Capstone Project - 2
Bike Sharing Demand Prediction

Let’s get the rented bike count:

1. Defining Problem Statement
2. Exploratory Data Analysis and Feature Selection
3. Feature Selection
4. Preparing dataset for modelling
5. Applying Model
6. Model Validation and Selection

What is Bike Sharing System?

A **bicycle-sharing system**, **public bicycle scheme**,or **public bike share** (PBS) scheme, is a shared transport service in which bicycles are made available for shared use to individuals on a short term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" and return it at another dock belonging to the same system. Docks are special bike racks that lock the bike, and only release it by computer control. The user enters payment information, and the computer unlocks a bike. The user returns the bike by placing it in the dock, which locks it in place. Other systems are dockless.For many systems, smartphone mapping apps show nearby available bikes and open docks. In July 2020, Google Maps began including bike shares in its route recommendations.People use bike-share for various reasons.

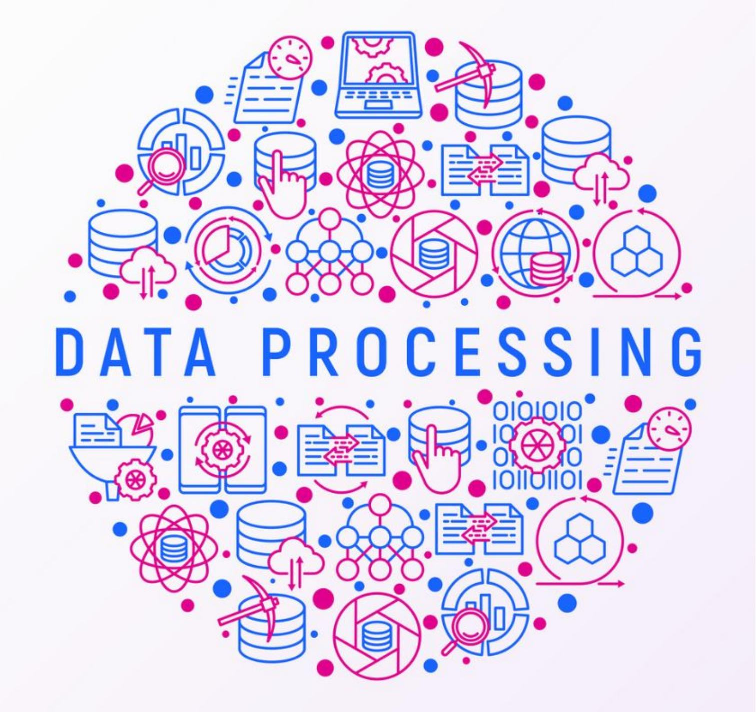
Most large-scale urban bike sharing programmes have numerous bike check-out stations, and operate much like public transit systems, catering to tourists and visitors as well as local residents. Their central concept is to provide free or affordable access to bicycles for short-distance trips in an urban area as an alternative to private vehicles, thereby reducing congestion, noise, and air pollution.

Data Processing

- **Data Preprocessing**: Deletion of NaN values and replacing it with the respective values to process the data machine readable for ML and DL purposes.

- **EDA**: Exploratory Data Analysis is done on the dataset to get inference from the data and to see the visible trends.

- **Create a model**: Experimenting with different models to get the best possible R2 Score as it explains the variance.



Data Preprocessing

As the **First step**, the dataset is checked for null values and those values are handled.

As the **Second step**, the EDA part is carried out for the trends and correlation in the dataset.

1. Firstly, the dataset is checked for the distribution by plotting distplots.
2. Then the dataset is treated for categorical variables which are needed to be replaced with the dummy variables in order to increase the correlation between the variables.
3. Homoscedasticity is checked.
4. The dataset is checked for the correlation and the VIF is determined for each of the features.
5. Skew of the model is checked and transformations are carried out to decrease the skew of the features.
6. Creating different models and selecting the best out of it.

