CAPSTONE PROJECT-2

BIKE SHARING DEMAND PREDICTION

(SUPERVISED MACHINE LEARNING REGRESSION)

BY

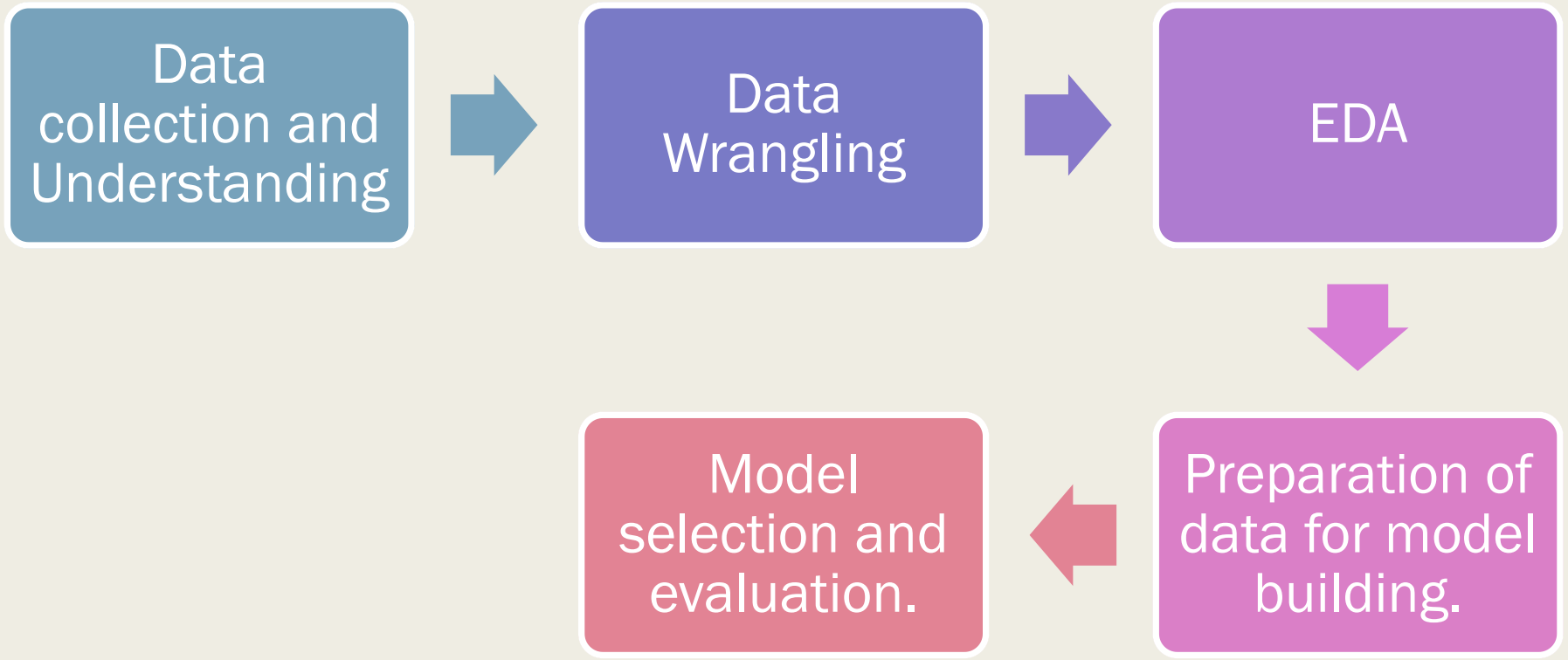
SHAISTA PARVEZ

❖ Problems

- Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. The client is Seoul Bike, which participates in a bike share program in Seoul, South Korea. An accurate prediction of bike count is critical to the success of the Seoul bike share program. 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 final aim of this project is the prediction of bike count required at each hour for the stable supply of rental bikes.

❖ Work Flow :

So we will divide our work flow into following steps.:-



❖ Data Collection and Understanding

We had a Seoul Bike Data for our analysis and model building

The dataset contains weather information (Temperature, Humidity, Wind speed, Visibility, Dew point, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

In this we had total 8760 observations and 14 features including target variable.

Data Description:

Date : year-month-day.

Hour - Hour of the day.

Temperature-Temperature in Celsius.

Humidity - %.

Wind speed - m/s.

Visibility - m.

Dew point temperature - Celsius.

Solar radiation - MJ/m2.

Rainfall - mm.

Snowfall - cm.

Seasons - Winter, Spring, Summer, Autumn.

❖ Data Wrangling and Feature Engineering:

As we know we had 8760 observations and 14 features.

Categorical Features: Seasons, Holiday and Functioning day.

Numerical Columns: Date, Hour, Temperature, Humidity, Wind speed, Visibility, Dew point temperature, Solar radiation, Rainfall, Snowfall, Rented Bike count .

