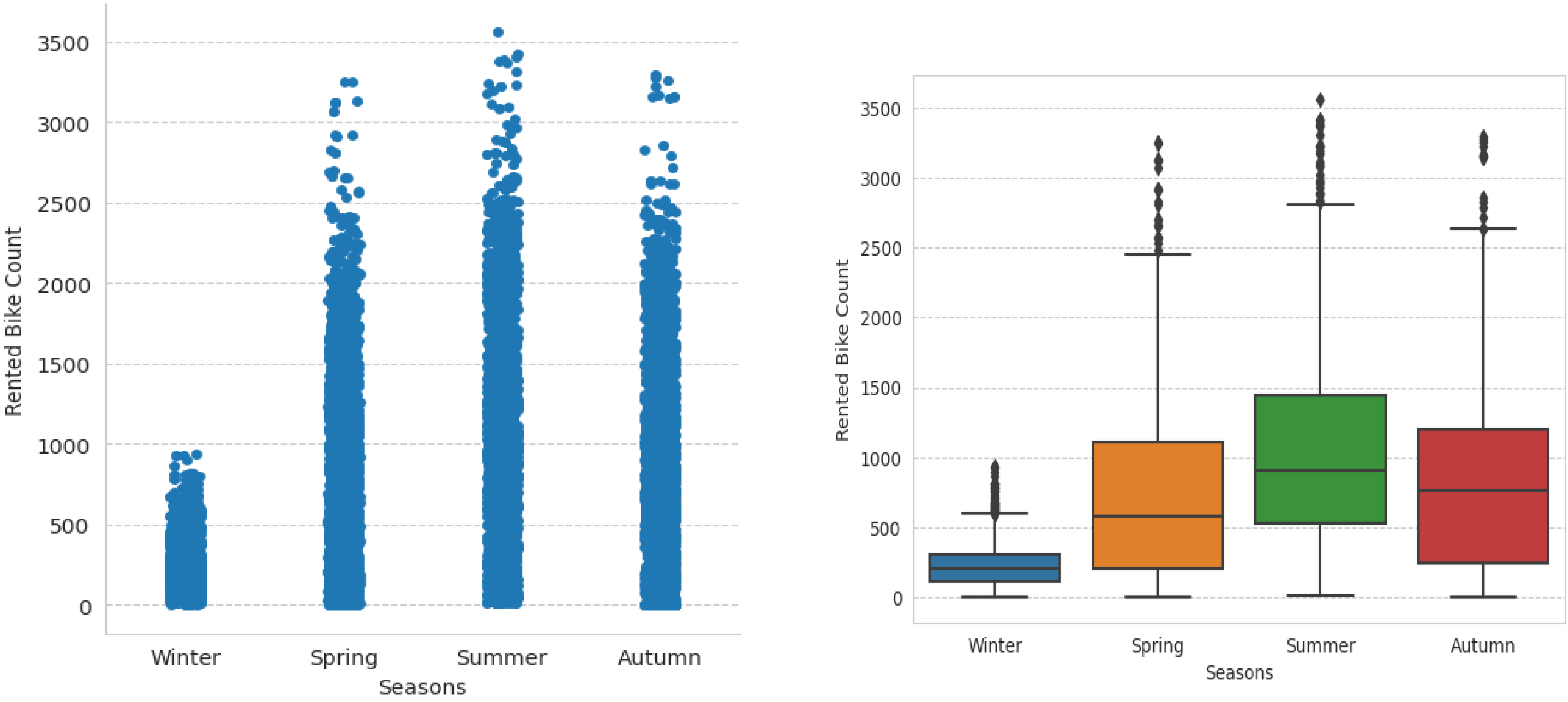
Square root transformation of rented bike count

EDA(contd..)



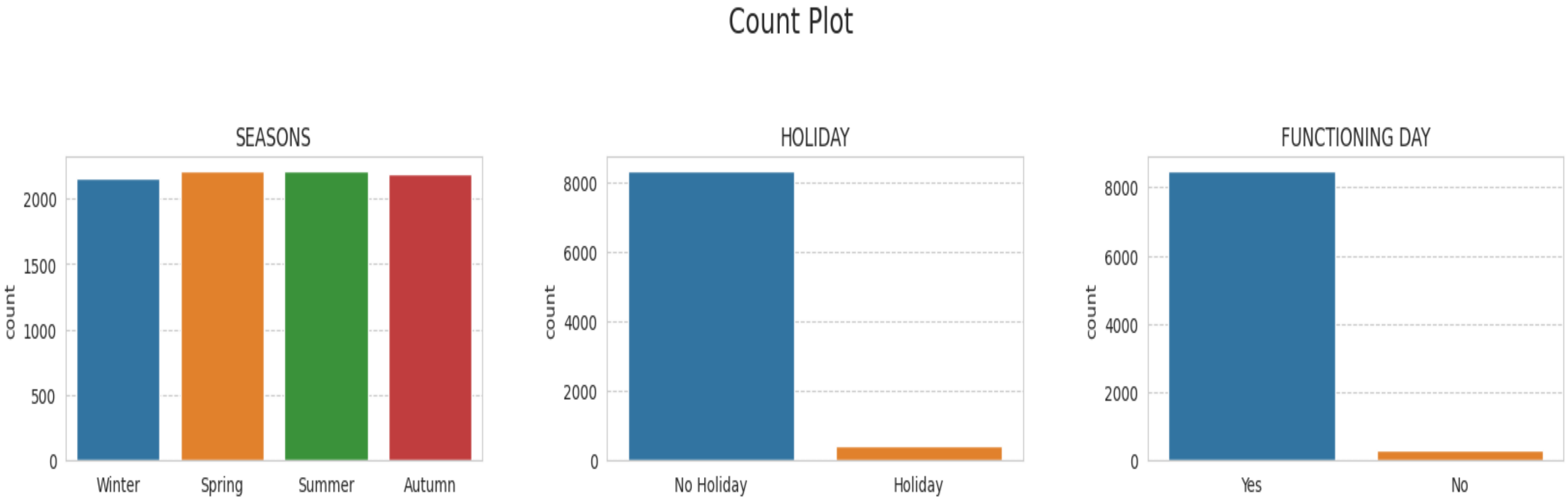
Line plot for the total interception of data

EDA(contd..)



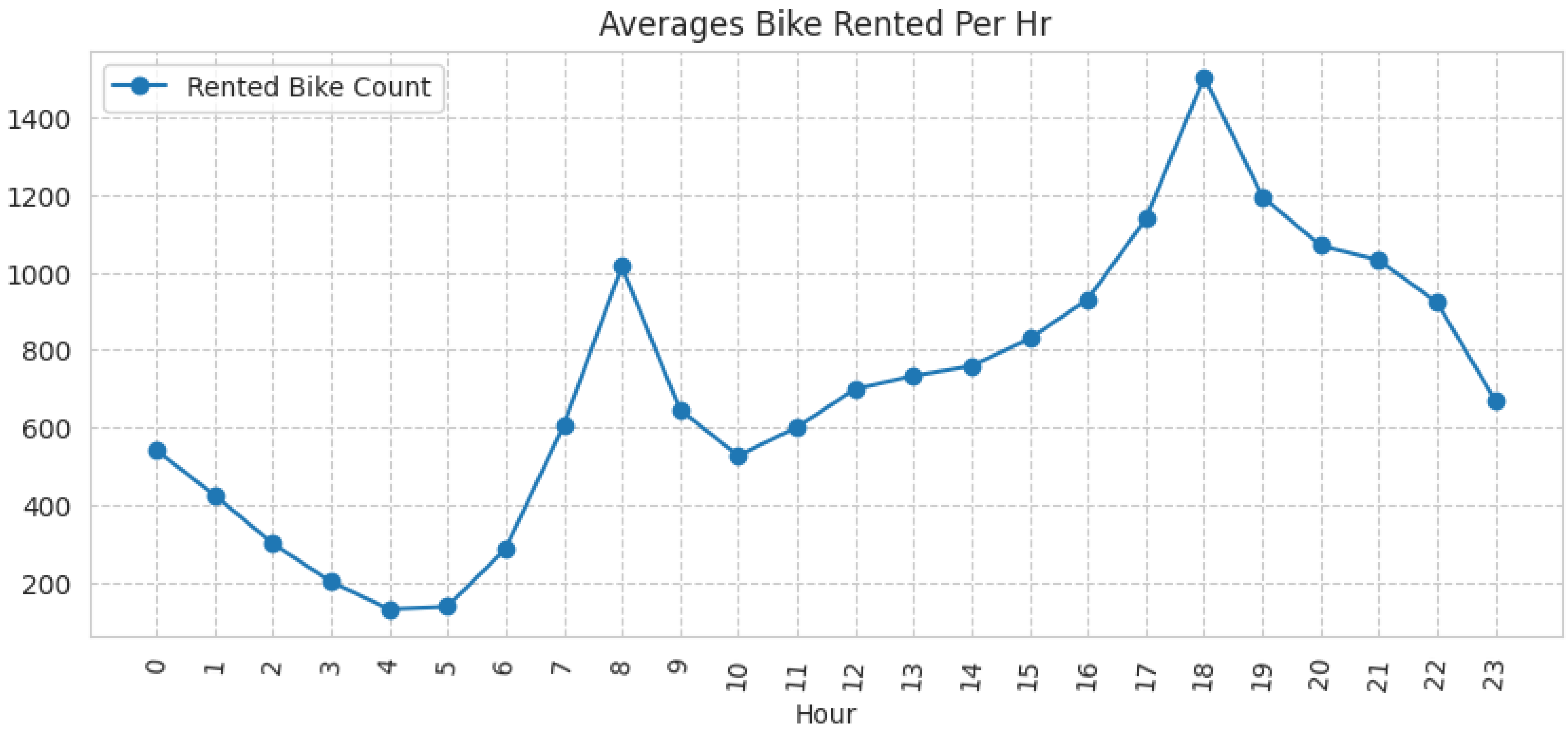
Plot for the demand of bikes in different season

EDA (contd...)



