Capstone Project

SEOUL BIKE SHARING DEMAND PREDICTION

By - Rohit Verma

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

Points for Discussion

❑ Data Pipeline

❑ Data Overview

❑ Data Preprocessing

❑ Exploratory Data Analysis

❑ Data Modeling

❑ Model Validation & Selection

❑ Evaluation Matrix of All the models

❑ Model Explainability

❑ Challenges

❑ Conclusion

Data Pipelines

Data Processing-1: In the first process we'have checking the unnecessary features and removed those features which are not relevant for this dataset.

Data Processing-2: In this part, I manually go through all the features selected from first part, then encoded their categorial features, change the columns which containing date time values and also their columns names.

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

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. By creating a model we can easily understand the internal point of the data and make the data to be more elaborative and simple to observe.

Data Summary

Dependent variable:

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

Temporal Features

- Date : year-month-day
- Hour - Hour of the day
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

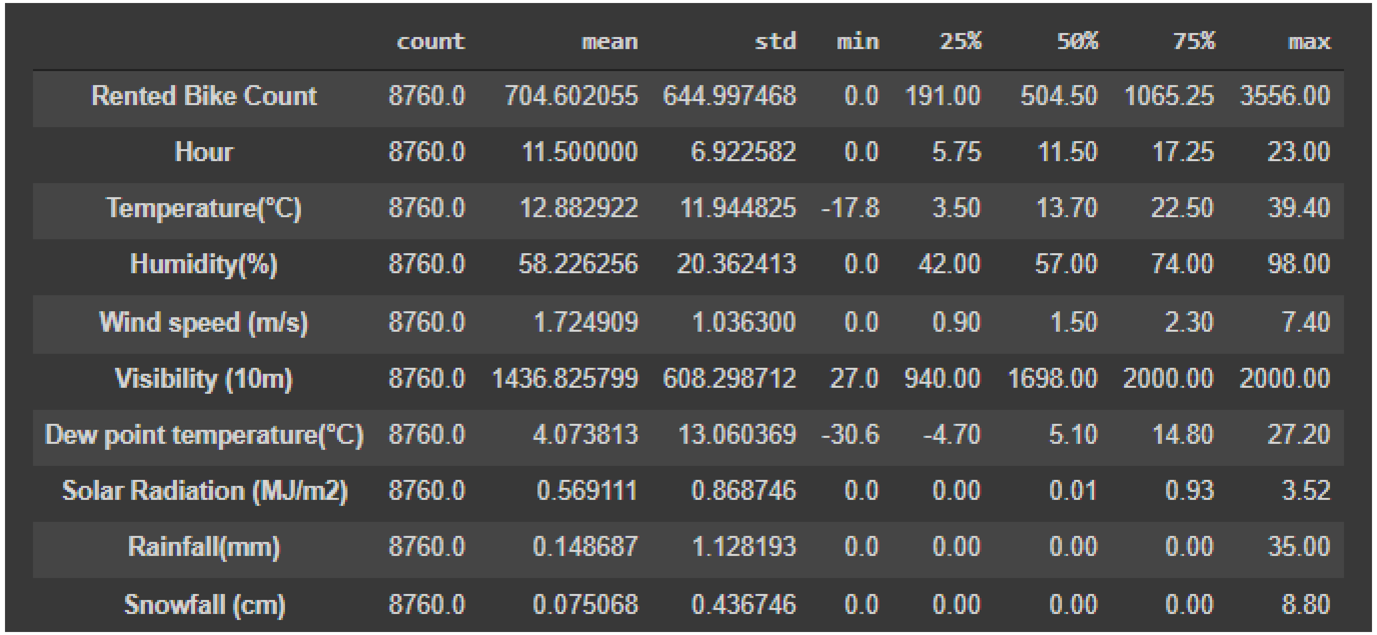
Weather Features

- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10 m
- Solar radiation - MJ/m2
- Rainfall - mm
- Snowfall - cm
- Seasons - Winter, Spring, Summer, Autumn

Data Preprocessing

The rented bike count is presented as a time series which represents the data with a step of an hour.

- There are 8760 rows and 14 columns
- It shows clearly that **the dataframe mainly represents weather conditions and information about the day.**



	count	mean	std	min	25%	50%	75%	max
Rented Bike Count	8760.0	704.602055	644.997468	0.0	191.00	504.50	1065.25	3556.00
Hour	8760.0	11.500000	6.922582	0.0	5.75	11.50	17.25	23.00
Temperature(°C)	8760.0	12.882922	11.944825	-17.8	3.50	13.70	22.50	39.40
Humidity(%)	8760.0	58.226256	20.362413	0.0	42.00	57.00	74.00	98.00
Wind speed (m/s)	8760.0	1.724909	1.036300	0.0	0.90	1.50	2.30	7.40
Visibility (10m)	8760.0	1436.825799	608.298712	27.0	940.00	1698.00	2000.00	2000.00
Dew point temperature(°C)	8760.0	4.073813	13.060369	-30.6	-4.70	5.10	14.80	27.20
Solar Radiation (MJ/m2)	8760.0	0.569111	0.868746	0.0	0.00	0.01	0.93	3.52
Rainfall(mm)	8760.0	0.148687	1.128193	0.0	0.00	0.00	0.00	35.00
Snowfall (cm)	8760.0	0.075068	0.436746	0.0	0.00	0.00	0.00	8.80

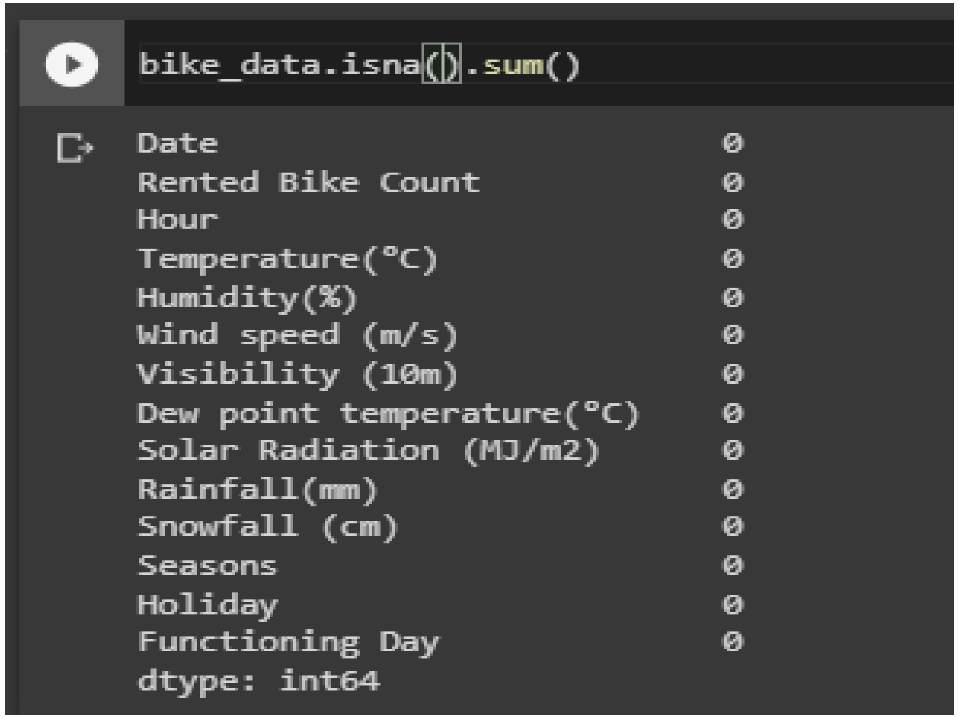
Data Preprocessing (Cont’d)

Data Cleaning

The dataset has 8760 rows with different types of data and columns.