Rename Columns: We renamed columns because they had units mentioned in brackets and was difficult to copy the feature name while working.

```
#Getting all the columns
print("Features of the dataset:")
bike_df.columns
```

```
Features of the dataset:
Index(['Date', '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)', 'Seasons',
       'Holiday', 'Functioning Day'],
      dtype='object')
```

```
#Rename the complex columns name
bike_df=bike_df.rename(columns={'Rented Bike Count':'Rented_Bike_Count',
                                'Temperature(°C)':'Temperature',
                                'Humidity(%)':'Humidity',
                                'Wind speed (m/s)':'Wind_speed',
                                'Visibility (10m)':'Visibility',
                                'Dew point temperature(°C)':'Dew_point_temperature',
                                'Solar Radiation (MJ/m2)':'Solar_Radiation',
                                'Rainfall(mm)':'Rainfall',
                                'Snowfall (cm)':'Snowfall',
                                'Functioning Day':'Functioning_Day'})
```

❖ Data Wrangling and Feature Engineering:

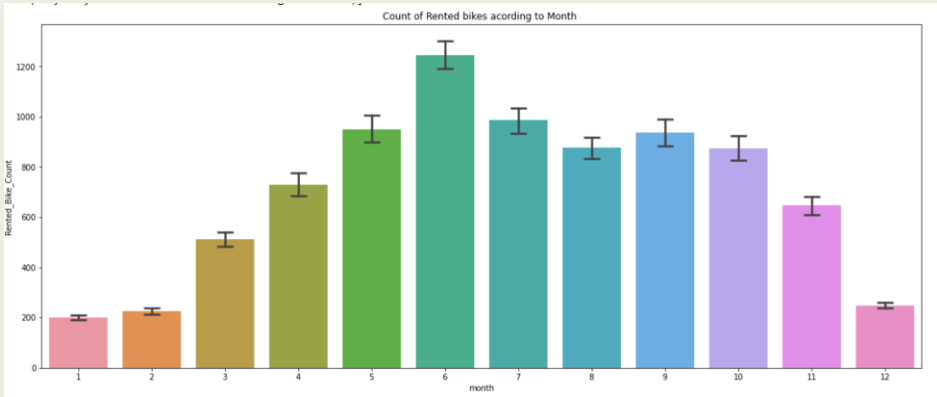
We had zero null values in our dataset.

Zero Duplicate entries found.

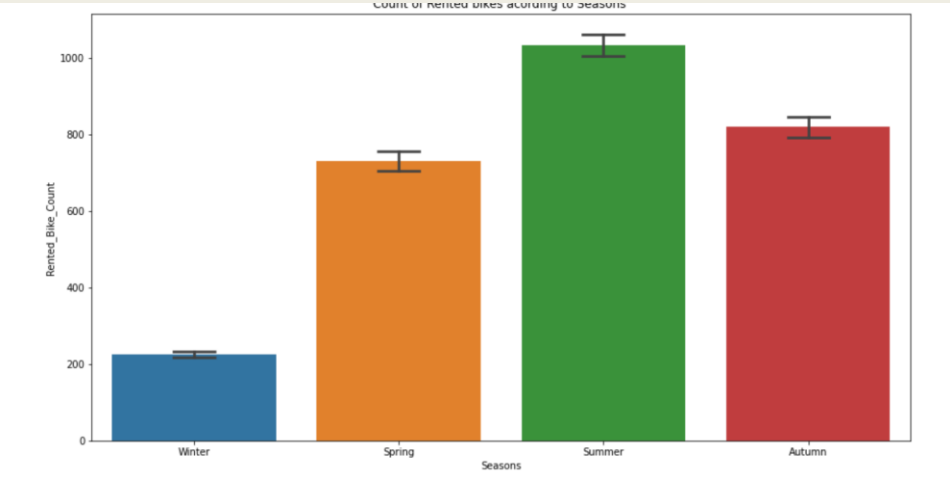
We changed the data type of Date column from ‘object’ to ‘datetime64[ns]’. This was done for feature engineering.

We Created two new columns with the help of Date column ‘Month’ and ‘Day’. Which were further used for EDA. And later we dropped Date column.

❖ EDA (Exploratory Data Analysis)



- From March Bike Rent Count started increasing and it was highest in June.



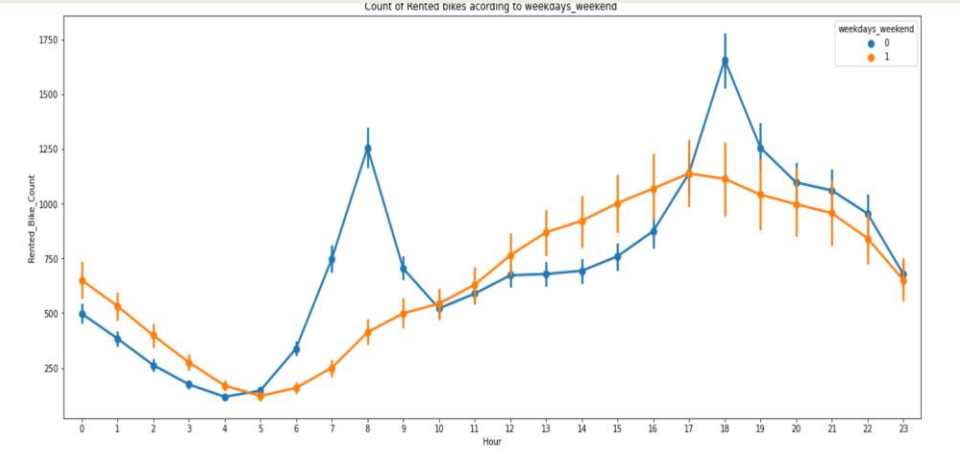
- Relation of rented bike count with categorical features:

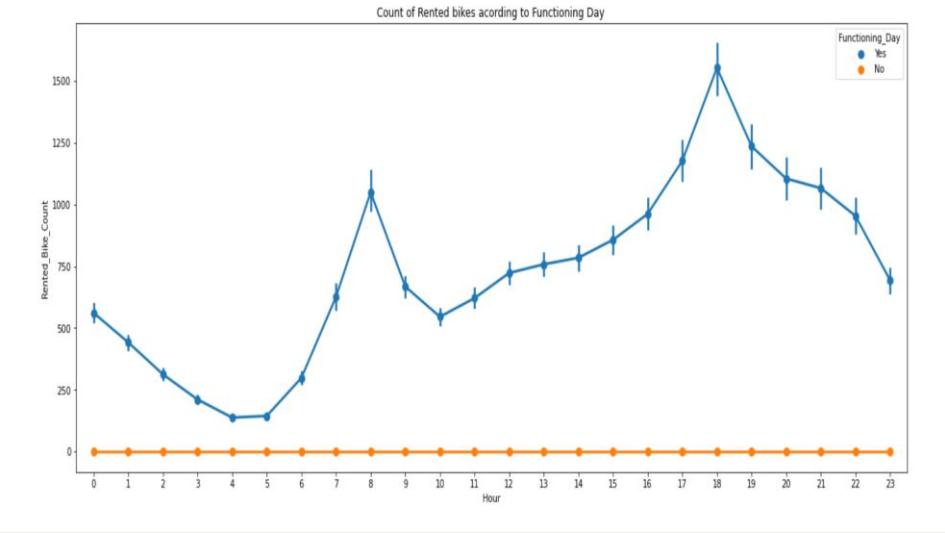
I. Summer season had the highest Bike Rent Count. People are more likely to take rented bikes in summer.

II. Bike rentals in winter is very less compared to other seasons.

❖ EDA (Exploratory Data Analysis)

- Bike Rent Trend according to hour in different scenarios.

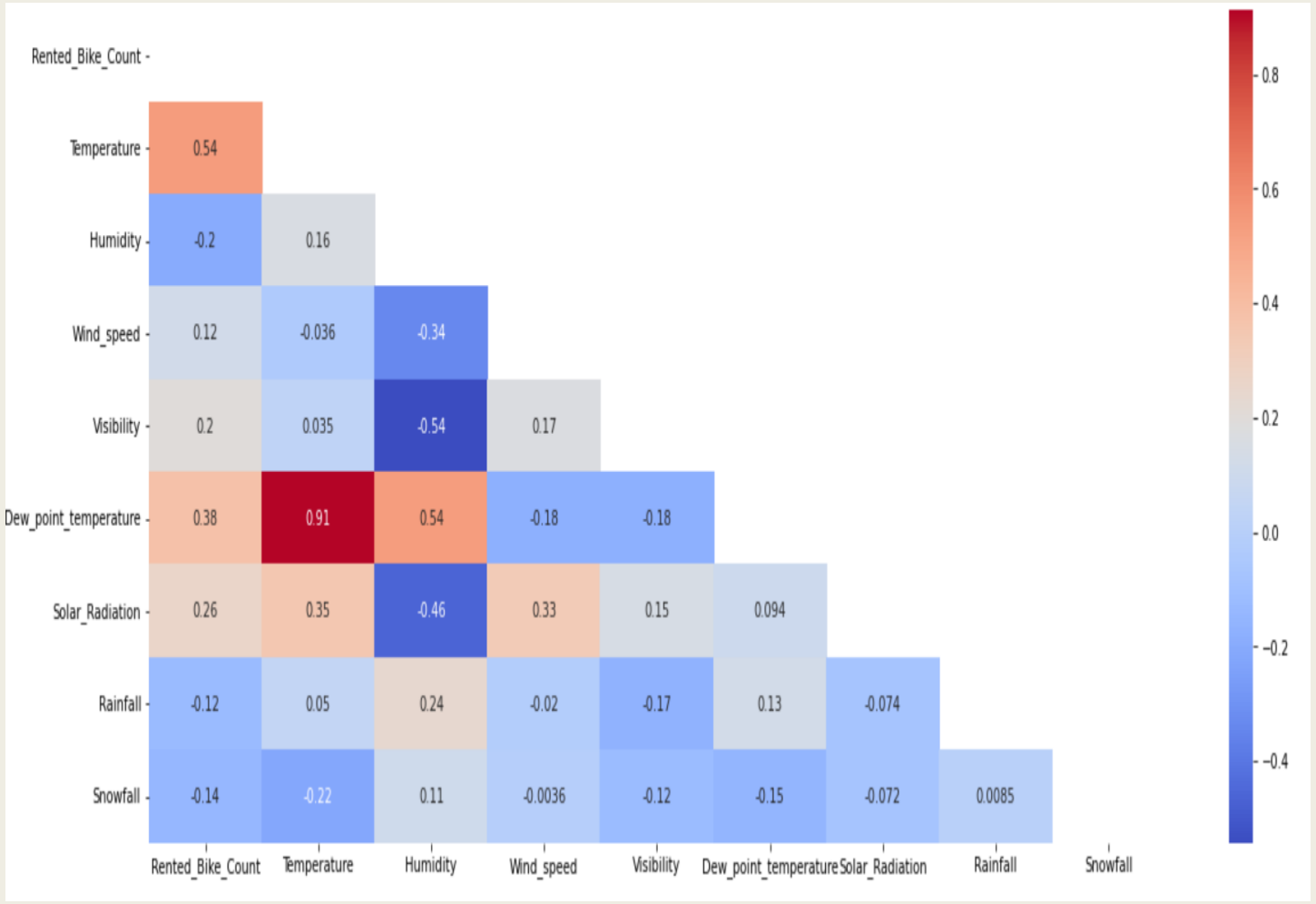




Observations:

1. Here we observed that, Bike rental trend according to hours is almost similar in all scenarios.
2. There is sudden peak between 6/7AM to 10 AM. Office /College going time could be the reason for this sudden peak on NO Holiday. But on Holiday the case is different,very less bike rentals happened.
3. Again there is peak between 4PM to 7 PM. may be its office leaving time for the above people.(NO Holiday).
4. Here the trend for functioning day is same as of No holiday. Only the difference is on No functioning day there were zero bike rentals

❖ Preparation of data for model building:



❖ Model Selection and Evaluation:

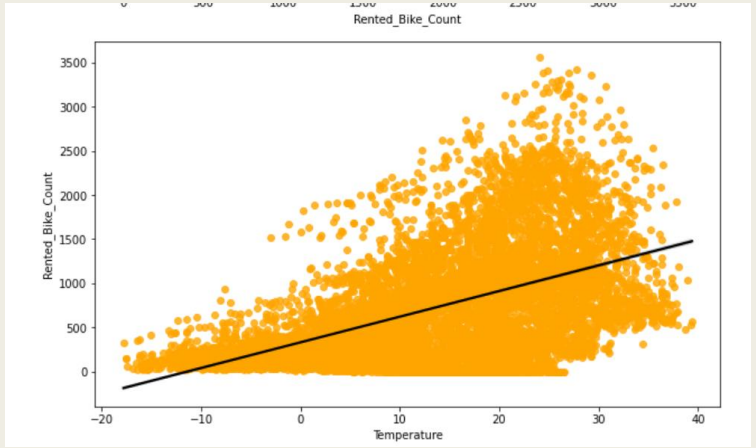
As this is the regression problem we are trying to predict continuous value. For this we used following regression models.

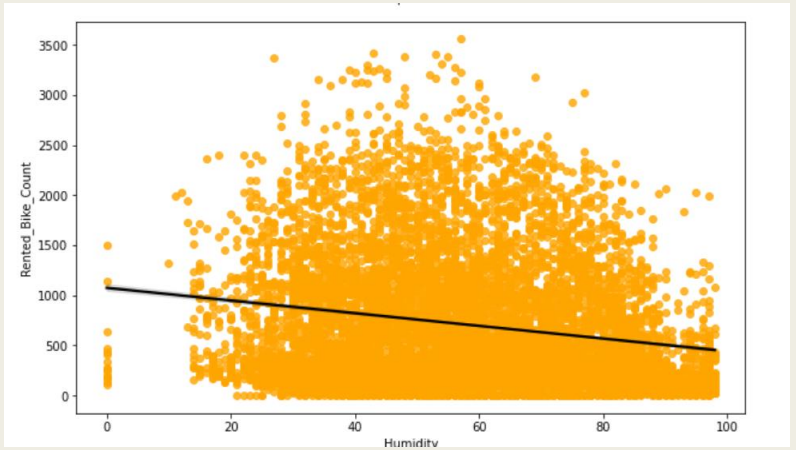
- Linear Regression
- Lasso regression (regularized regression)
- Ridge Regression(regularized regression)
- Decision Tree regression.
- Random forest regression
- Gradient Boosting regression.

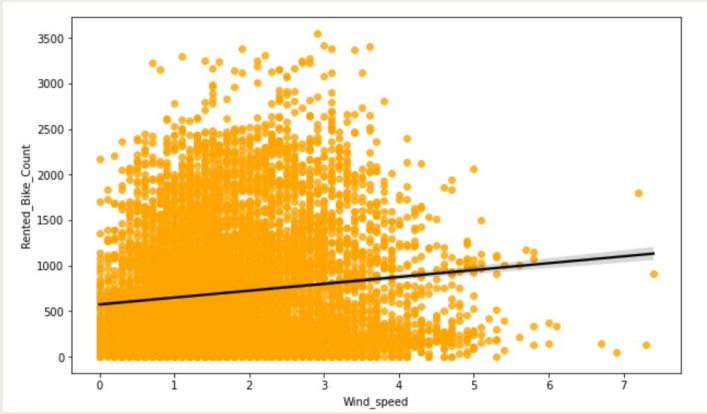
Assumptions of regression line:

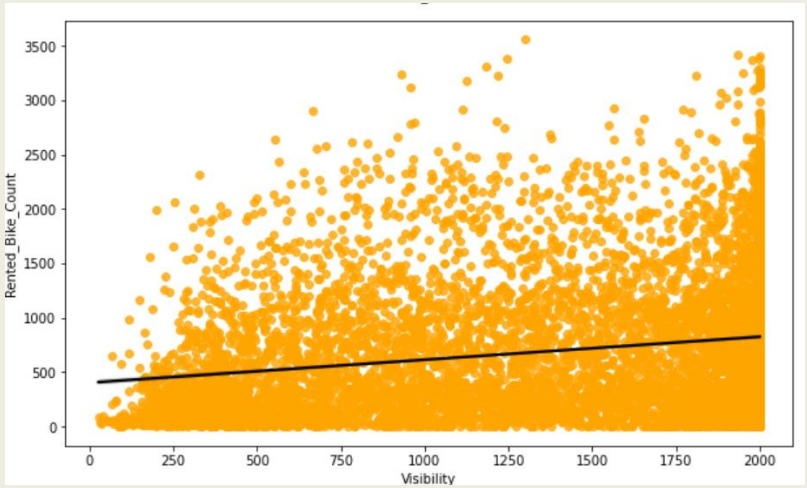
1. The relation between the dependent and independent variables should be almost linear.
2. Mean of residuals should be zero or close to 0 as much as possible. It is done to check whether our line is actually the line of “best fit”.
3. There should be homoscedasticity or equal variance in a regression model. This assumption means that the variance around the regression line is the same for all values of the predictor variable (X).
4. There should not be multicollinearity in regression model. Multicollinearity generally occurs when there are high correlations between two or more independent variables.
5. Before and after applying these models we checked our regression assumptions by distribution of residuals, scatter plot of actual and predicted values, removing multi-colinearity among independent variables