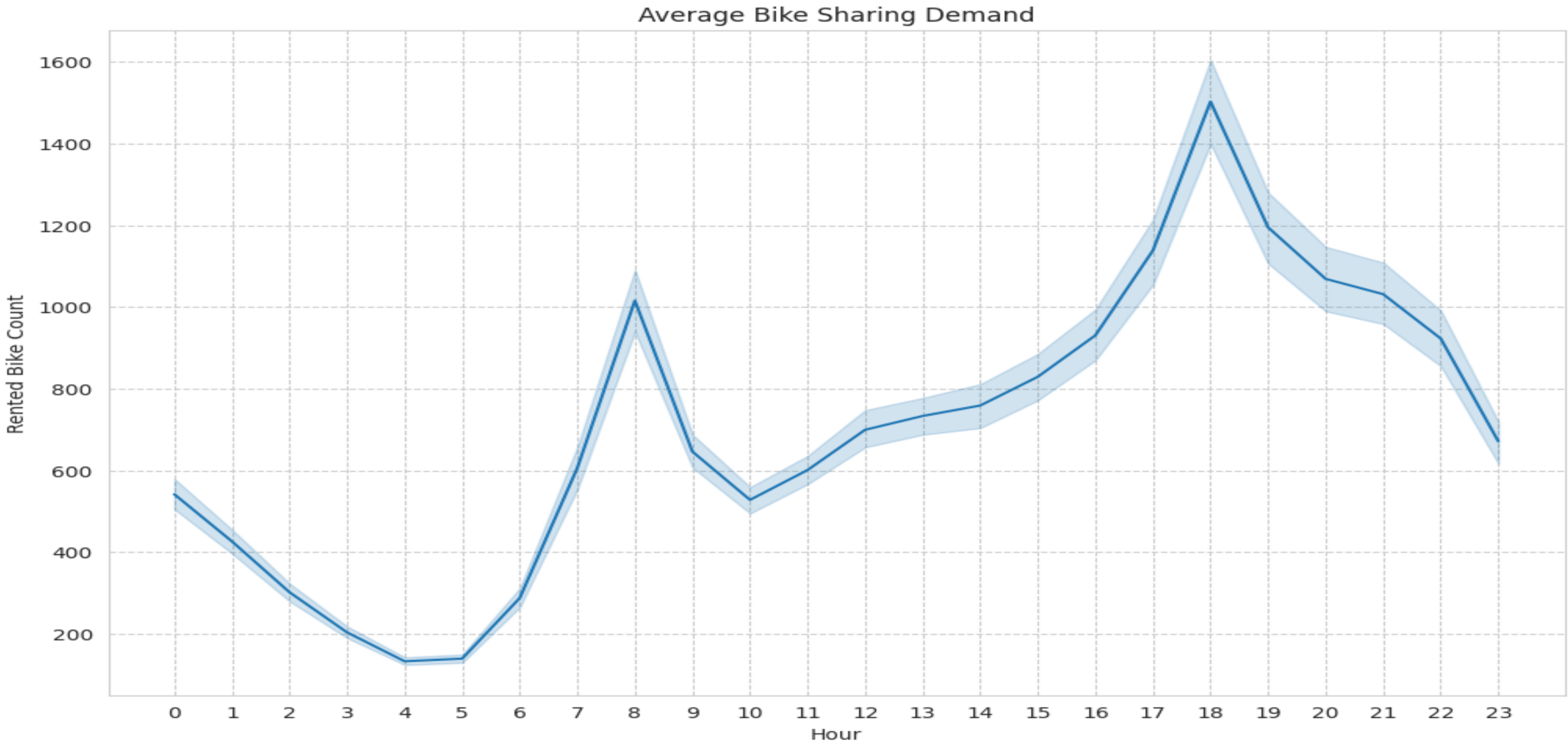
- Less demand on winter seasons
- Slightly Higher demand during Non holidays
- Almost no demand on non functioning day

EDA(contd..)



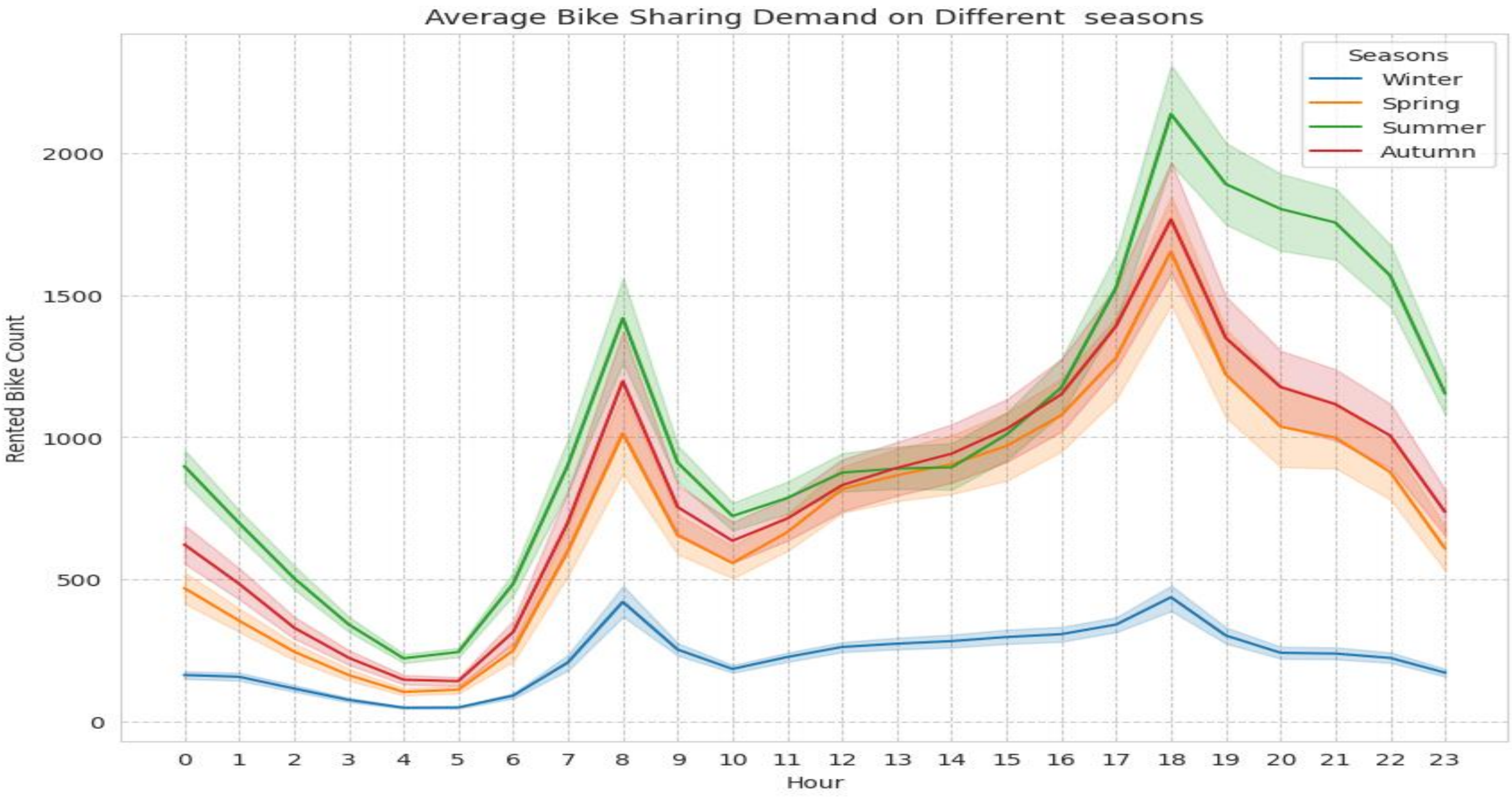
Averages Bike Rented Per Hr
Rented Bike Count
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour

EDA(contd..)