Exploratory Data Analysis

And the second step is the **Exploratory Data Analysis(EDA)** part, where the data is correlated and the trends in the data are discussed. The statistics obtained are as follows:

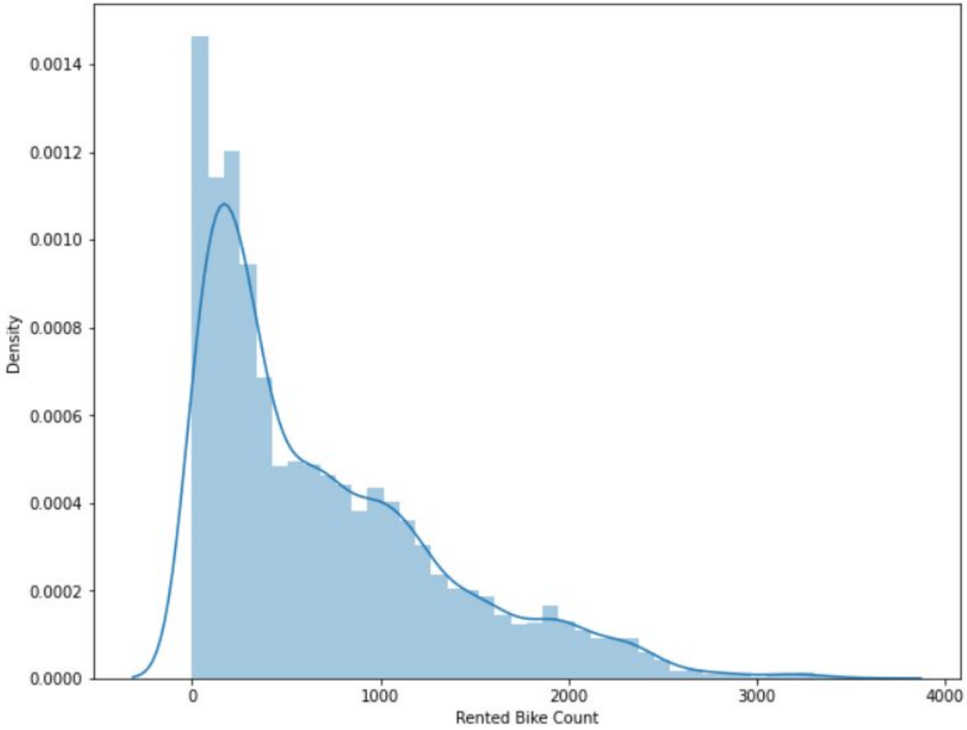
- Checking the distribution
- Replacing Categorical Variables with dummy variables
- Transformations to maintain the skew of the variables.
- Checking Correlation
- Multicollinearity
- Using OLS from statsmodels to get the model summary

Data Summary

Seoul Bike Sharing Dataset:

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	Date	8760 non-null	object
1	Rented Bike Count	8760 non-null	int64
2	Hour	8760 non-null	int64
3	Temperature(°C)	8760 non-null	float64
4	Humidity(%)	8760 non-null	int64
5	Wind speed (m/s)	8760 non-null	float64
6	Visibility (10m)	8760 non-null	int64
7	Dew point temperature(°C)	8760 non-null	float64
8	Solar Radiation (MJ/m2)	8760 non-null	float64
9	Rainfall(mm)	8760 non-null	float64
10	Snowfall (cm)	8760 non-null	float64
11	Seasons	8760 non-null	object
12	Holiday	8760 non-null	object
13	Functioning Day	8760 non-null	object