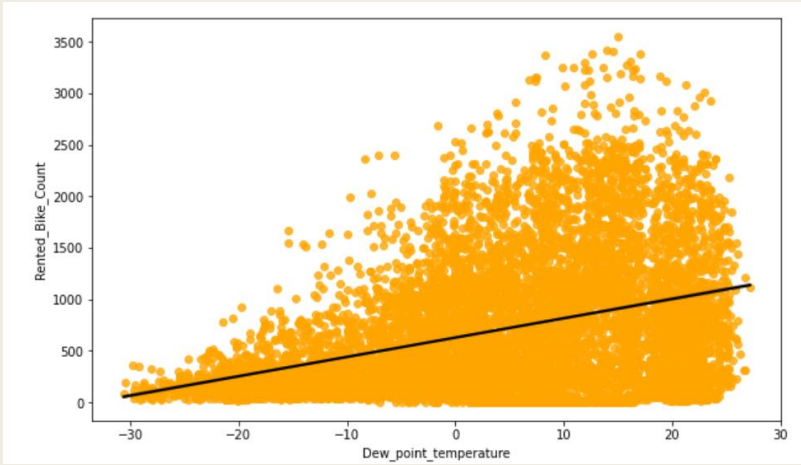
❖ Model Selection and Evaluation

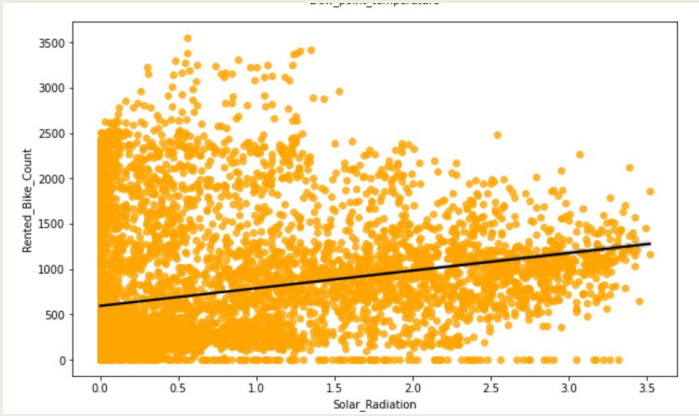




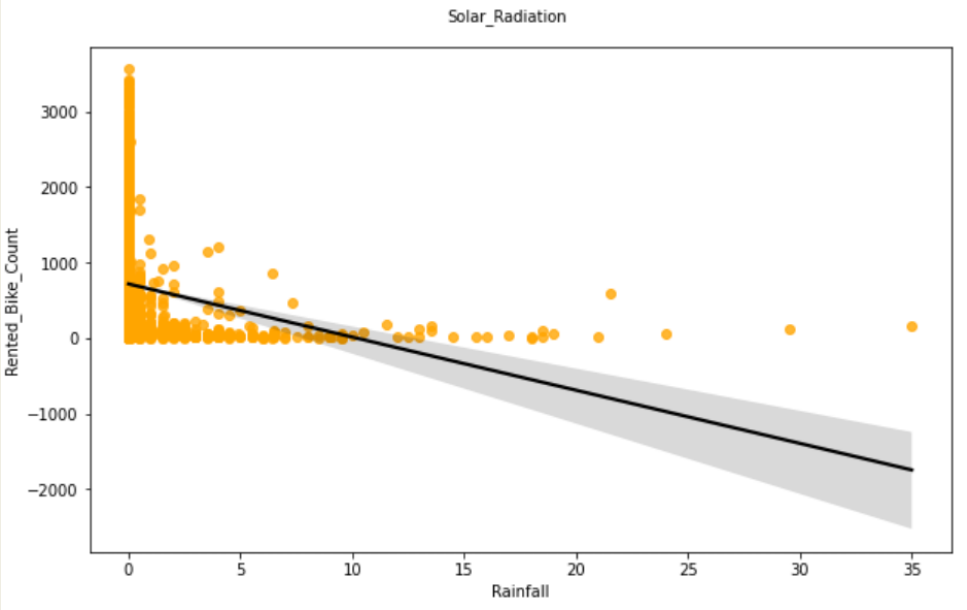


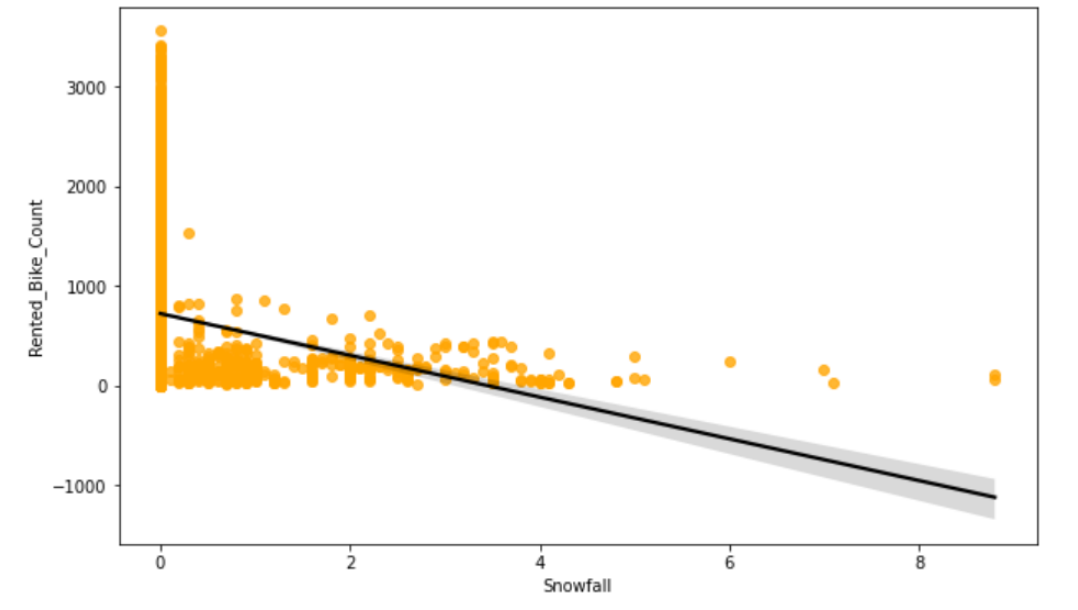






From the above regression plot of all numerical features we see that the columns 'Temperature', 'Wind_speed', 'Visibility', 'Dew_point_temperature', 'Solar_Radiation' are positively relation to the target variable,which means the rented bike count increases with increase of these features