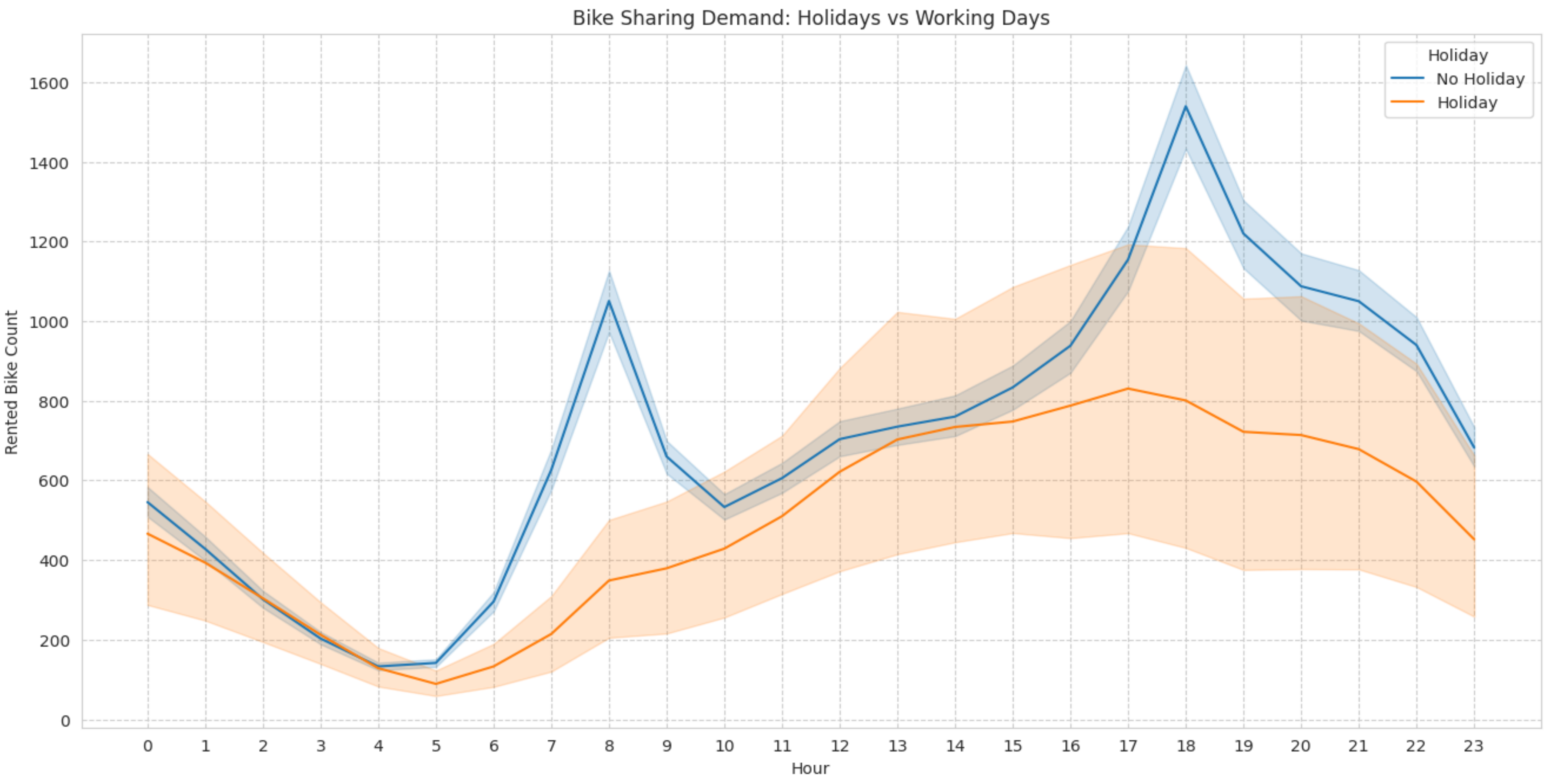
Average Bike Sharing Demand
Rented Bike Count
1600
1400
1200
1000
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour

EDA(contd..)

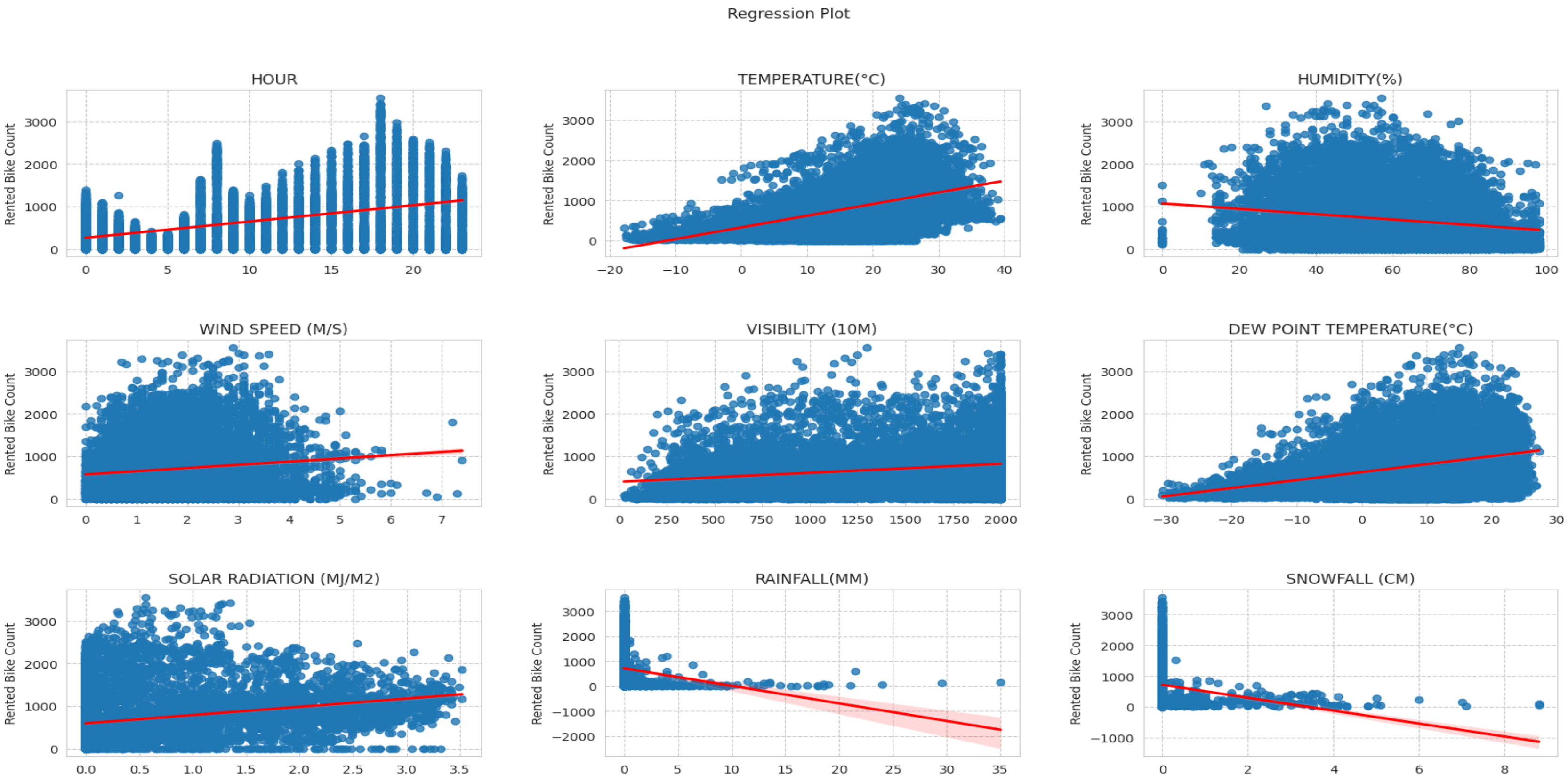


Average Bike Sharing Demand on Different seasons
Seasons
Winter
Spring
Summer
Autumn
Rented Bike Count
0
500
1000
1500
2000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour

EDA(contd..)

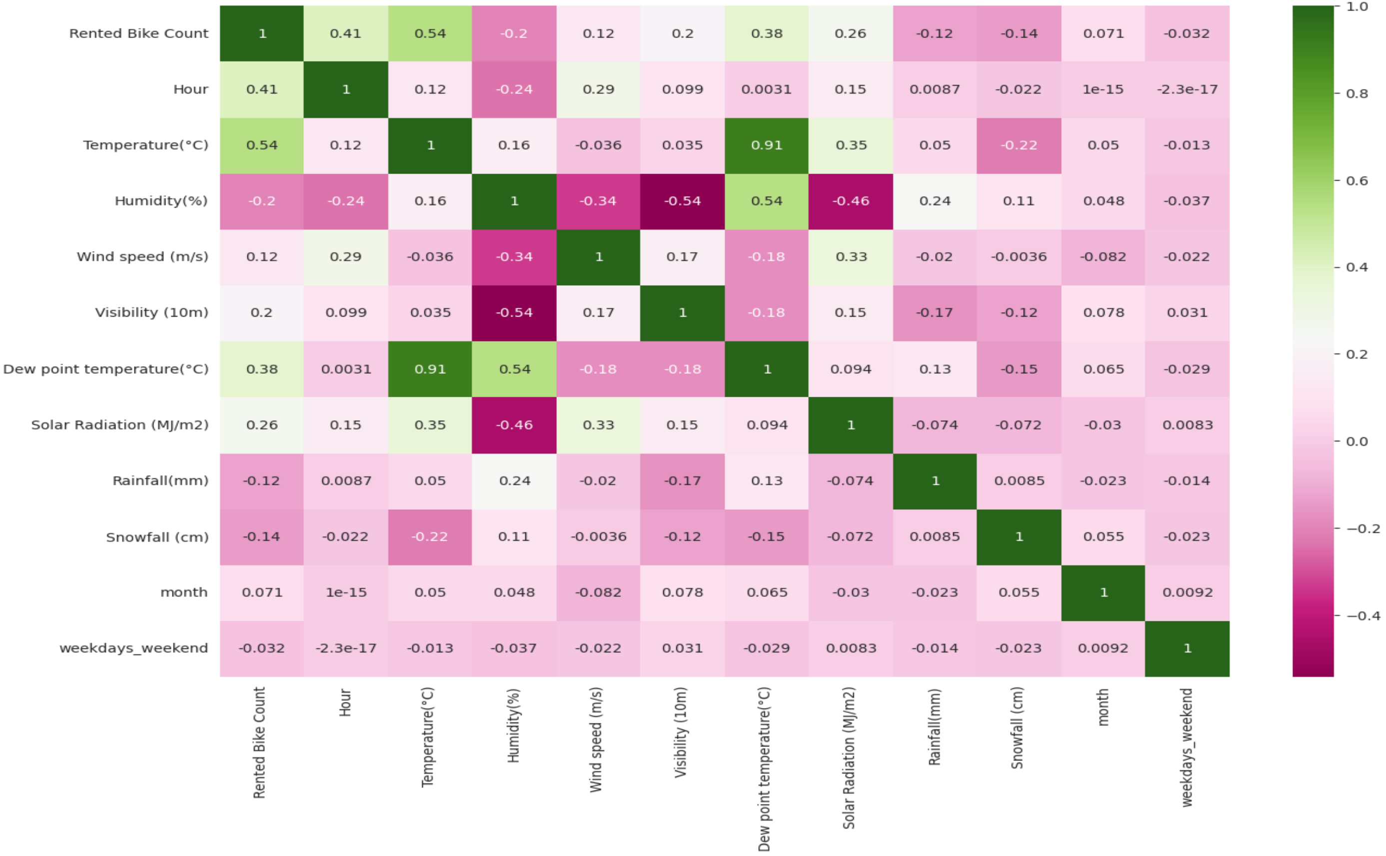

Bike Sharing Demand: Holidays vs Working Days
Holiday
No Holiday
Holiday
Rented Bike Count
Hour
0
200
400
600
800
1000
1200
1400
1600
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

EDA (contd...)