Exploratory Data Analysis



Distribution Plot of Rented Bike Count

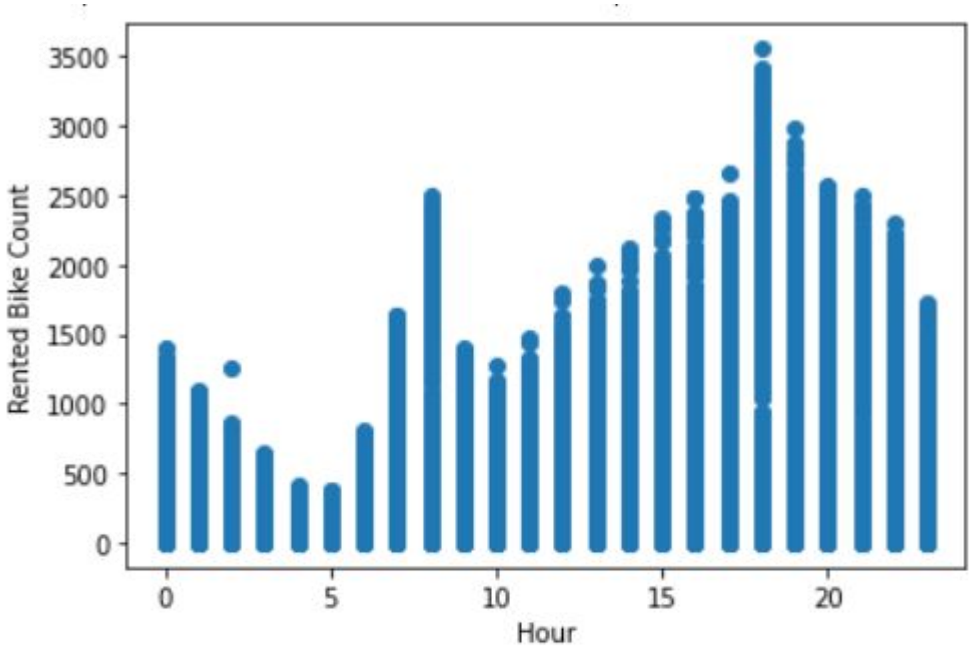
Exploratory Data Analysis

Dummy Variables

Seasons_Autumn	Seasons_Spring	Seasons_Summer	Seasons_Winter	Holiday_Holiday	Holiday_No Holiday	Functioning Day_No	Functioning Day_Yes
0	0	0	1	0	1	0	1
0	0	0	1	0	1	0	1
0	0	0	1	0	1	0	1
0	0	0	1	0	1	0	1
0	0	0	1	0	1	0	1