'Rainfall',‘ Snowfall’, 'Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.

❖ Model Selection and Evaluation

Linear regression, Lasso and Ridge Regression:

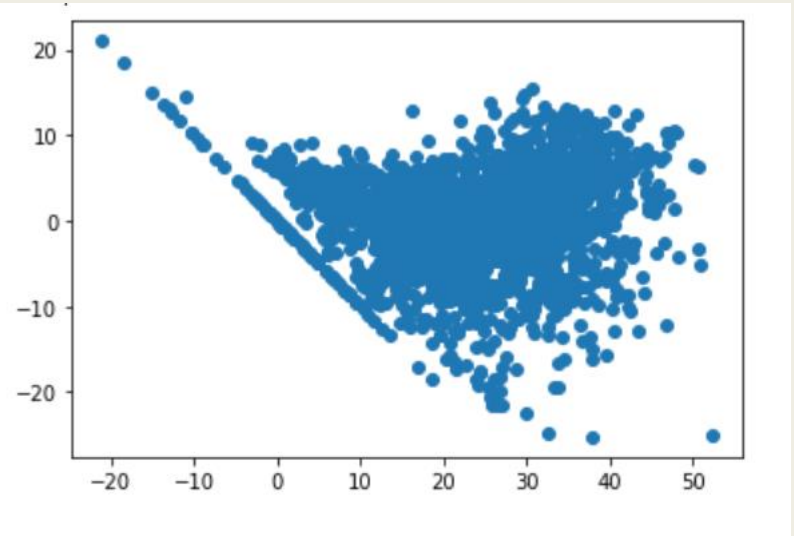
- Linear Regression:

Scores on Train set

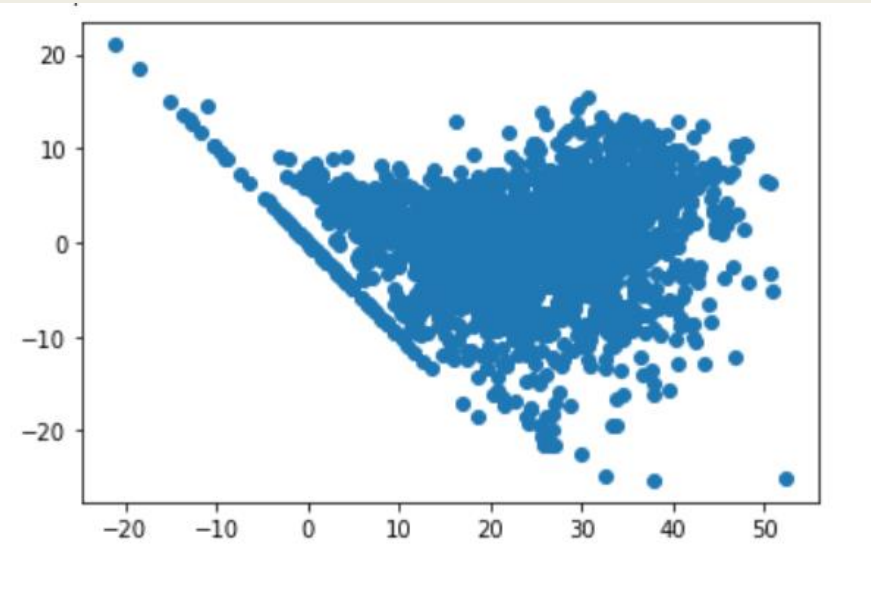
```
MSE : 35.07751288189292
RMSE : 5.9226271942350825
MAE : 4.474024092996788
R2 : 0.7722101548255267
Adjusted R2 : 0.7672119649454145
```