- We start our exploration by reducing the number of rows & columns based on the previous criteria

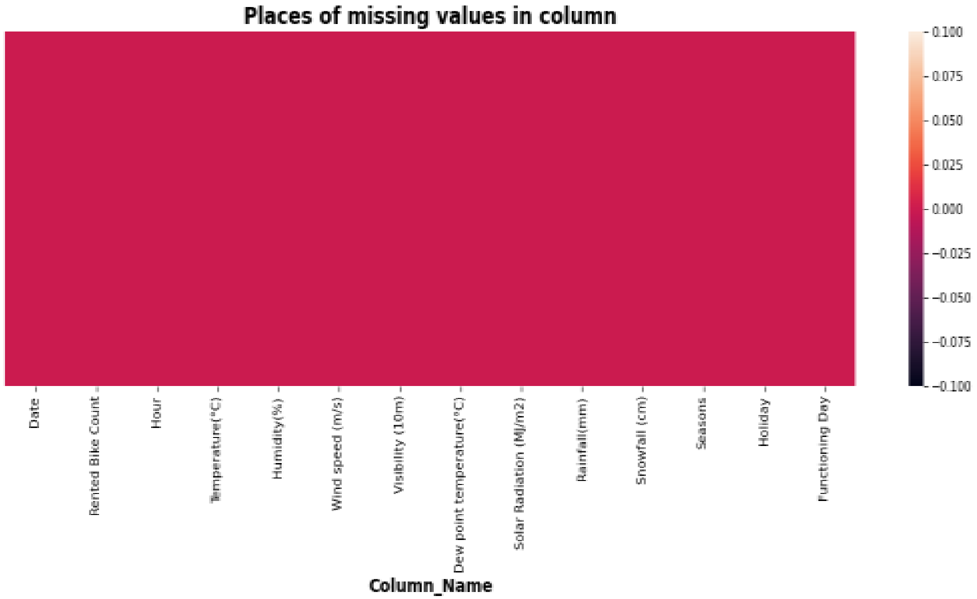
- Checking null or unnecessary values



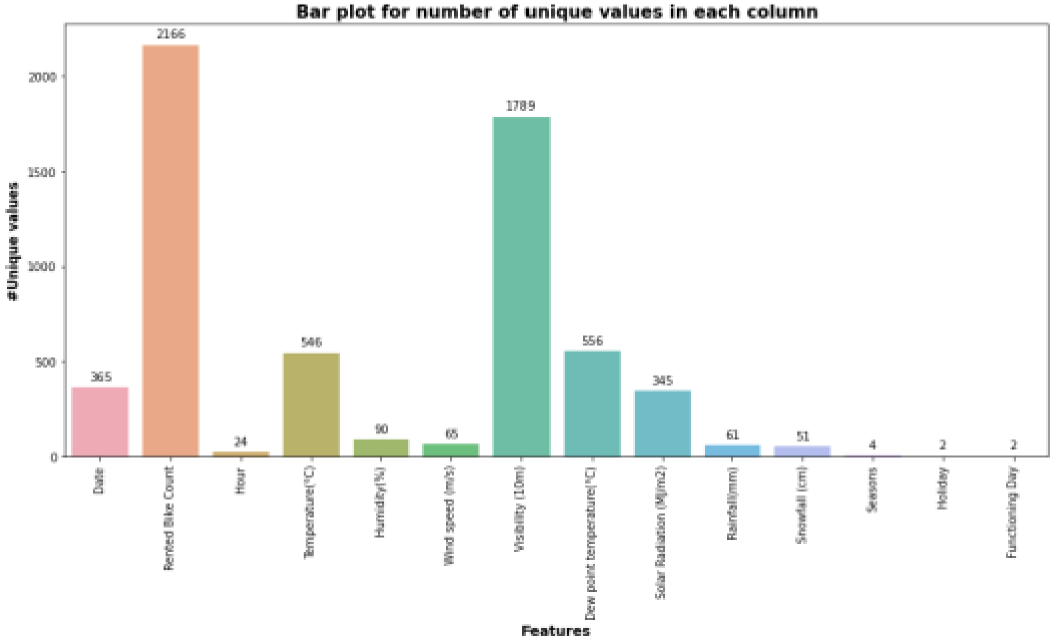
```
bike_data.isna().sum()
```

```
Date                         0
Rented Bike Count            0
Hour                         0
Temperature(°C)              0
Humidity(%)                  0
Wind speed (m/s)             0
Visibility (10m)             0
Dew point temperature(°C)    0
Solar Radiation (MJ/m2)      0
Rainfall(mm)                 0
Snowfall (cm)                0
Seasons                      0
Holiday                      0
Functioning Day              0
dtype: int64
```

Data Preprocessing (Cont’d)



This plot shows there is no null values in our dataset



This plot shows that the maximum no of unique values are present on Rented Bike Count

Correlation Analysis

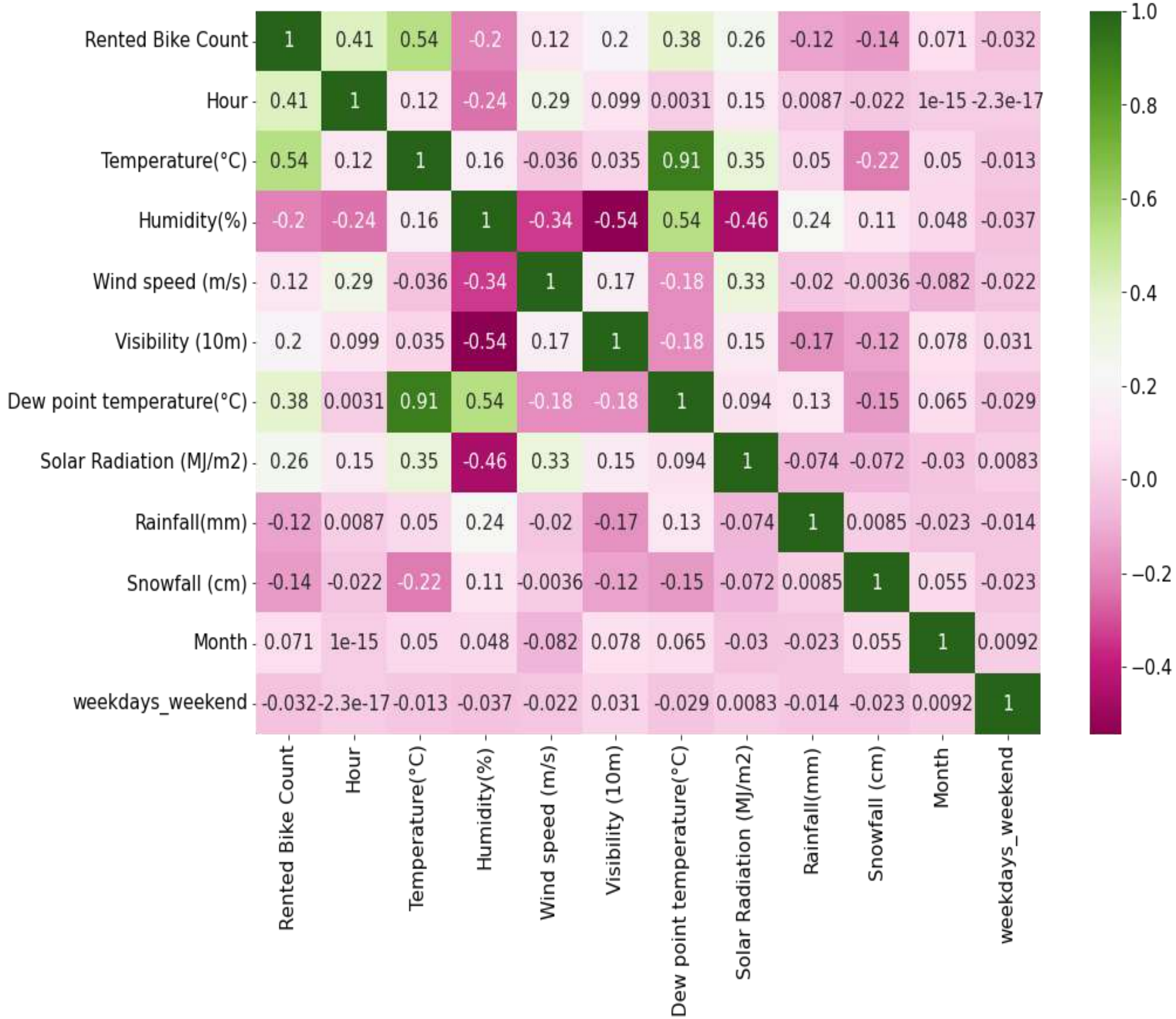
We can see that on the target variable line the most correlated variables to the rent are

1.the hour

2.the temperature

3.the dew point temperature

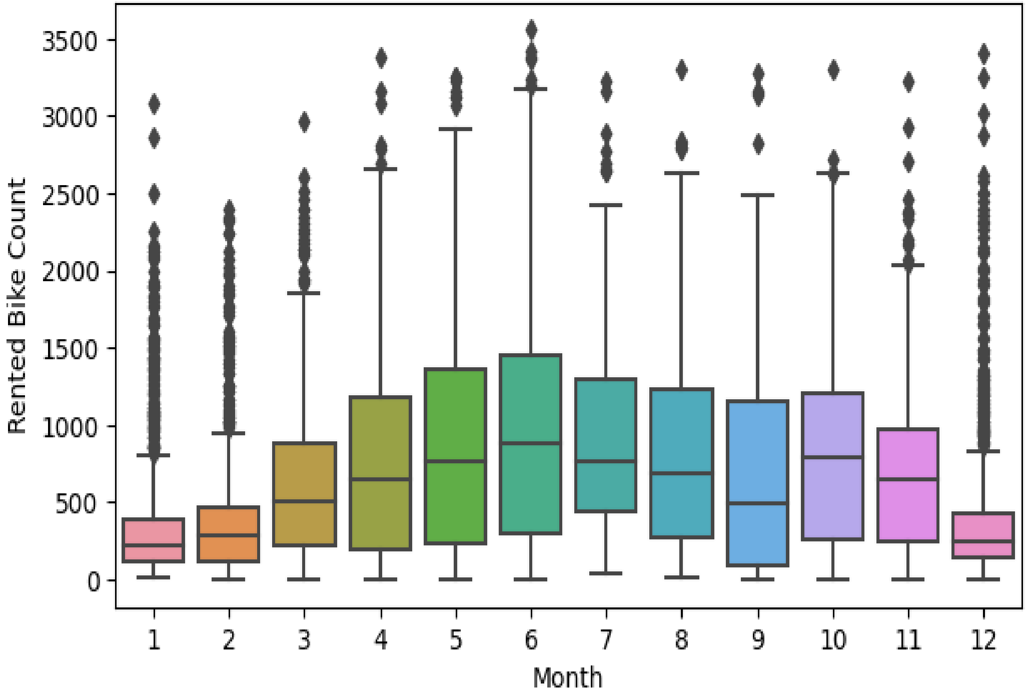
4.the solar radiation (The dew point is a temperature which is so cold that the vapor becomes liquid)