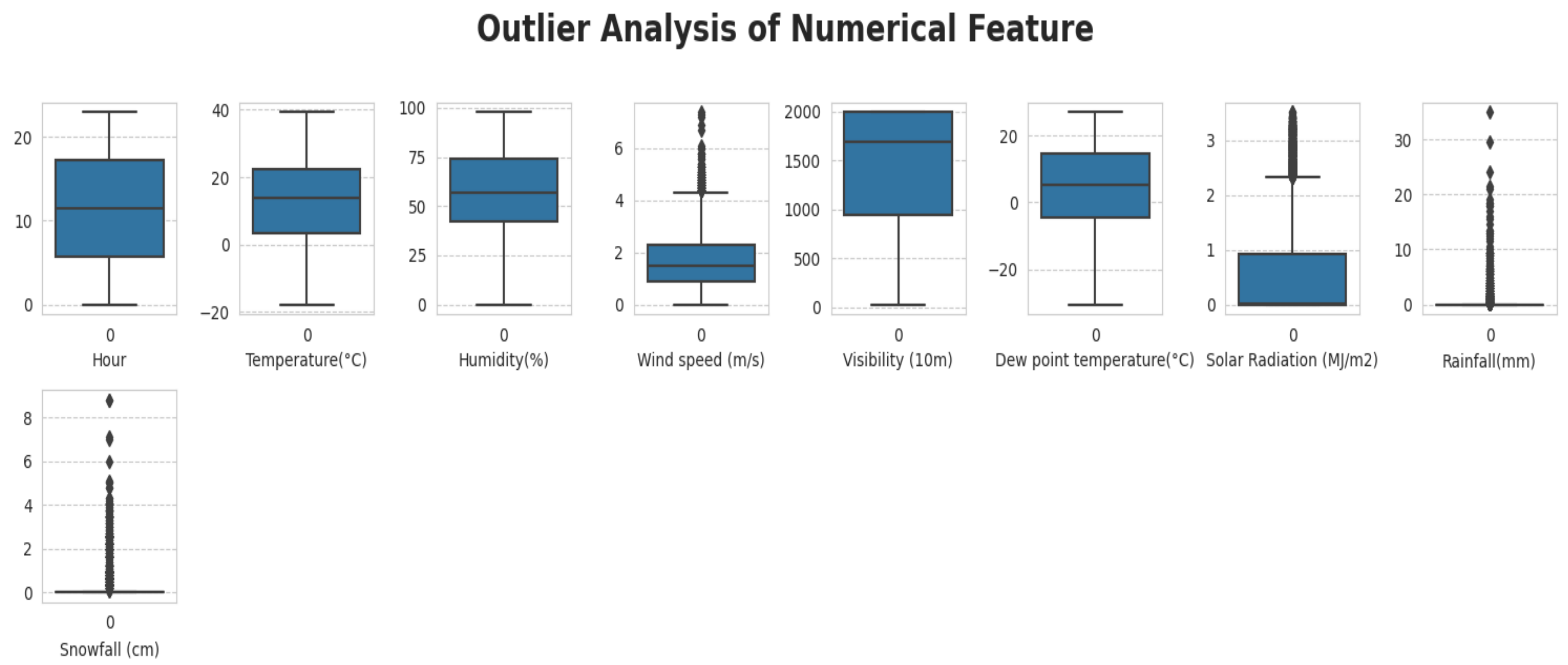


Regression Plot
HOUR
TEMPERATURE(°C)
HUMIDITY(%)
WIND SPEED (M/S)
VISIBILITY (10M)
DEW POINT TEMPERATURE(°C)
SOLAR RADIATION (MJ/M2)
RAINFALL(MM)
SNOWFALL (CM)
Rented Bike Count