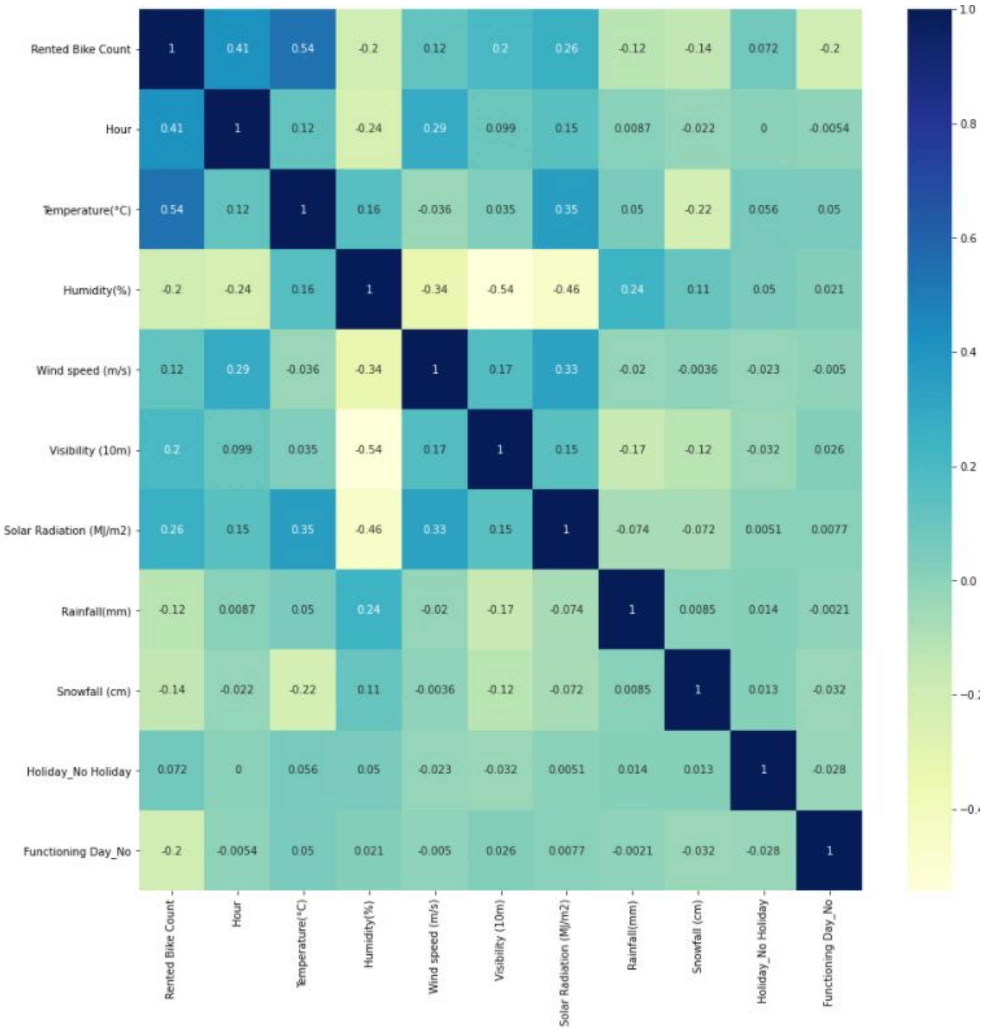
Exploratory Data Analysis

Checking Homoscedasticity



EDA

Checking Correlation



EDA

Checking Multicollinearity

	feature	VIF
0	Hour	3.863762
1	Humidity(%)	4.970480
2	Wind speed (m/s)	4.826903
3	Visibility (10m)	4.943015
4	Solar Radiation (MJ/m2)	1.912428
5	Rainfall(mm)	1.081362
6	Snowfall (cm)	1.128083
7	Seasons_Spring	2.051839
8	Seasons_Summer	2.244963
9	Seasons_Winter	1.988709
10	Functioning Day_No	1.109417

EDA

Checking Skew

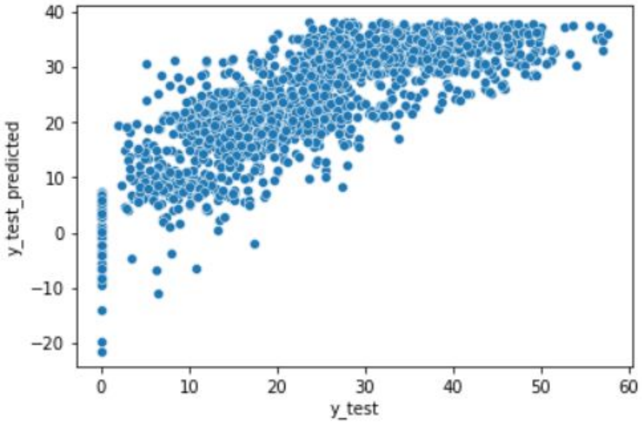
```
Rented Bike Count              0.239782
Hour                           0.000000
Temperature(°C)               -0.198326
Humidity(%)                    0.059579
Wind speed (m/s)              -0.005369
Visibility (10m)              -0.701786
Dew point temperature(°C)     -0.367298
Solar Radiation (MJ/m2)        0.807503
Rainfall(mm)                  -3.700812
Snowfall (cm)                  4.336966
Seasons_Autumn                 1.159123
Seasons_Spring                 1.142294
Seasons_Summer                 1.142294
Seasons_Winter                 1.176139
Holiday_Holiday                4.163603
Holiday_No Holiday            -4.163603
Functioning Day_No             5.170969
Functioning Day_Yes           -5.170969
dtype: float64
```

Linear Regression Model

Naive Model Metrics:

```
MSE is 56.82077363475076 5
RMSE is 7.537955534145234
RMSE is 0.6387867490855245
MAE is 5.951473944047419
MAPE is 231.57717086456583
adjusted_r2 is 0.6431056244410687
```

Naive Model Scatterplot:

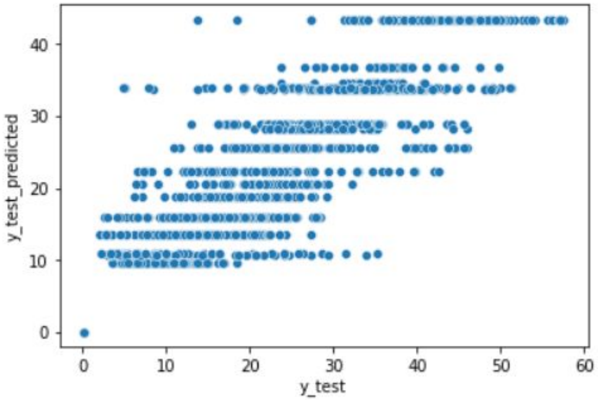


Decision Tree Regressor Model

Decision Tree Model Metrics:

```
MSE is 34.3458546989 2762
```

Decision Tree Scatterplot:

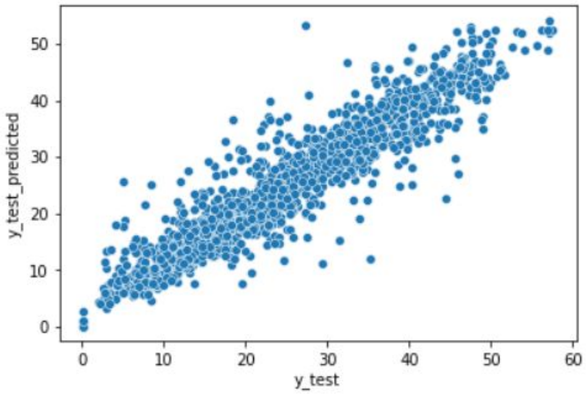


Random Forest Regressor Model

RF Model Metrics:

```
MSE is 16.06661846767959
RMSE is 4.008318658450148
MAE is 2.705508364434621
R2 is 0.8978634904688491
Adjusted R2 is 0.8968621521401123
```

RF Model Scatterplot:

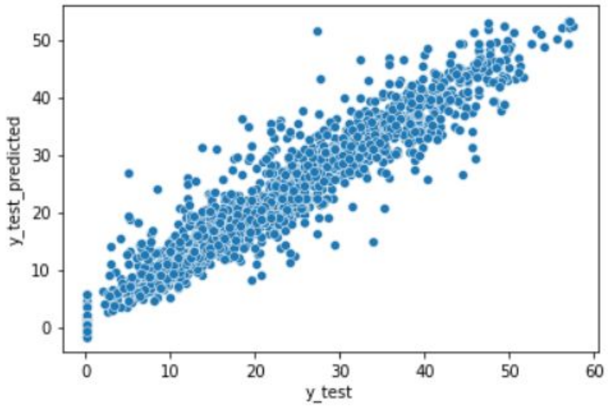


Gradient Boosting Machine Regressor Model

GBM Model Metrics:

```
MSE is 15.107309114738104
RMSE is 3.886812204717138 7
MAE is 2.642082519736794 3
R2 is 0.9039618806850064
Adjusted R2 is 0.9030203304956437
```

GBM Model Scatterplot:

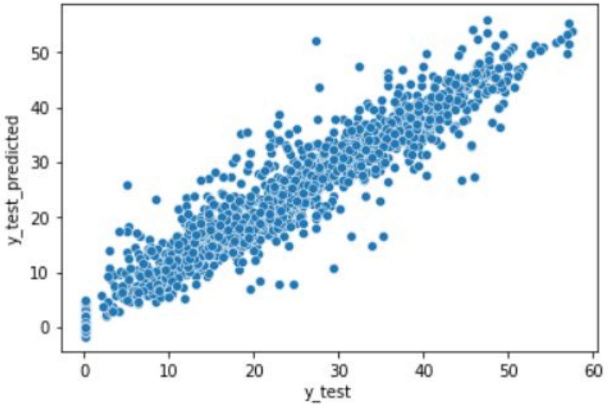


XGBoost Regressor Model

XGBoost Model Metrics:

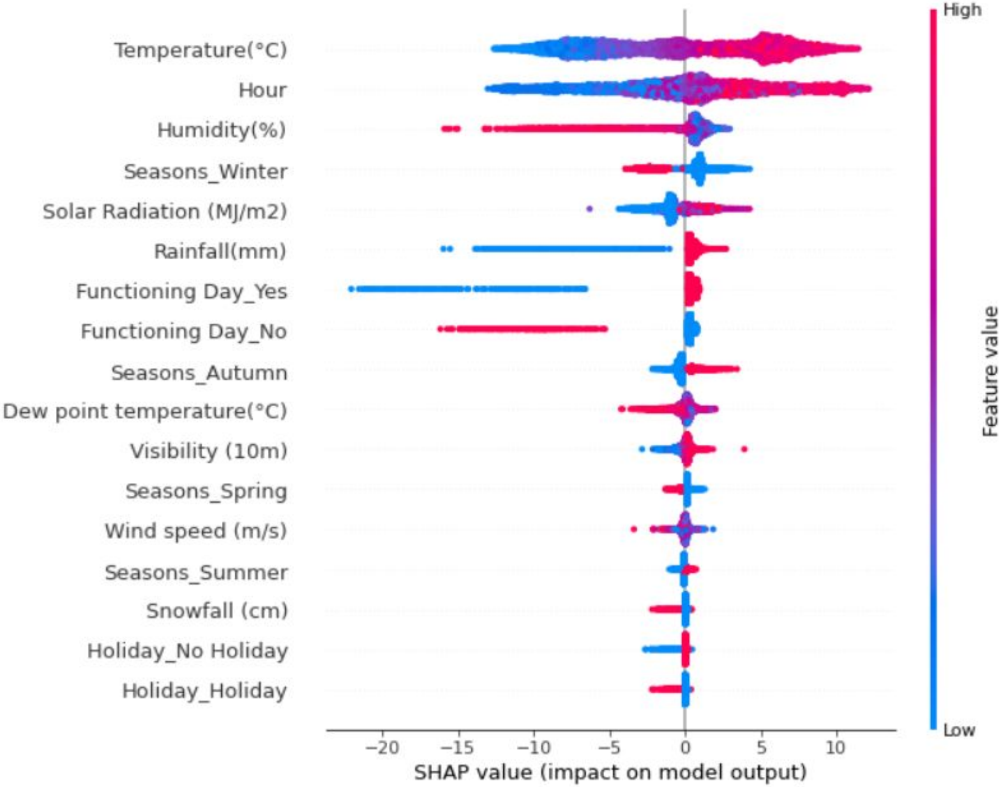
```
MSE is 14.994580560497097
RMSE is 3.872283636369771
MAE is 2.610419284721725
R2 is 0.9046785032324224
Adjusted R2 is 0.9037439787543089
```

XGBoost Model Scatterplot:



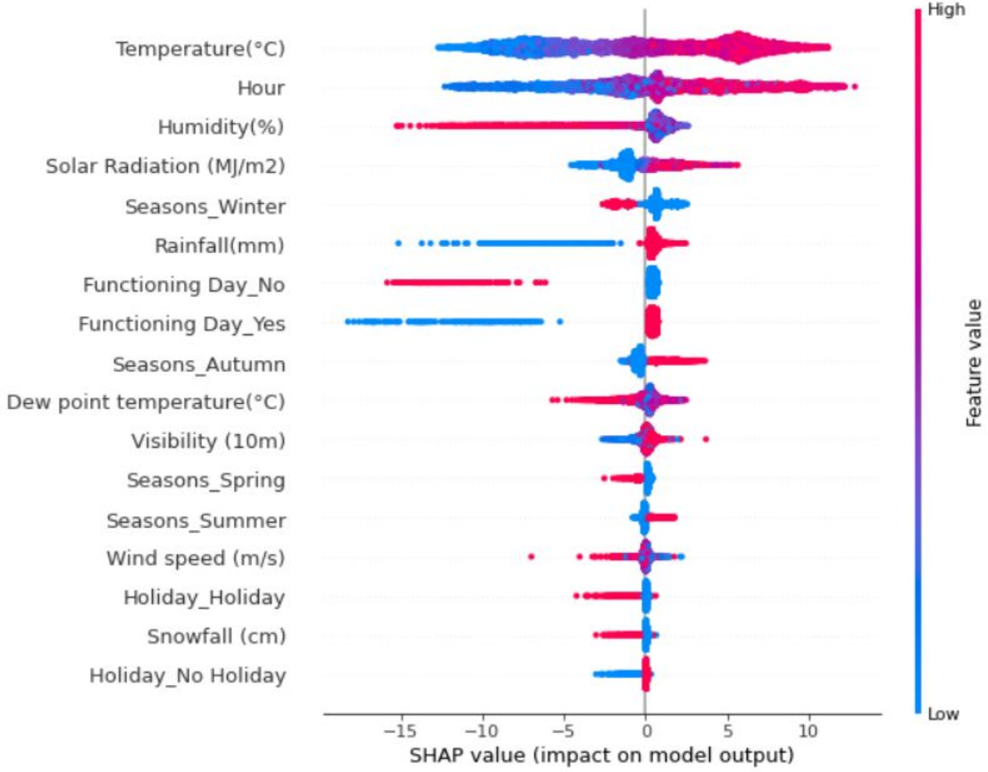
Feature Importances(Using Shap Library)

RF Model:



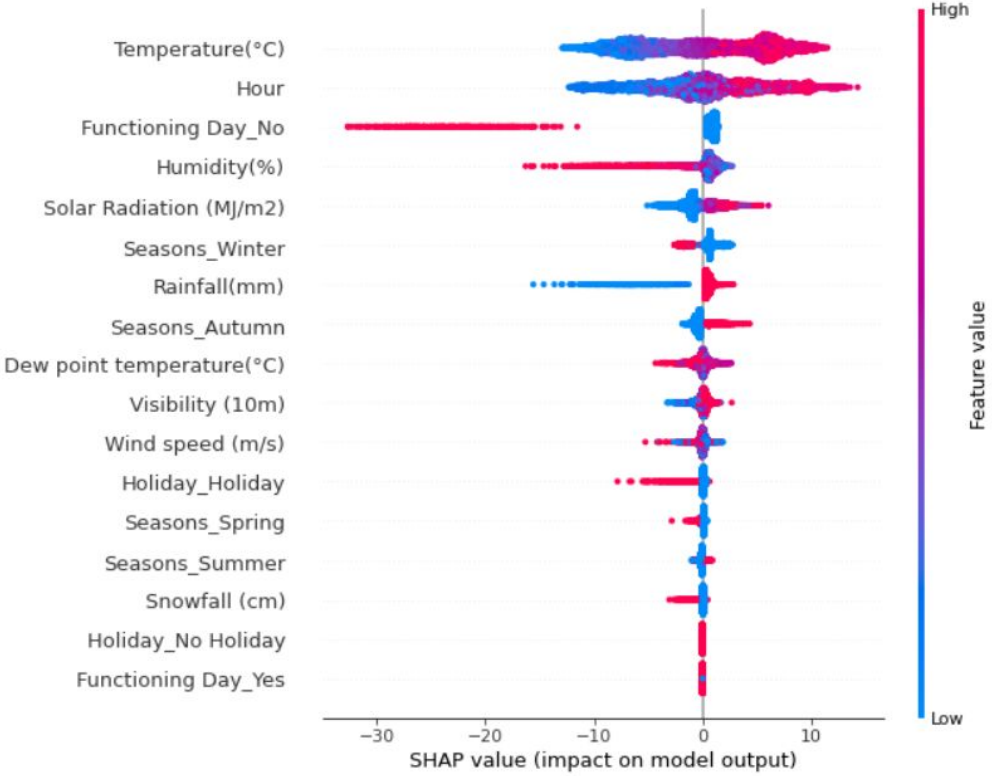
Feature Importances(Using Shap Library)

GBM Model:



Feature Importances(Using Shap Library)

XGBoost Model:

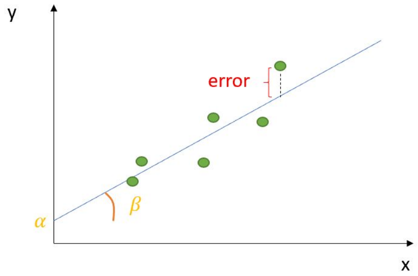


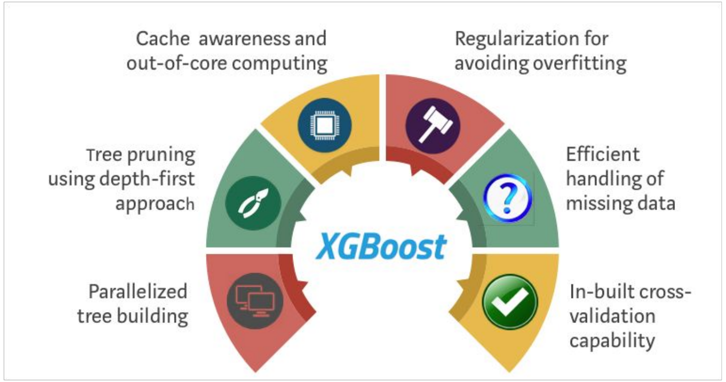
Conclusion

- Upon the tested models XGBoost gives the highest R2 Score which the amount of explained variance by the model. XGBoost gives out the score as 0.9 that means that the variance that can be explained by the model is around 90%.

- Therefore, the best tested model is the XGBoost model and the OLS model which also has a slight edge over the XGBoost model in both R2 Score(i.e., around 0.922 that mean that the variance that can be explained by the model is around 92.2%) and the less model complexity.

- OLS model can be used in this case because it yields more simplicity to the system yet providing very good explained variance(i.e.,R2 Score).





The End