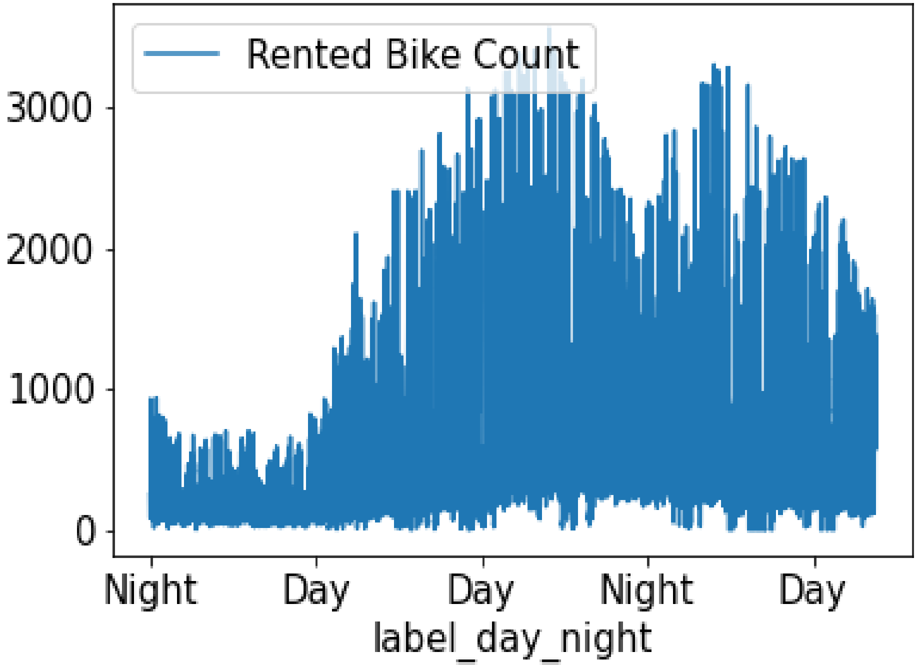


	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Month	weekdays_weekend
Rented Bike Count	1	0.41	0.54	-0.2	0.12	0.2	0.38	0.26	-0.12	-0.14	0.071	-0.032
Hour	0.41	1	0.12	-0.24	0.29	0.099	0.0031	0.15	0.0087	-0.022	1e-15	-2.3e-17
Temperature(°C)	0.54	0.12	1	0.16	-0.036	0.035	0.91	0.35	0.05	-0.22	0.05	-0.013
Humidity(%)	-0.2	-0.24	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11	0.048	-0.037
Wind speed (m/s)	0.12	0.29	-0.036	-0.34	1	0.17	-0.18	0.33	-0.02	-0.0036	-0.082	-0.022
Visibility (10m)	0.2	0.099	0.035	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12	0.078	0.031
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	-0.18	-0.18	1	0.094	0.13	-0.15	0.065	-0.029
Solar Radiation (MJ/m2)	0.26	0.15	0.35	-0.46	0.33	0.15	0.094	1	-0.074	-0.072	-0.03	0.0083
Rainfall(mm)	-0.12	0.0087	0.05	0.24	-0.02	-0.17	0.13	-0.074	1	0.0085	-0.023	-0.014
Snowfall (cm)	-0.14	-0.022	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	1	0.055	-0.023
Month	0.071	1e-15	0.05	0.048	-0.082	0.078	0.065	-0.03	-0.023	0.055	1	0.0092
weekdays_weekend	-0.032	-2.3e-17	-0.013	-0.037	-0.022	0.031	-0.029	0.0083	-0.014	-0.023	0.0092	1

Explorative Data Analysis (EDA)

DATE:





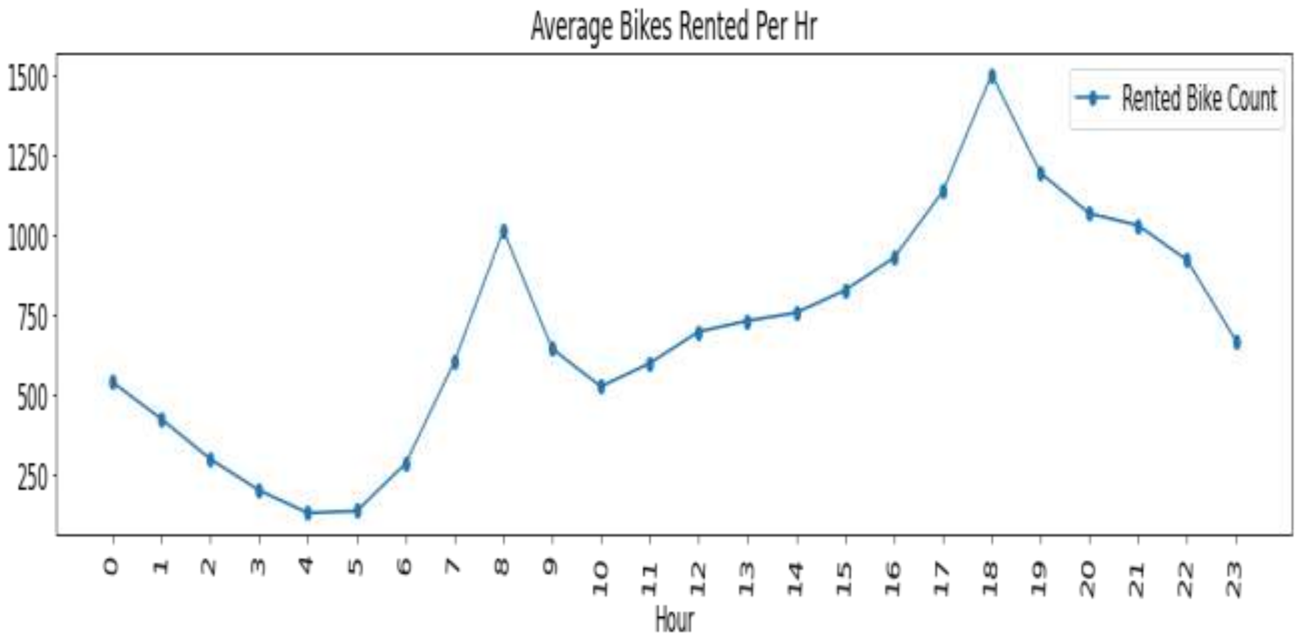
We can see a high demand in April to autumn of bikes rent and there less demand of Rented bike in the month of December, January and February

There are much more rents during the day than the night

EDA (Cont’d)

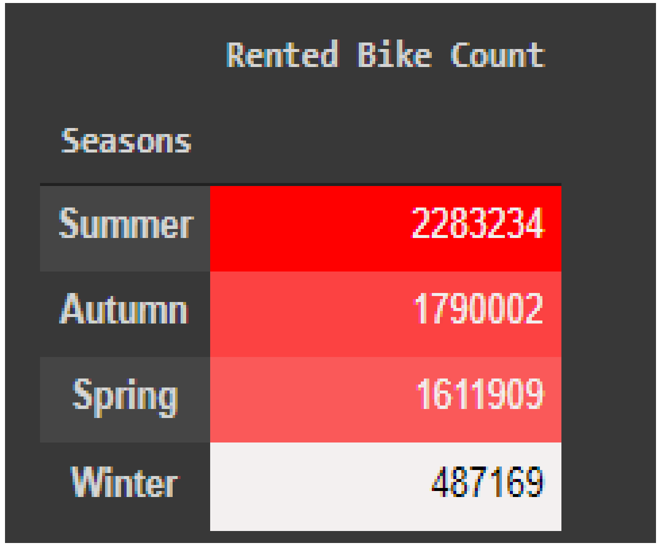
HOUR:

- High rise of Rented Bikes from 8:00 a.m to 9:00 p.m means people prefer rented bike during rush hour.
- We can clearly see that demand rises most at 8 a.m and 6:00 p.m so we can say that that during office opening and closing time there is much high demand

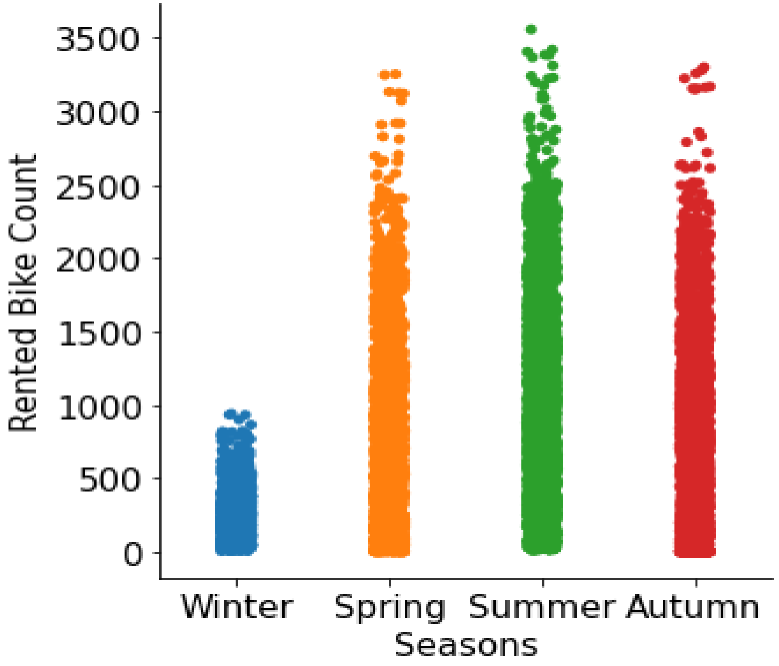


EDA (Cont’d)

SEASON:



Seasons	Rented Bike Count
Summer	2283234
Autumn	1790002
Spring	1611909
Winter	487169

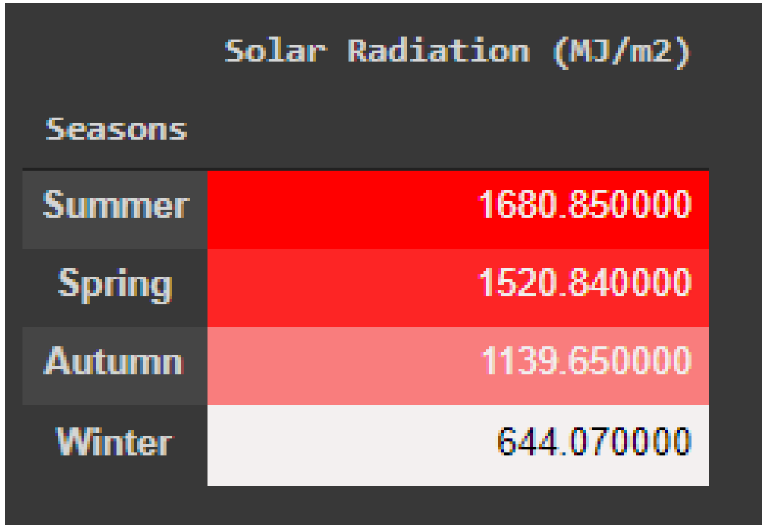


We can see that as we expected, summer is the season in which we have the most rents

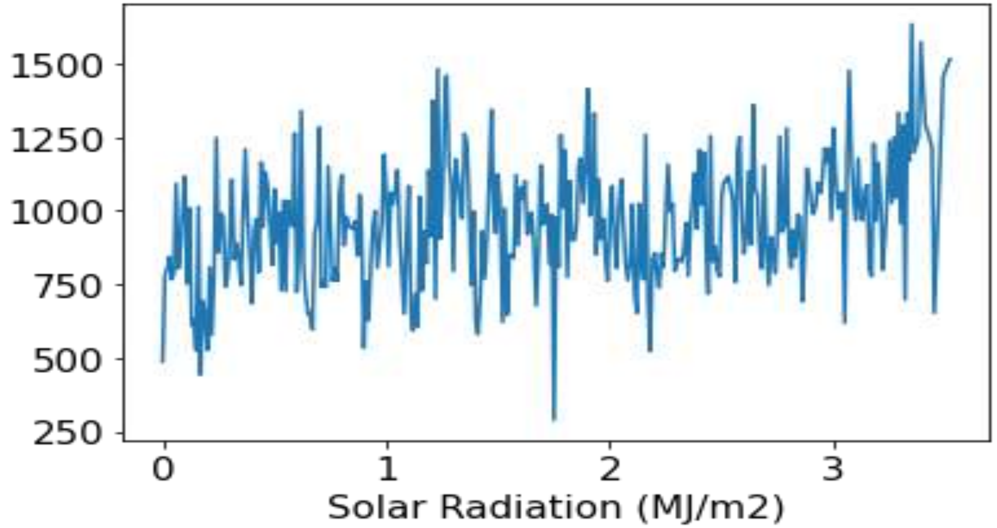
We can clearly see that there is less demand of rented bike during winter season

EDA (Cont’d)

SOLAR RADIATION:



	Solar Radiation (MJ/m2)
Seasons	
Summer	1680.850000
Spring	1520.840000
Autumn	1139.650000
Winter	644.070000



It shows high correlation of bikes with the feature 'solar Radiation' and 'Summer

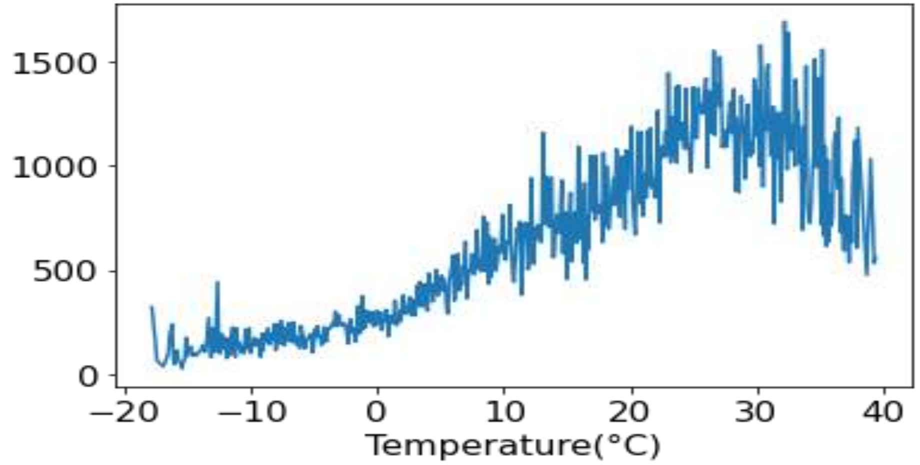
Here the amount of rented bikes is huge, when there is solar radiation

EDA (Cont'd)

TEMPERATURE:

- Steady increase in bike count with temperature.

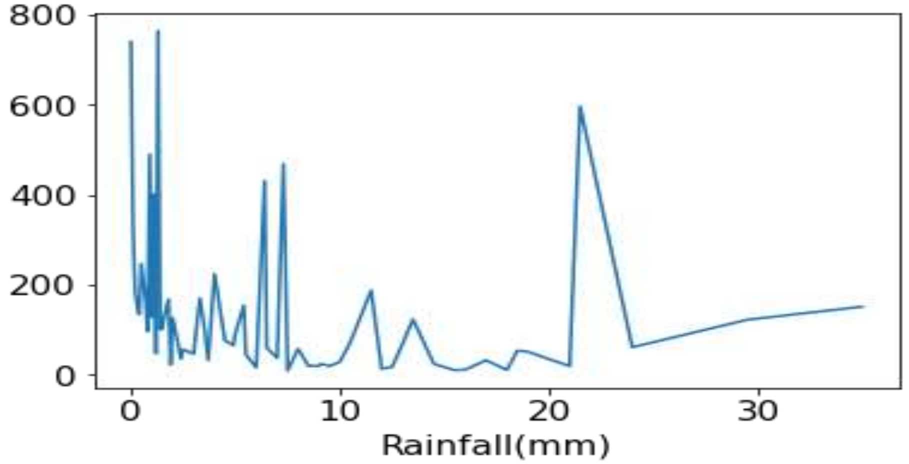
- Ideal temperature for biking is between 32 and 36 degree.