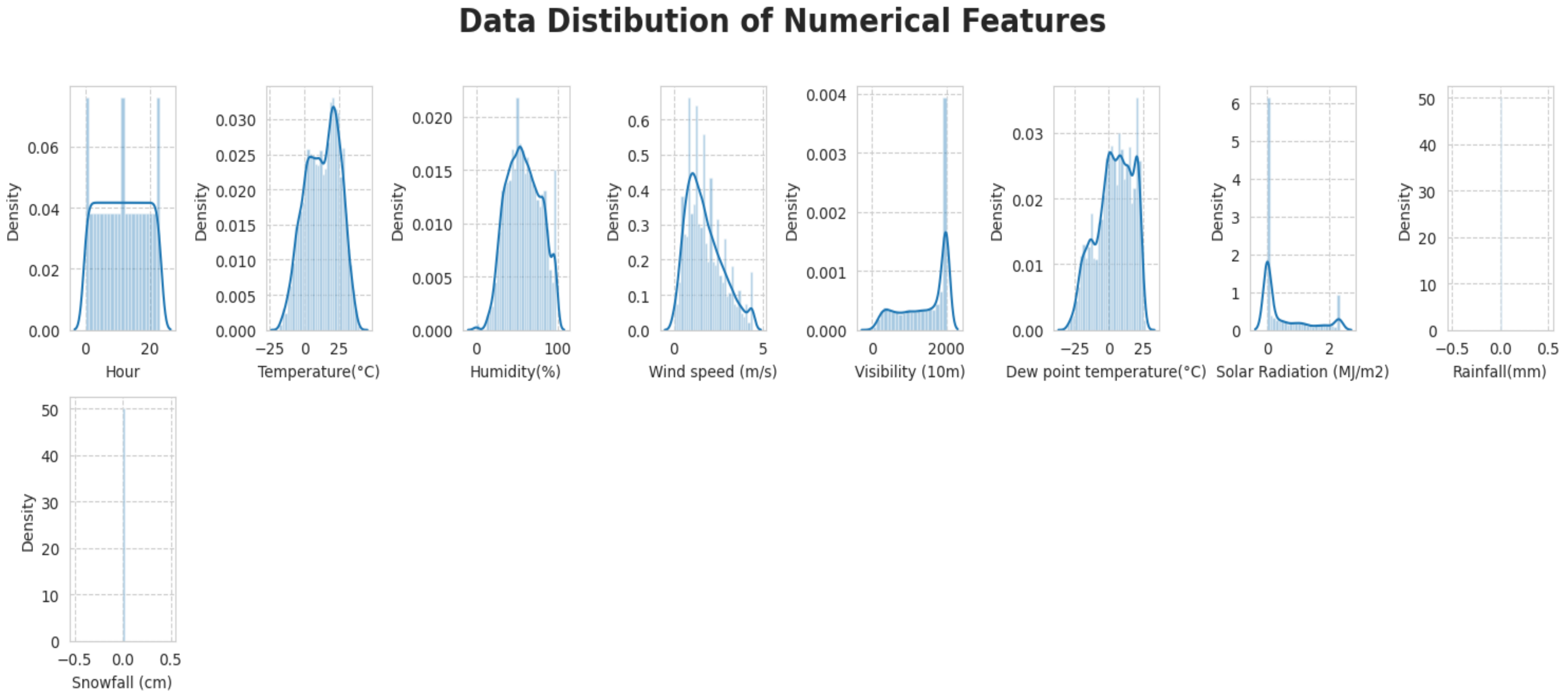
EDA - Feature Correlation



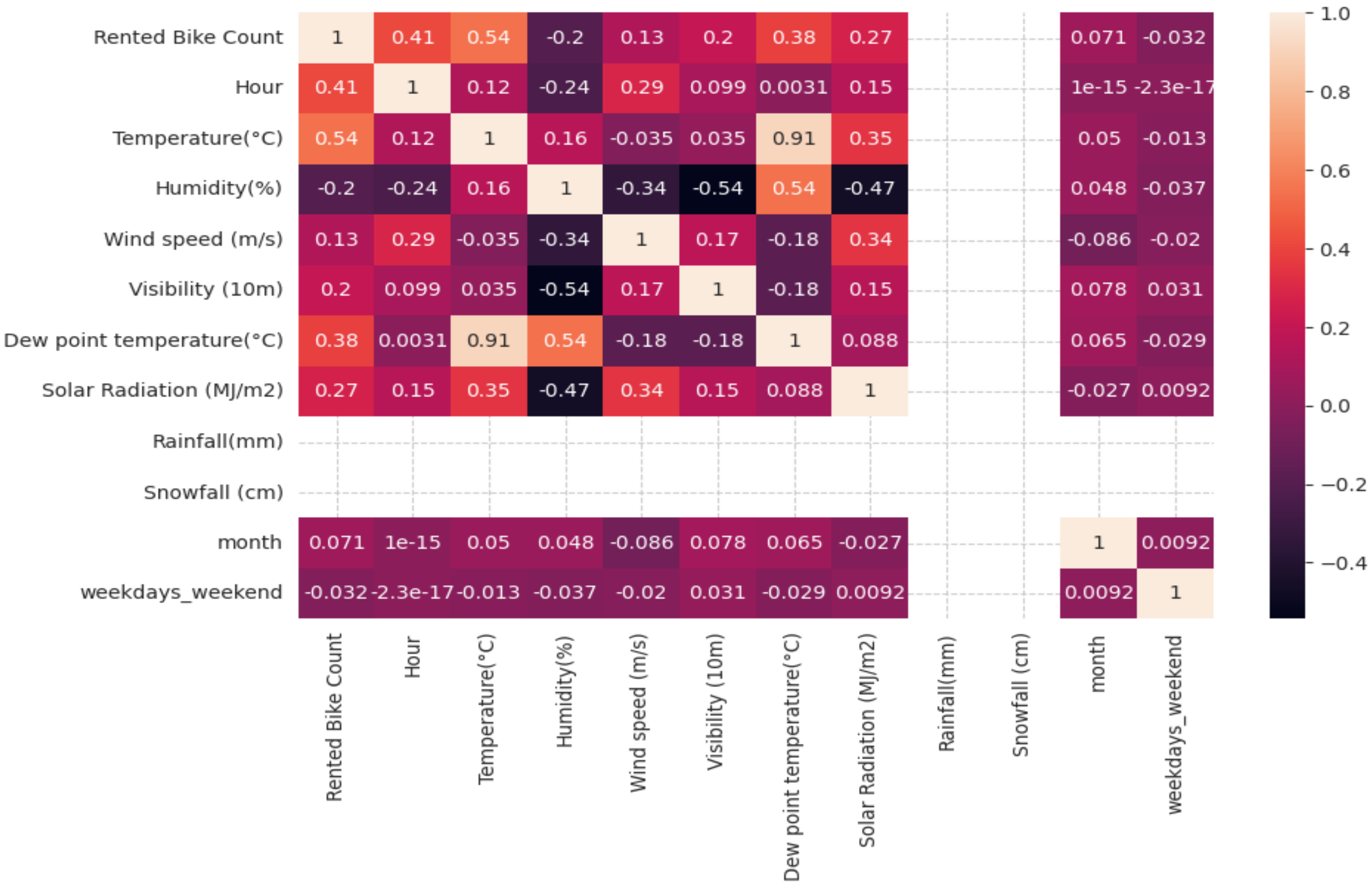
Handling outliers



Checking for distribution for new outliers

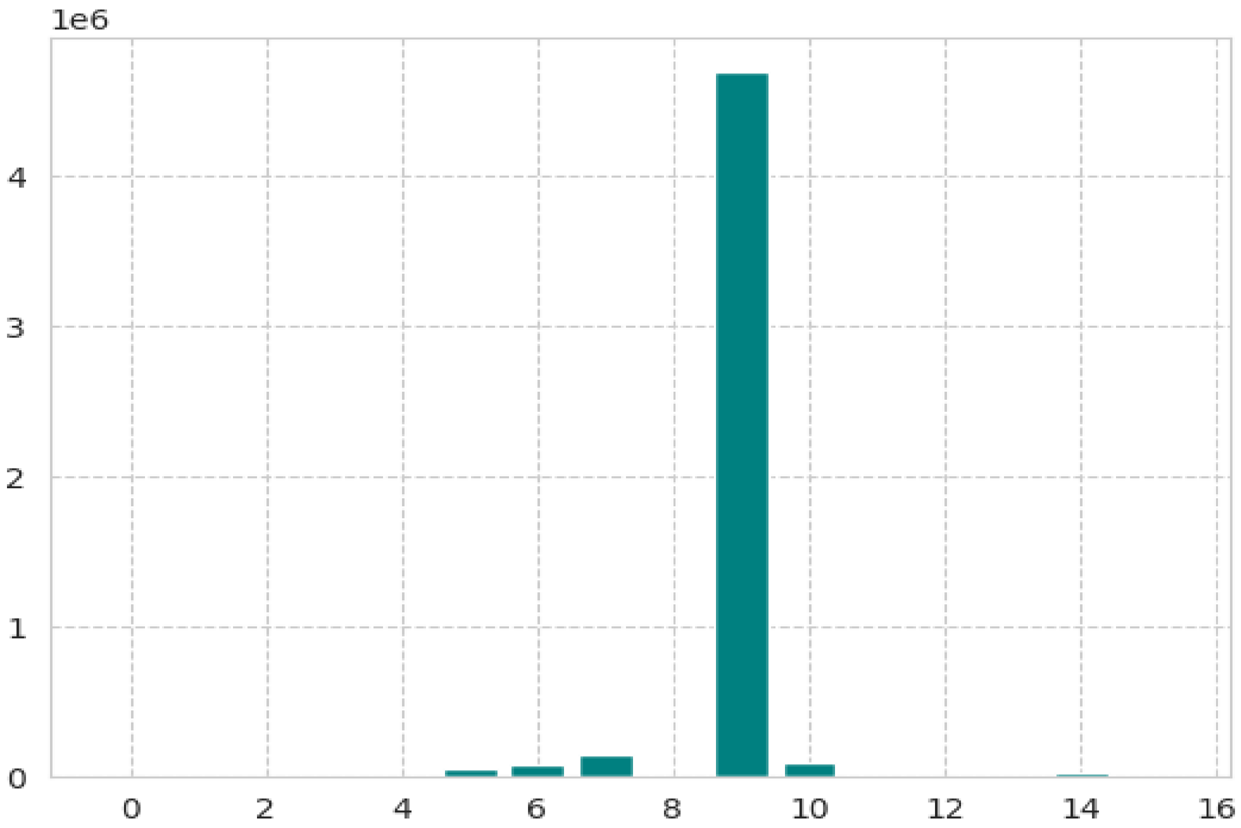


Feature manipulation and selection



Correlation Graph

Feature selection bar graph



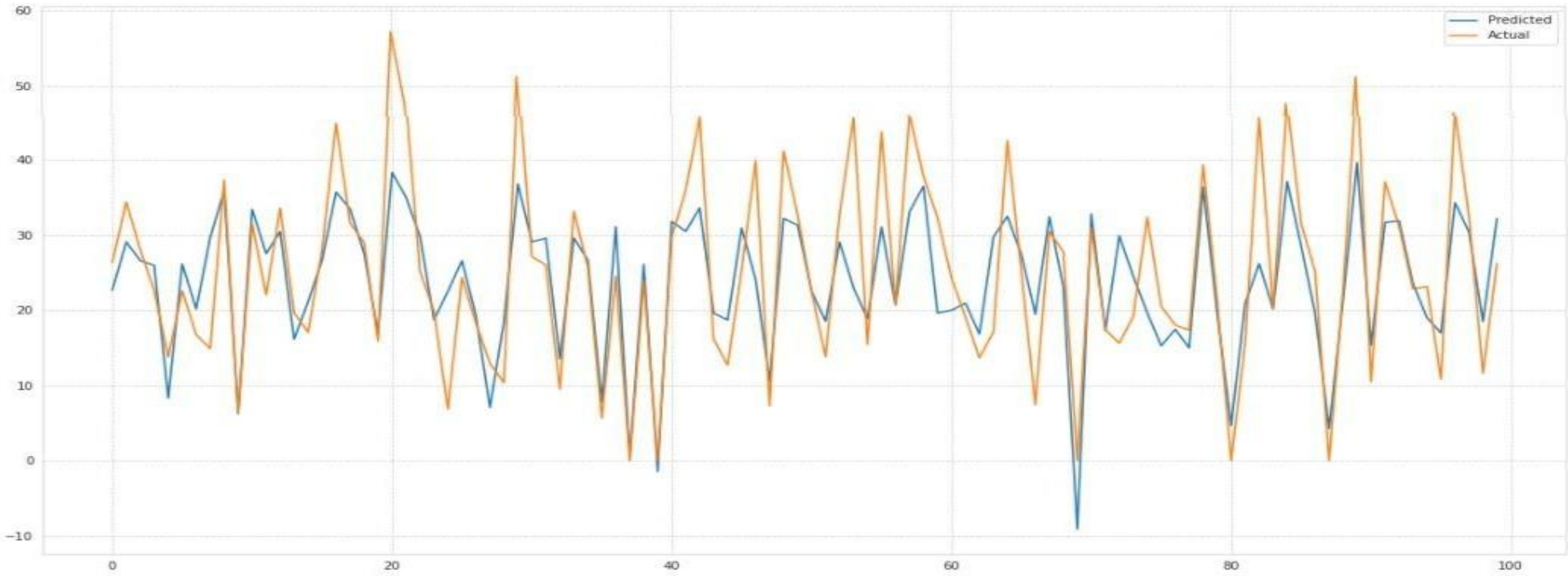
Model’s Performed

- Linear Regression with regularizations
- Lasso Regression
- Ridge Regression
- Decision tree
- Random forest
- Gradient Boosting
- Catboosting
- Bagging
- lightGBM Regressor
- K nearest neighbours