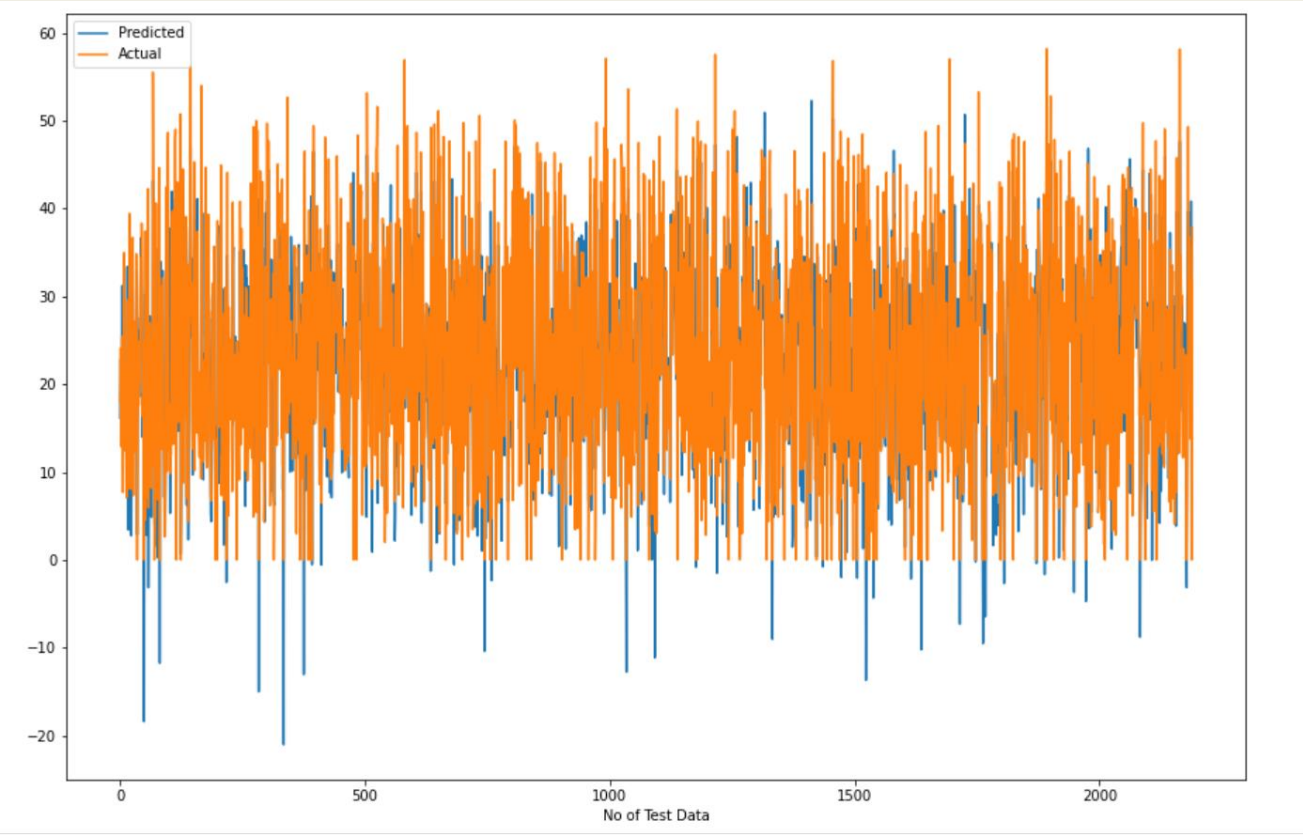
Scores on Test set

```
MSE : 33.27533089591926
RMSE : 5.76847734639907
MAE : 4.410178475318181
R2 : 0.7893518482962683
Adjusted R2 : 0.7847297833429184
```



Mean of residuals should be zero or close to 0 as much as possible. It is done to check whether our line is actually the line of “best fit”.





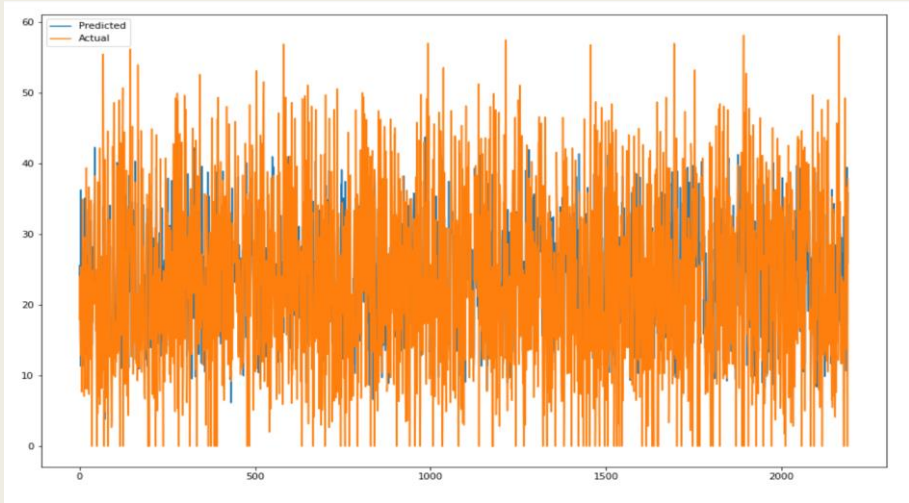
❖ Model Selection and Evaluation :