RAINFALL:

- We can see that even if it rains a lot of Korean rent a bikes

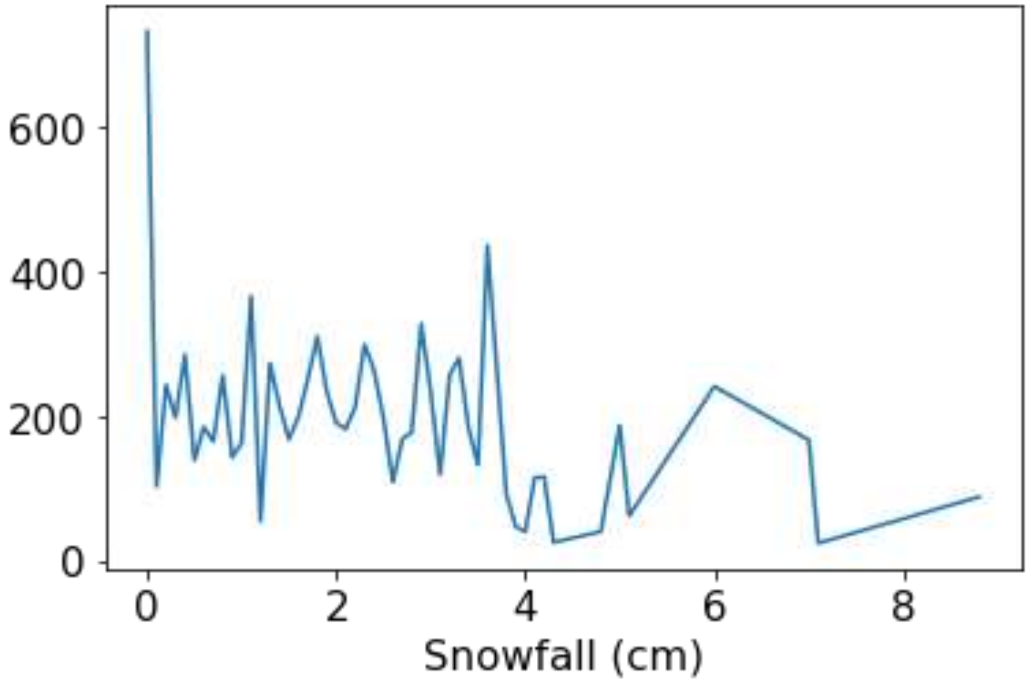
- This raise between 20 and 25 mm of rainfall seems very contradictory



EDA (Cont'd)

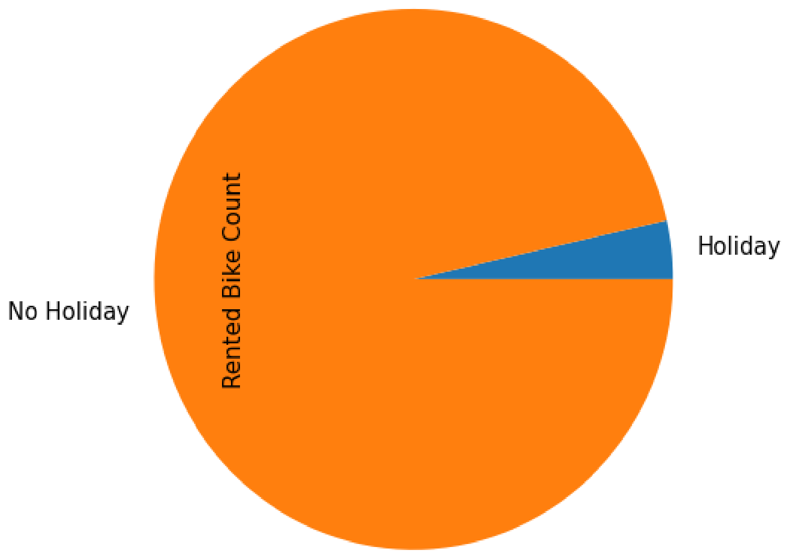
SNOWFALL:

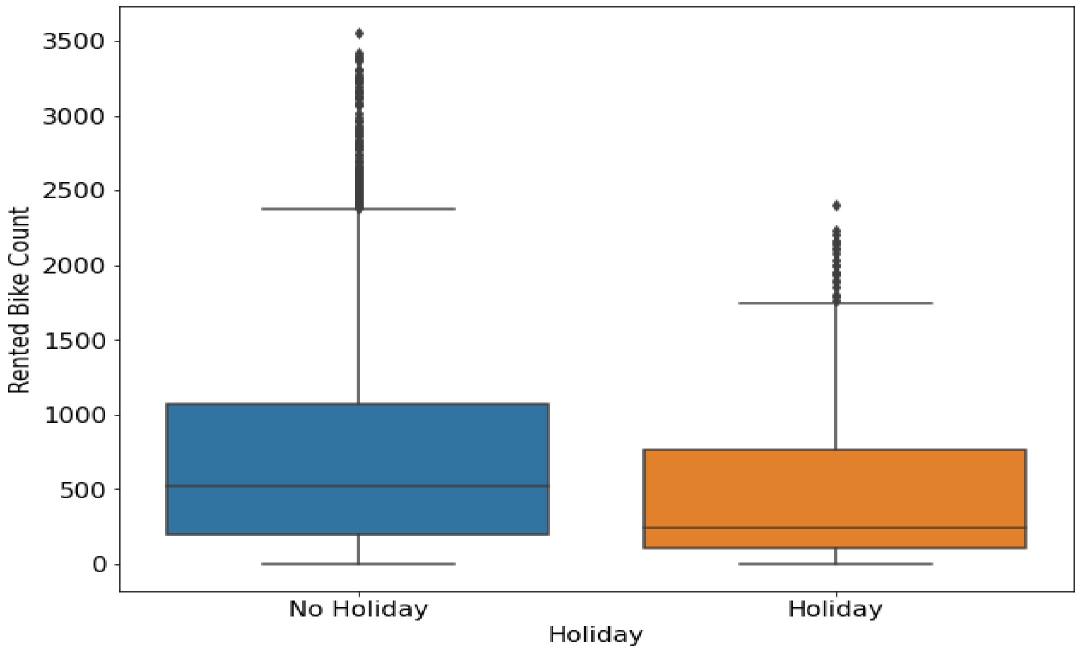
- We can see on the y-axis, the amount of rents is very low.
- So when we have more than 4 cm of snow, the bike rents is much lower



EDA (Cont'd)

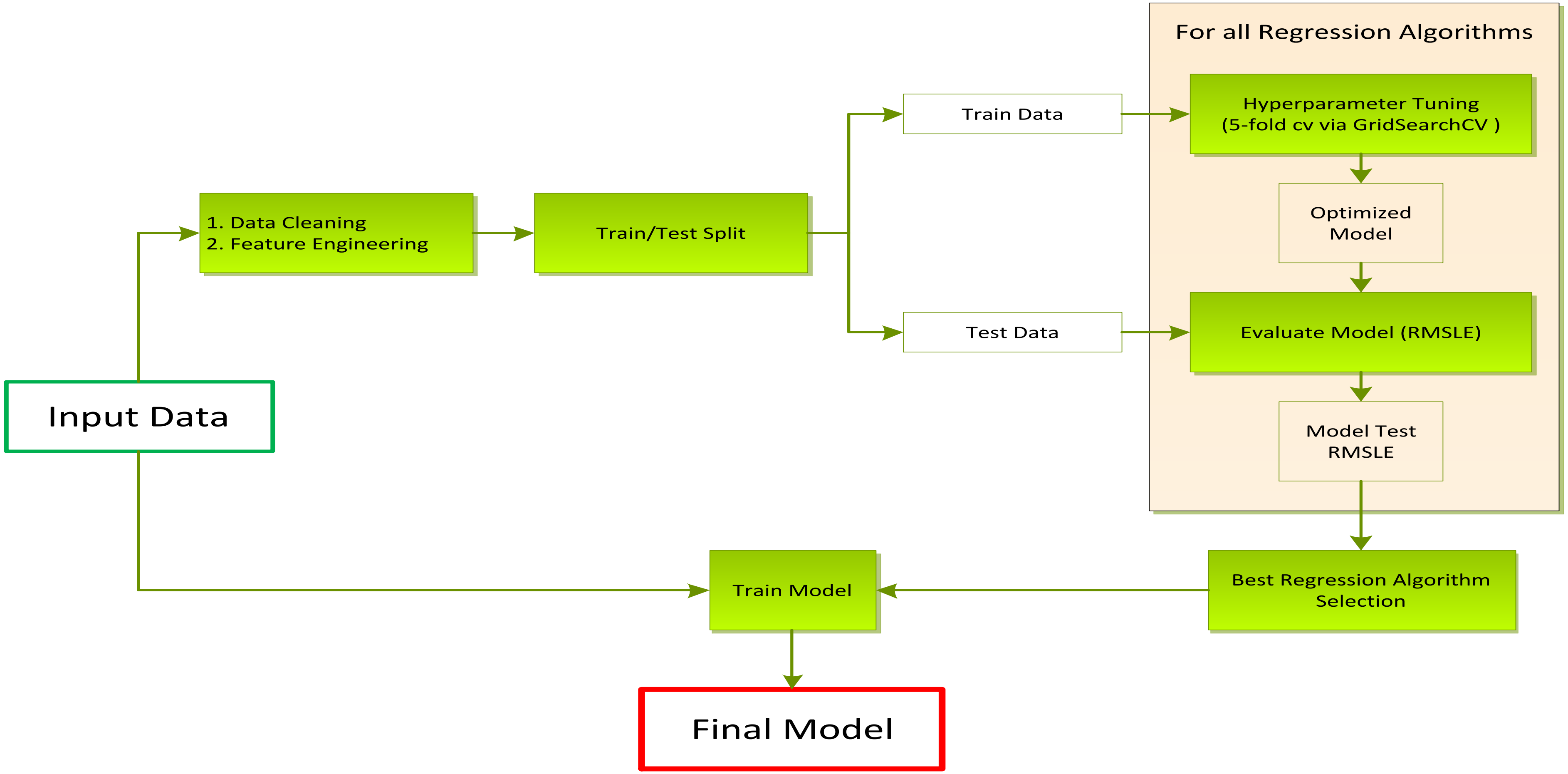
HOLIDAY:





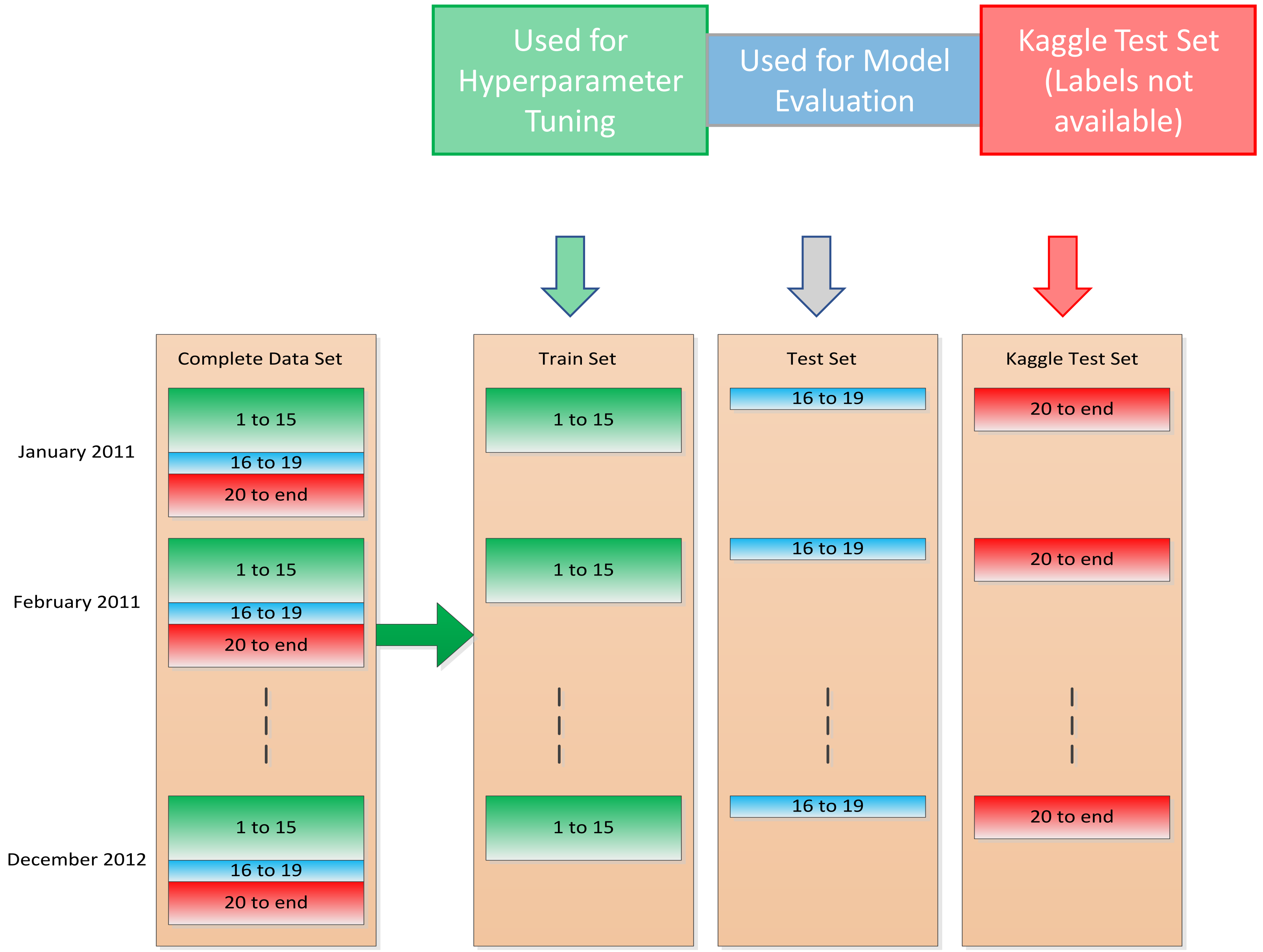
The shape of this DF is 432 lines and 16 columns, it means that there are only 18 days of holidays in Korea which is very short. And that explains why the proportion of rented bike during holiday is low

Data Modeling



Data Modeling (Cont’d)

TRAIN/TEST SPLIT



Used for Hyperparameter Tuning

Used for Model Evaluation

Kaggle Test Set (Labels not available)

Complete Data Set

January 2011

1 to 15

16 to 19

20 to end

February 2011

1 to 15

16 to 19

20 to end

December 2012

1 to 15

16 to 19

20 to end

Train Set

1 to 15

1 to 15

1 to 15

Test Set

16 to 19

16 to 19

16 to 19

Kaggle Test Set

20 to end

20 to end

20 to end

Data Modeling (Cont’d)

Hyperparameter Tuning

5 Hyperparameter turned

- N_estimators= number of bikes on rent
- max_features= max number of features considerwd for splitting a node= ‘auto’= all features used
- min_sample_leaf= min number of samples allowed in a leaf node
- max_depth= max numbers of level in each steps
- min_sample_split=min number of data points placed in a node before the node is split^2

Data Modeling (Cont’d)