Linear Regression

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

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

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

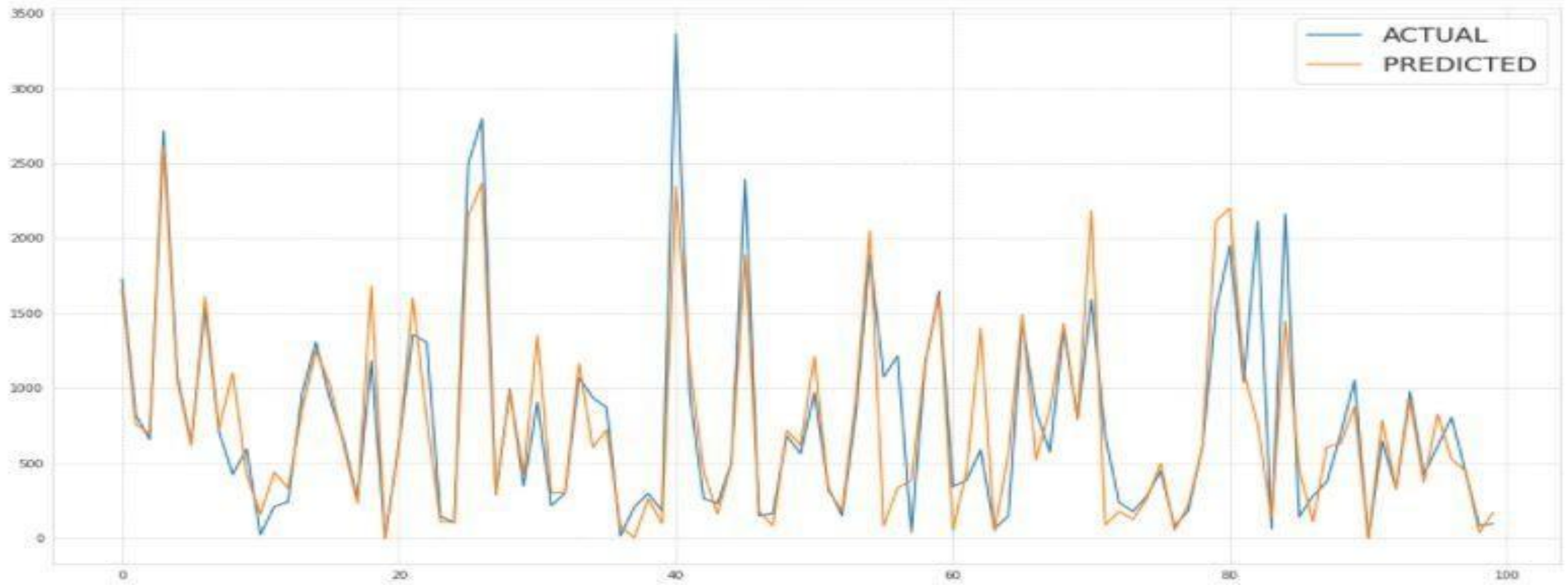
```
 =========================================================  Evalution Graph  ===================================================
```