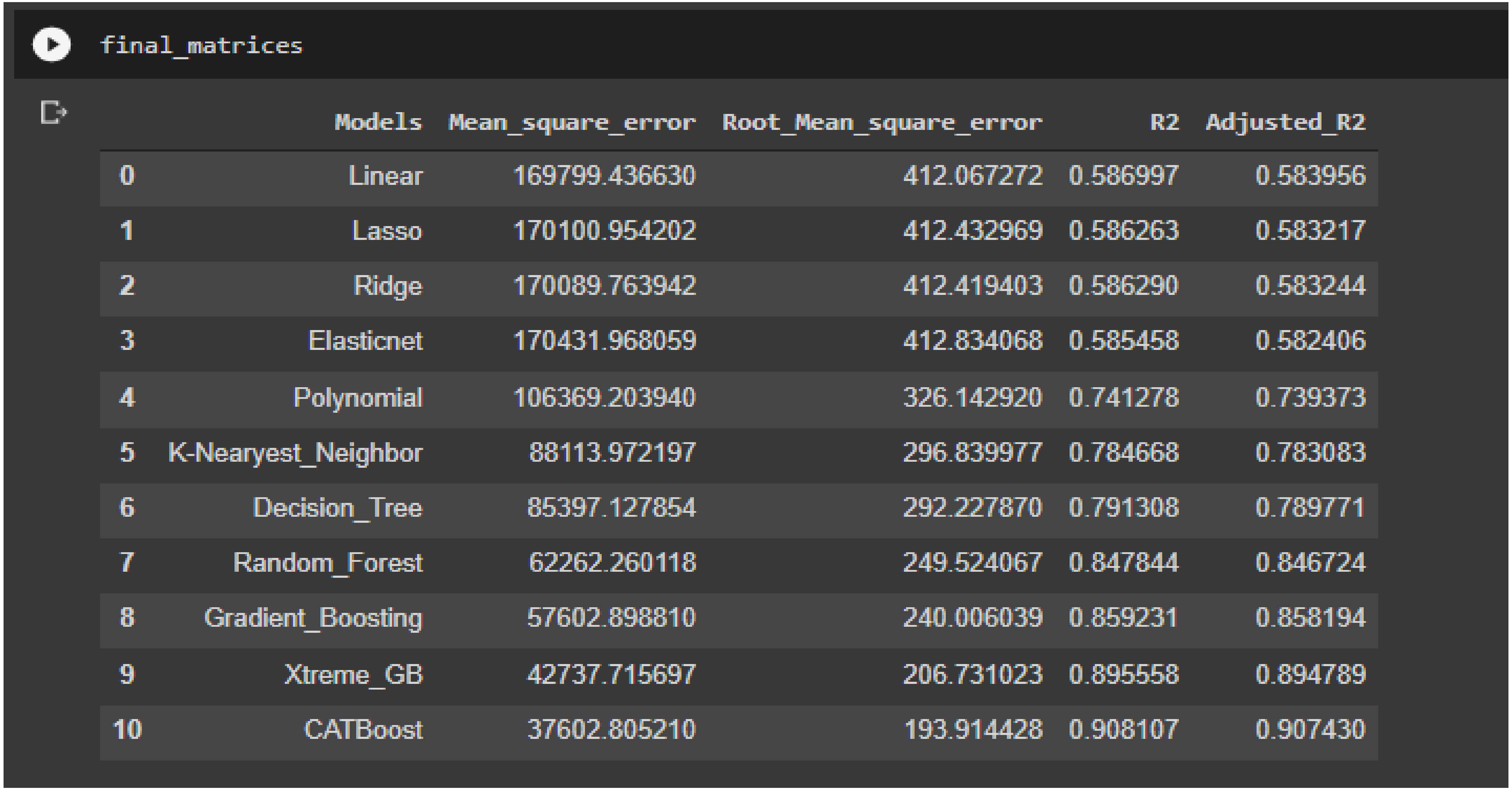
EVALUATION METRIC - RMSLE

- RMSLE = Root Mean Square Log Error
- RMSLE = $\sqrt{\frac{1}{n}\sum_i^n(\log\ (p_i+1)-\log(a_i+1))^2}$
  - n is the number of hours in the test set
  - p_i is the predicted count
  - a_i is the actual count
  - $\log(x)$ is the natural logarithm

Model Validation & Selection

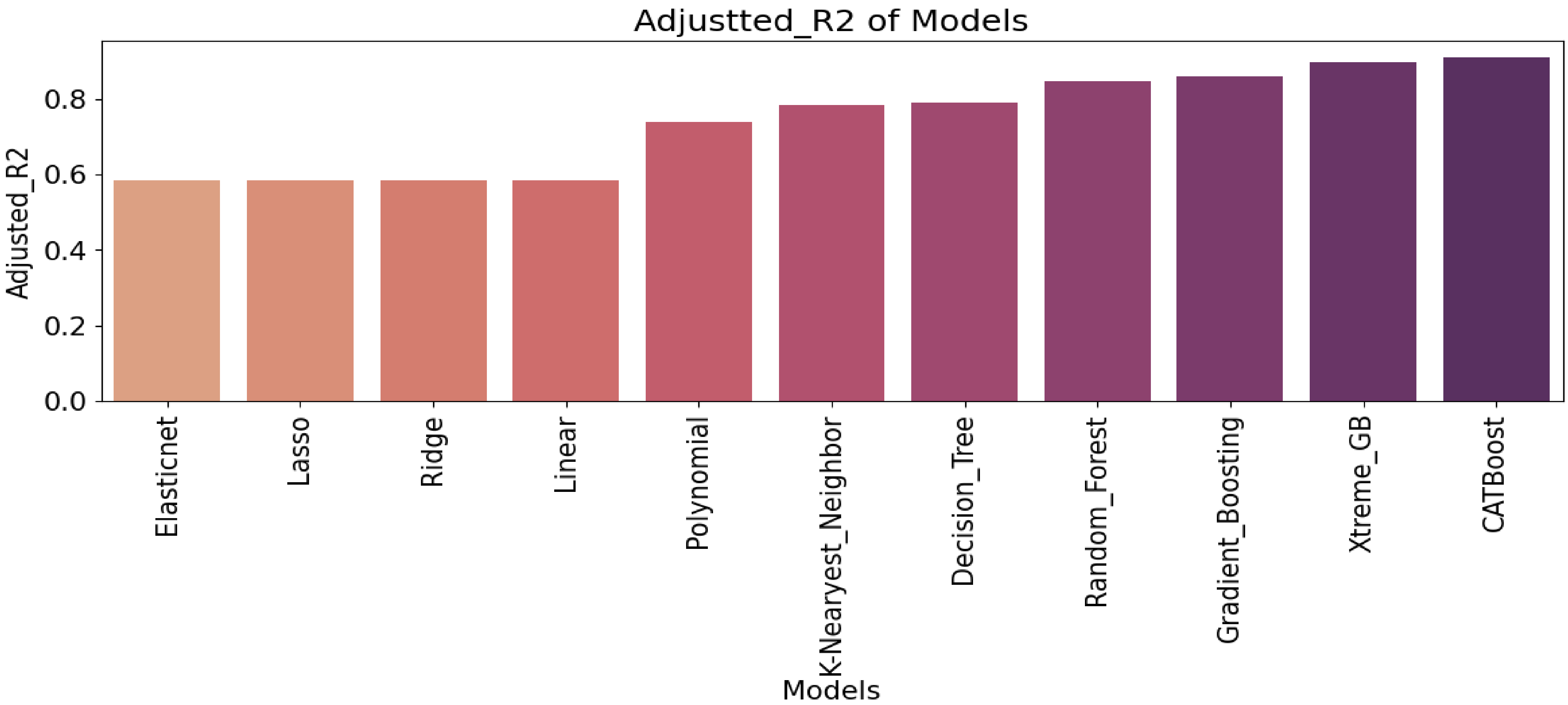
- LINEAR REGRESSION
- LASSO
- RIDGE
- POLYNOMIAL REGRESSION
- DECISION TREE
- K- NEAREST NEIGHBOUR
- ELASTICNET REGRESSION
- RANDOM – FOREST REGRESSION
- GRADIENT BOOSTING
- XTREME_GB
- CATBOOST