Decision Tree Regression

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

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

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

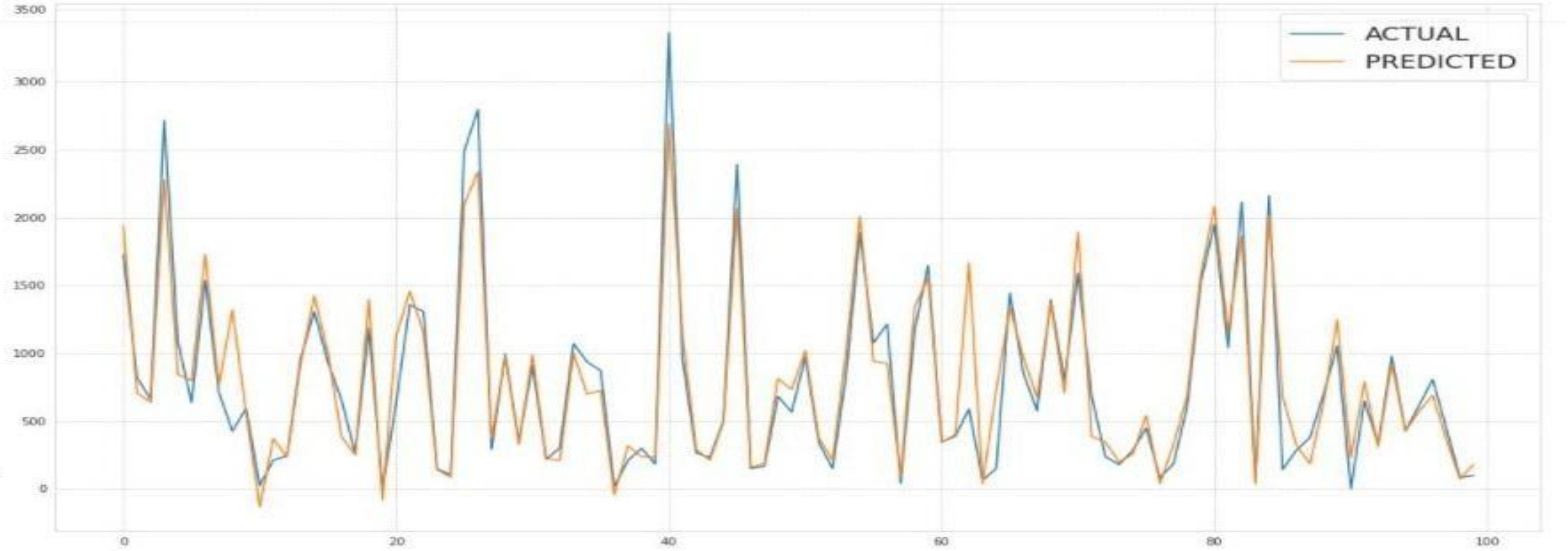
=== Evalution Graph ==



CatBoost

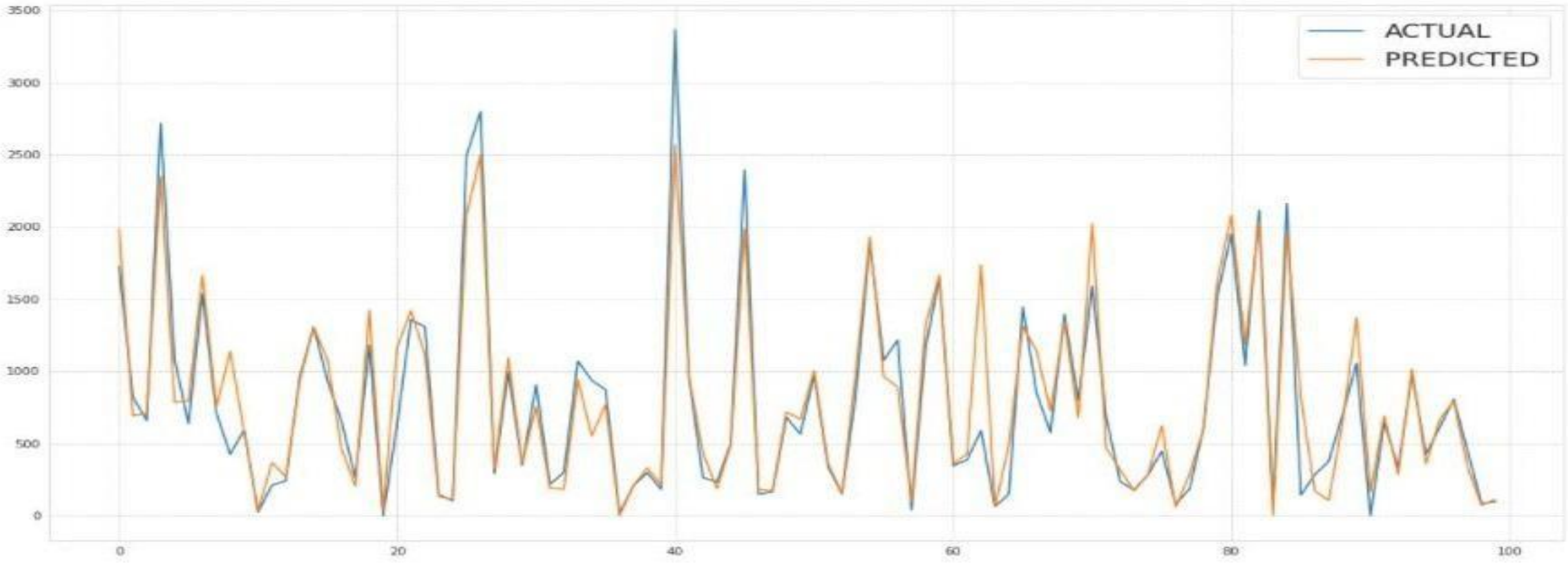
```
================Evalution Matrix========================

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

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

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

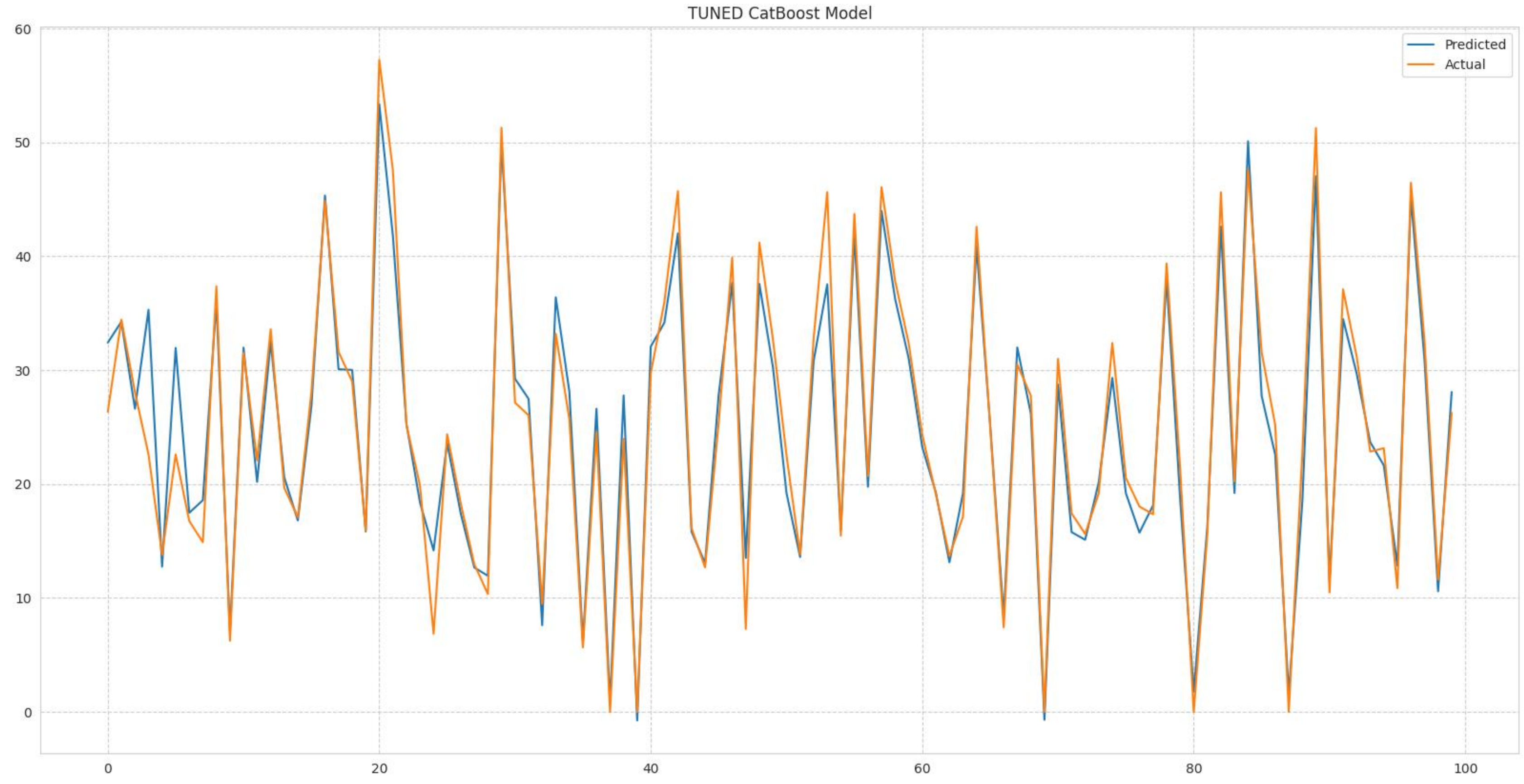


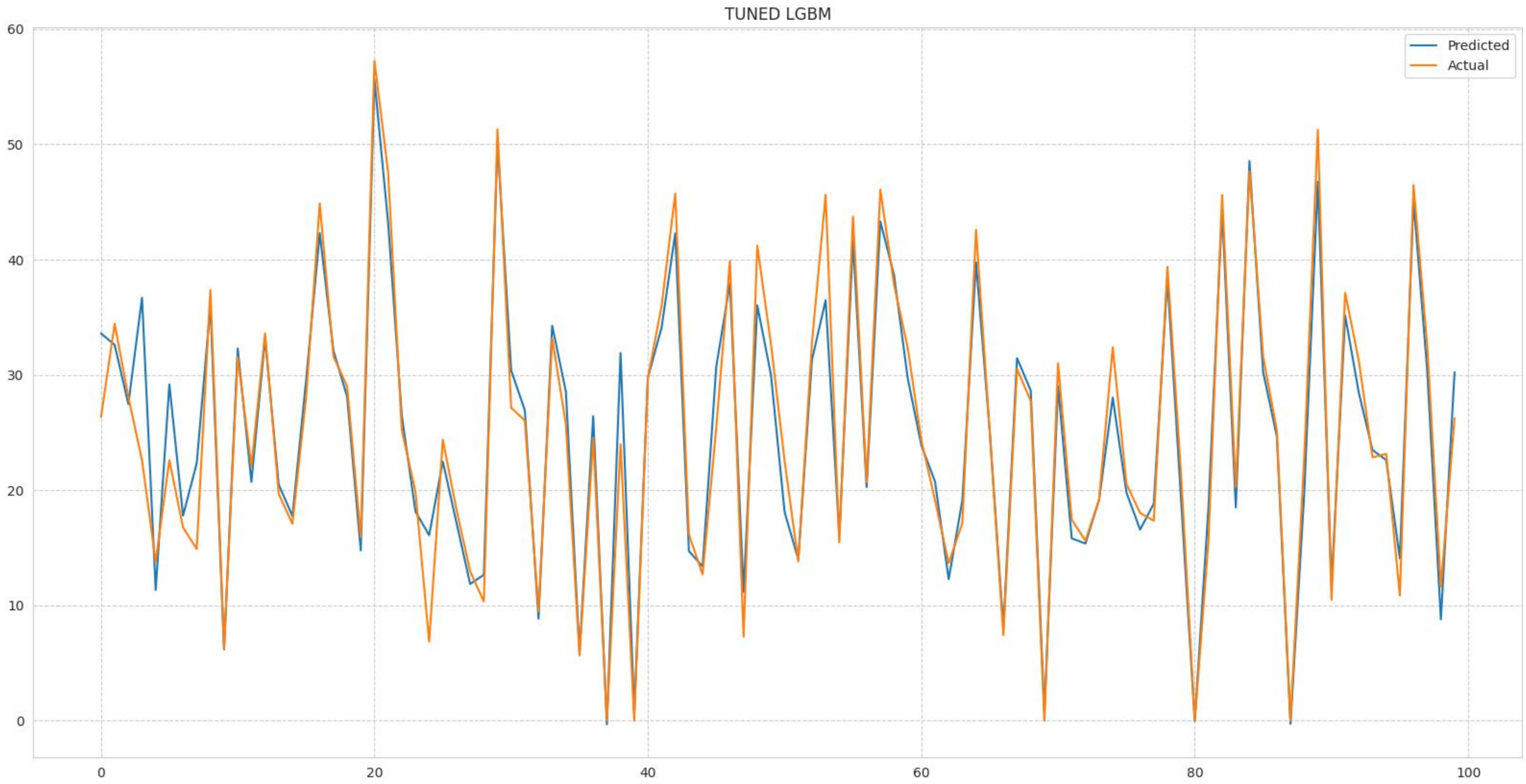
lightGBM

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

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

================Evalution Matrix=========================

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



Cross validation and Hyperparameter tunning




TUNED LGBM
Predicted
Actual
60
50
40
30
20
10
0
0
20
40
60
80
100

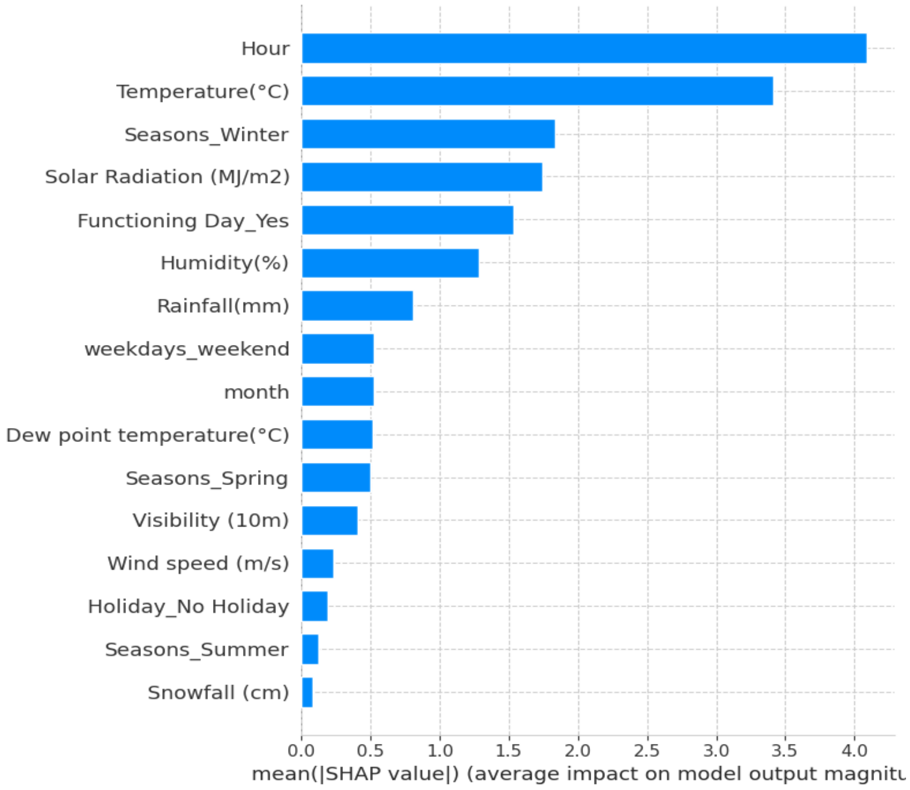
Model Validation & Selection(continued)

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

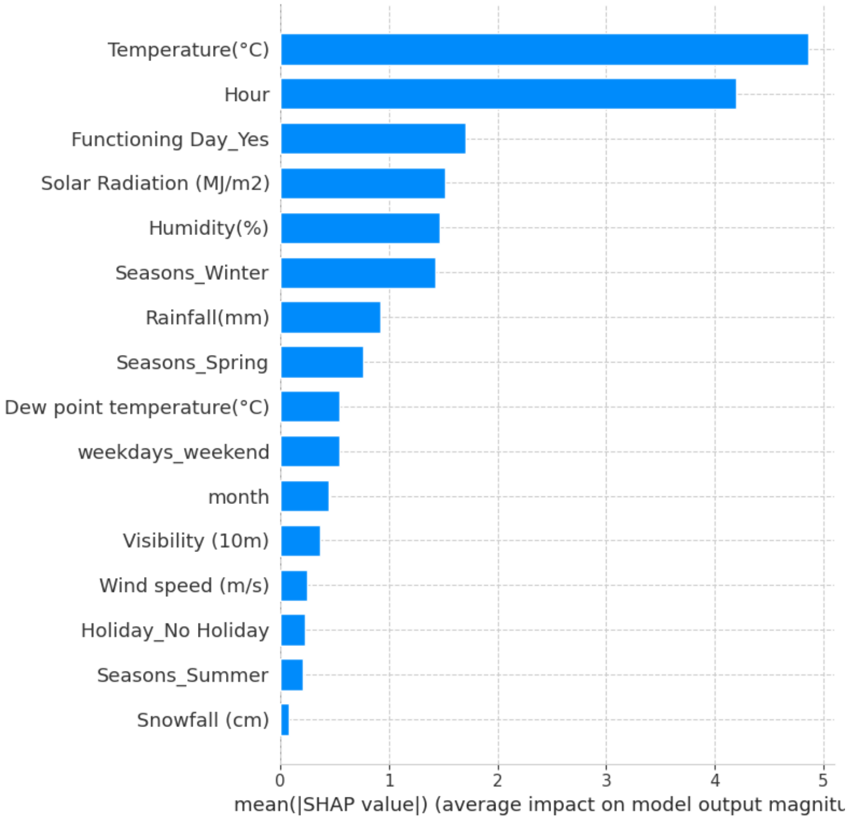
- **Observation 2:** Random forest & GBR have performed equally good in terms of adjusted r2.

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

Feature Importance

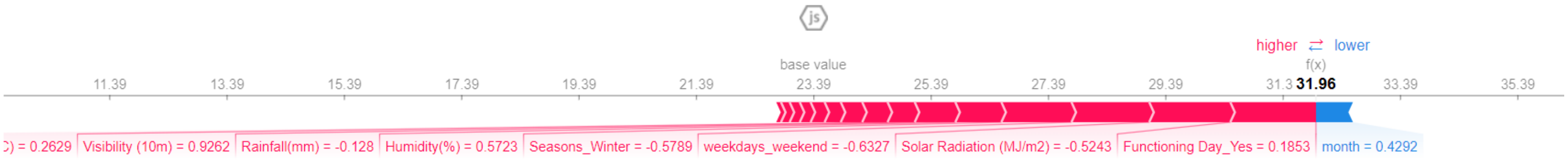


Catboost

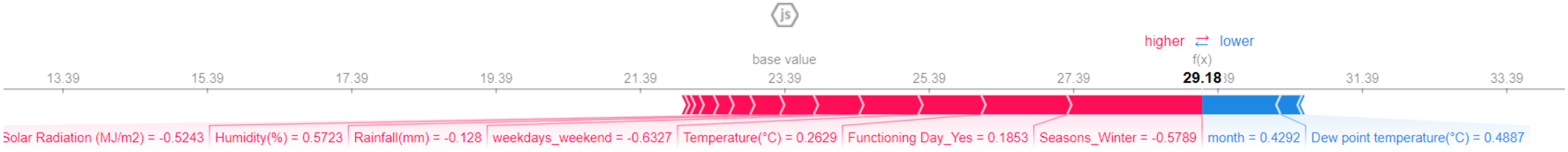


Lightgbm

Model Explainability - SHAP



catboost

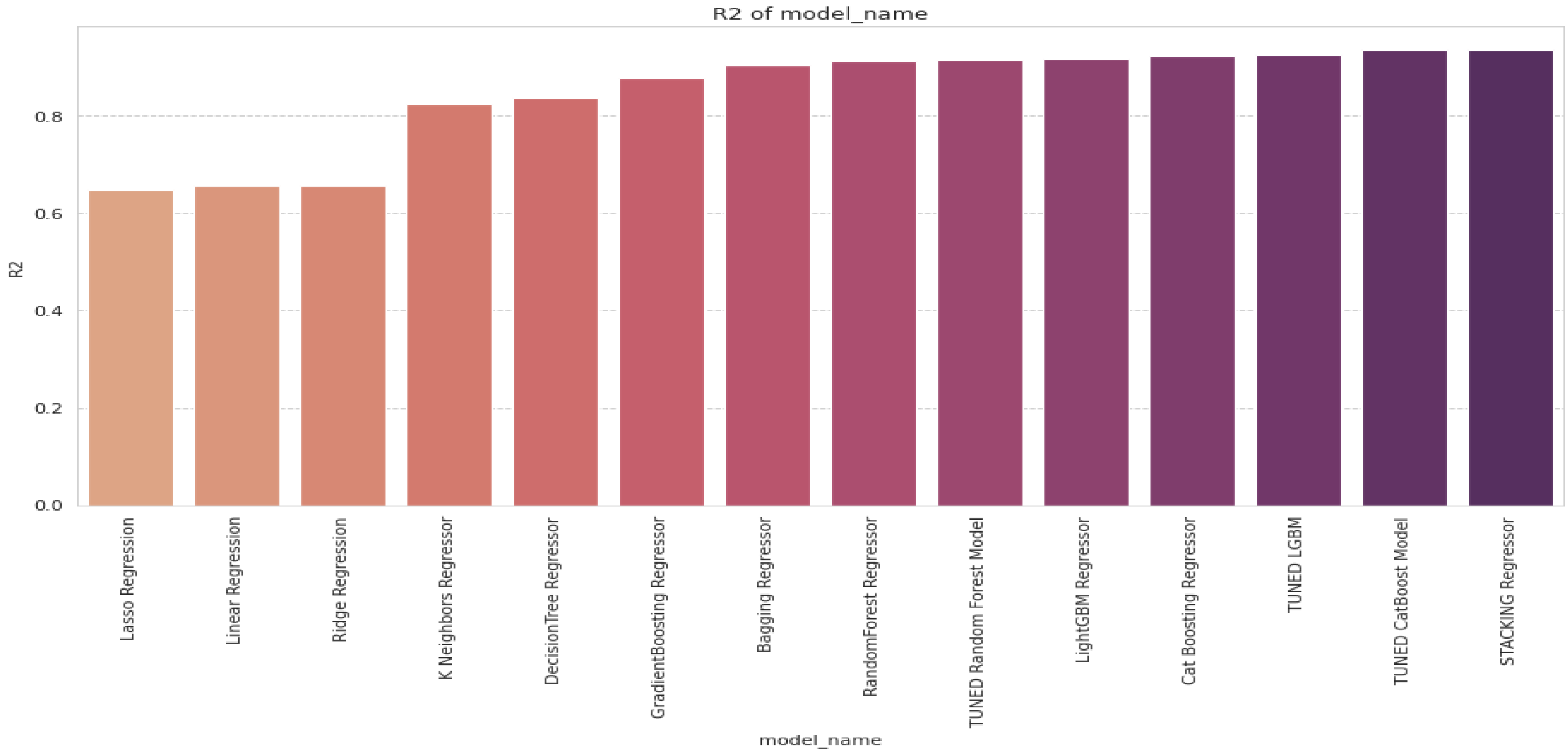


lightgbm

Model’s Evaluation Matrices

	model_type	model_name	rmse	mae	R2	adjusted R2
13	Ensemble Method	STACKING Regressor	3.103857	1.915789	0.937108	0.936645
10	Ensemble Method	TUNED CatBoost Model	3.126705	1.946292	0.936179	0.935709