Evaluation Matrix of All the models



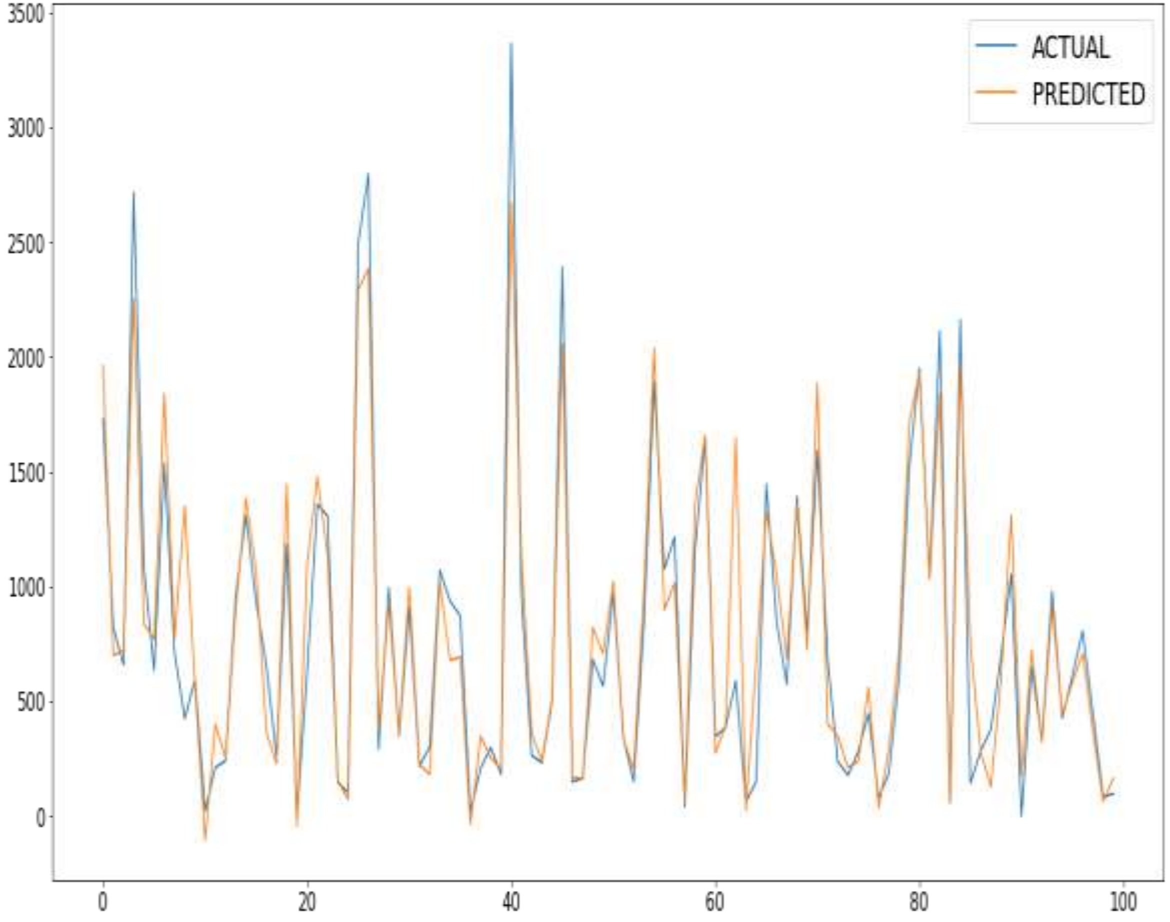
```
final_matrices
```

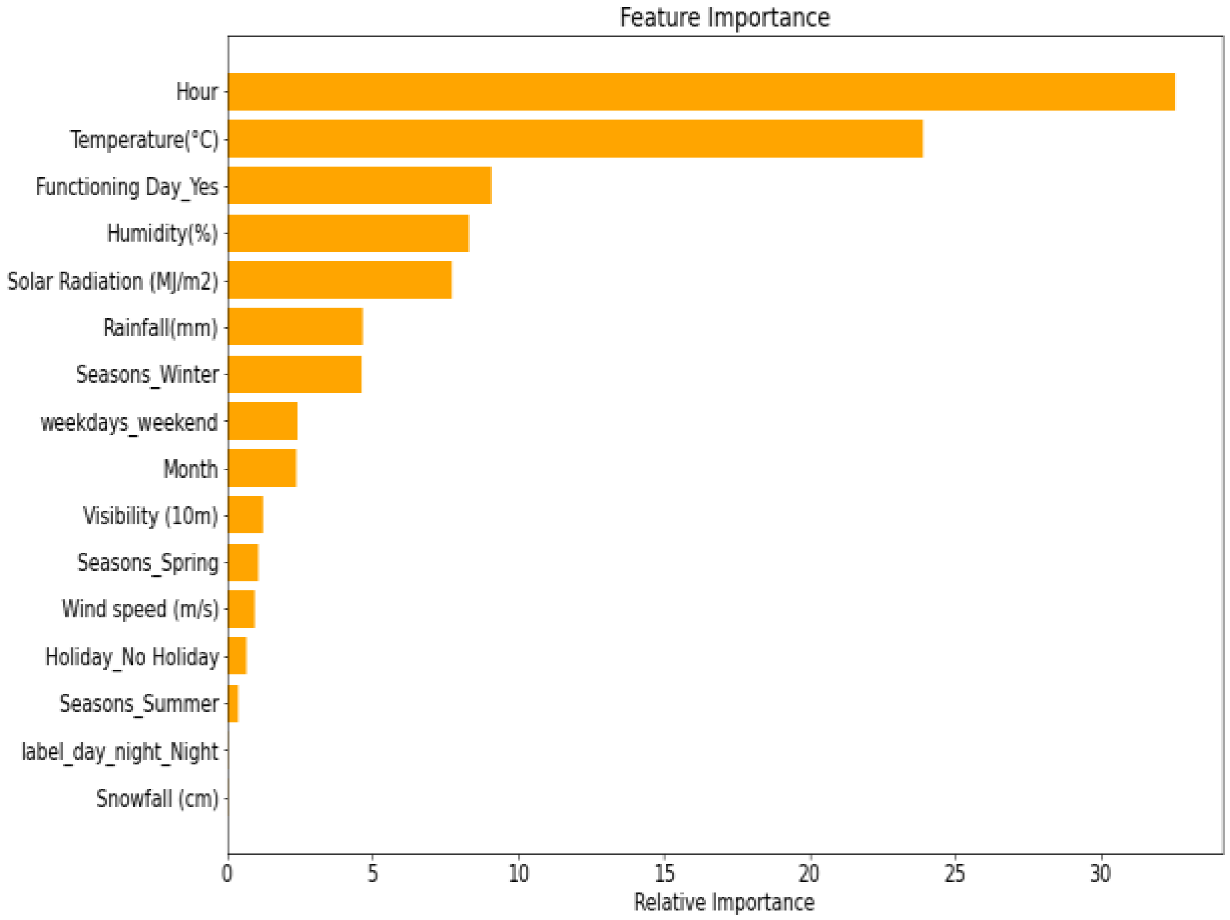
	Models	Mean_square_error	Root_Mean_square_error	R2	Adjusted_R2
0	Linear	169799.436630	412.067272	0.586997	0.583956
1	Lasso	170100.954202	412.432969	0.586263	0.583217
2	Ridge	170089.763942	412.419403	0.586290	0.583244
3	Elasticnet	170431.968059	412.834068	0.585458	0.582406
4	Polynomial	106369.203940	326.142920	0.741278	0.739373
5	K-Nearyest_Neighbor	88113.972197	296.839977	0.784668	0.783083
6	Decision_Tree	85397.127854	292.227870	0.791308	0.789771
7	Random_Forest	62262.260118	249.524067	0.847844	0.846724
8	Gradient_Boosting	57602.898810	240.006039	0.859231	0.858194
9	Xtreme_GB	42737.715697	206.731023	0.895558	0.894789
10	CATBoost	37602.805210	193.914428	0.908107	0.907430



After performing the various models the Catboost model found to be 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) shows a higher value for the CatBoost models !

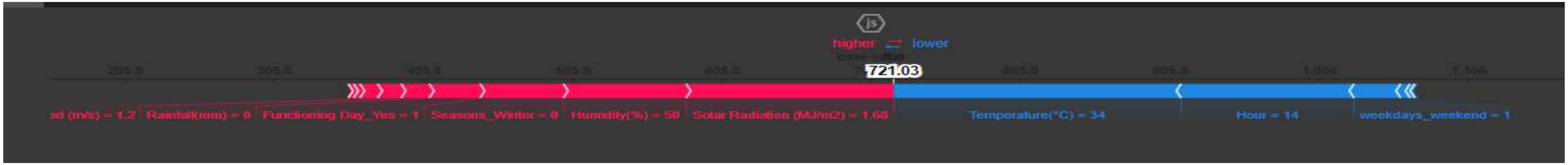
CATBOOST MODEL





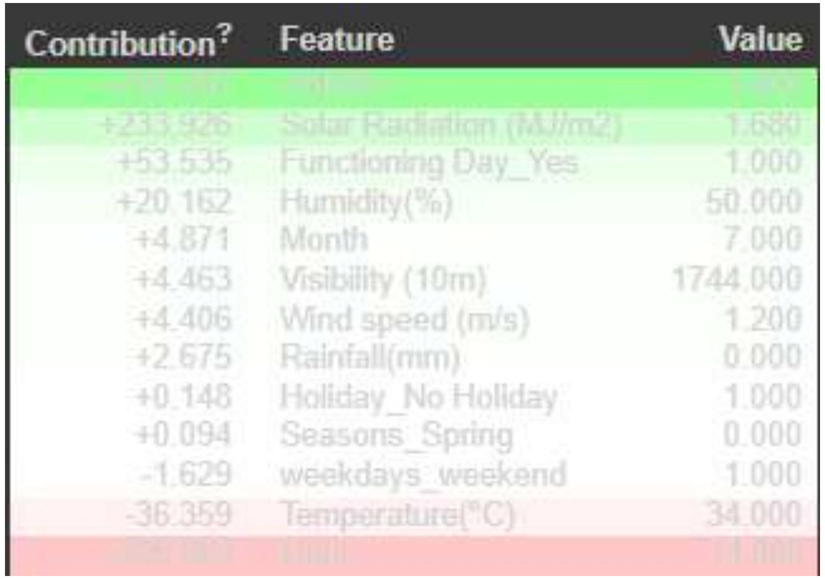
Model Explainability

SHAP:



ELI5:

- The green color shows how much the feature contributes to the prediction of the respective class and the weights are positive for the green color.
- The red color has negative weights that indicate the feature isn't contributing to the prediction of that class.
- It can be observed from the above output, eli5 shows us the contribution of each feature in predicting the output.



Contribution?	Feature	Value
[illegible]	[illegible]	[illegible]
+233.926	Solar Radiation (MJ/m2)	1.680
+53.535	Functioning Day_Yes	1.000
+20.162	Humidity(%)	50.000
+4.871	Month	7.000
+4.463	Visibility (10m)	1744.000
+4.406	Wind speed (m/s)	1.200
+2.675	Rainfall(mm)	0.000
+0.148	Holiday_No Holiday	1.000
+0.094	Seasons_Spring	0.000
-1.629	weekdays_weekend	1.000
-36.359	Temperature(°C)	34.000
[illegible]	Hour	14.000

Challenges

- A huge amount of data needed to be deal while project which is doing the quite an important task and also even small inferences need to be kept in mind.

- As dataset was enough which computation time.

Conclusion

- In holiday or non-working days there is demands in rented bikes.
- There is a surge of high demand in the morning 8AM and in evening 6PM as the people might be going to their work at morning 8AM and returing from their work at the evening 6PM.
- People prefered more rented bikes in the morning than the evening.
- When the rainfall was less, people have booked more bikes except some few cases.
- The Temperature, Hour & Humidity are the most important features that positively drive the total rented bikes count.
- After performing the various models the Catboost model found to be 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) shows a higher value for the Catboost models !

THANK YOU