12	Ensemble Method	TUNED LGBM	3.341796	2.106997	0.927096	0.926560
6	Ensemble Method	Cat Boosting Regressor	3.445109	2.198265	0.922519	0.921949
8	Ensemble Method	LightGBM Regressor	3.536183	2.279485	0.918368	0.917767
11	Ensemble Method	TUNED Random Forest Model	3.608341	2.340785	0.915003	0.914377
4	Ensemble Method	RandomForest Regressor	3.653117	2.330923	0.912880	0.912239
7	Ensemble Method	Bagging Regressor	3.820242	2.450598	0.904727	0.904025
5	Ensemble Method	GradientBoosting Regressor	4.301828	3.066305	0.879192	0.878303
3	CART	DecisionTree Regressor	4.981970	3.144459	0.837972	0.836778
9	Neighbours	K Neighbors Regressor	5.196705	3.656671	0.823703	0.822405
2	Regularized Linear (Ridge)	Ridge Regression	7.253217	5.564770	0.656560	0.654032
0	Linear	Linear Regression	7.253736	5.564918	0.656511	0.653982
1	Regularized Linear (Lasso)	Lasso Regression	7.330763	5.661869	0.649177	0.646594

R2 of Model’s Performed



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

- Upon Exploratory Data Analysis, we found that the bike rentals follow an hourly trend where it hits the first peak in the morning and the highest peak later in the evening.
- We also found that these trends are prominent only during weekdays and working days, leading us to make a safe assumption that office-goers make a notable contribution to bike sharing demand.
- In addition, seasons were observed to have a notable effect on bike rentals with high traffic during summer and a significantly lower demand in winter.
- 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 use either lightGBM or catboost model for the above problem
- Also it can be concluded that the lightGBM and CatBoost models are the best performing models for our project.

THANK YOU