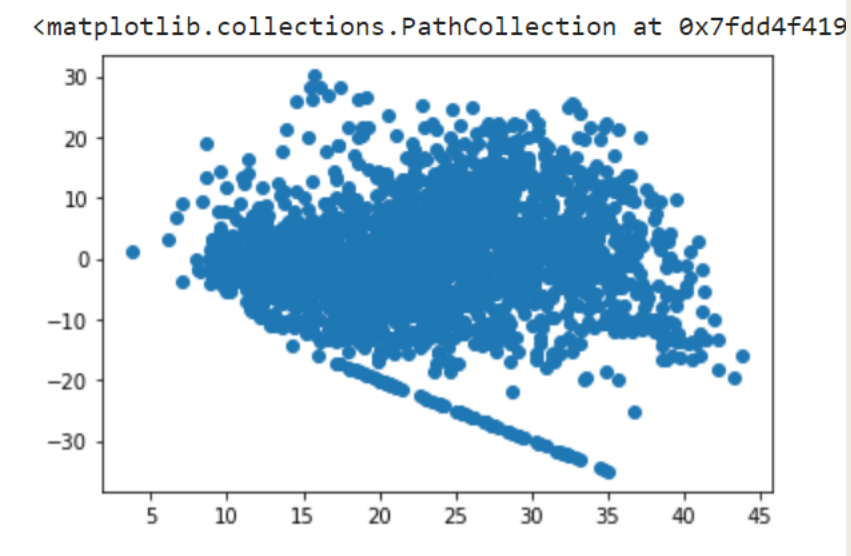
- Lasso (Hyper-parameter tuned- alpha=0.01)
- Scores on Train set

```
MSE : 91.59423336097032
RMSE : 9.570487623991283
MAE : 7.255041571454952
R2 : 0.40519624904934015
Adjusted R2 : 0.3921449996120475
```

Scores on Test set

```
MSE : 96.7750714044618
RMSE : 9.837432155011886
MAE : 7.455895061963607
R2 : 0.3873692800799008
Adjusted R2 : 0.37392686932535146
```





❖ Model Selection and Evaluation

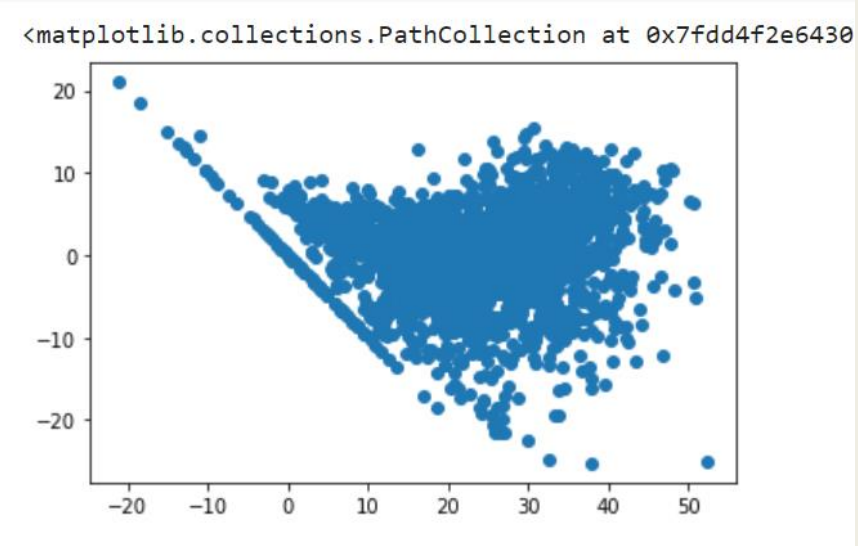
Ridge (Hyper-parameter tuned- alpha=0.1)

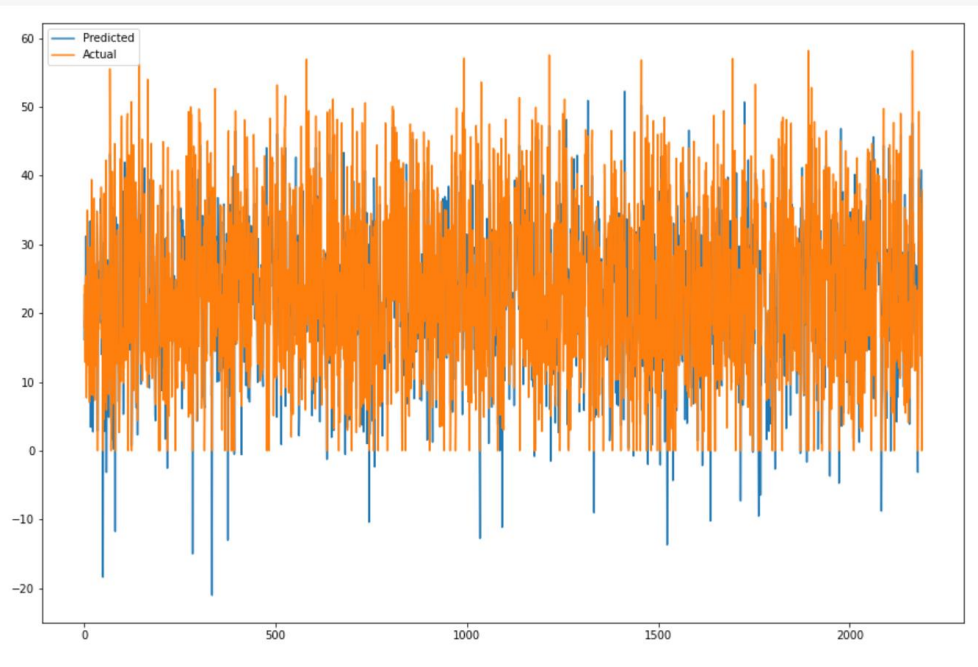
Scores on Train set

```
MSE : 35.07752456136463
RMSE : 5.922628180239296
MAE : 4.474125776125378
R2 : 0.7722100789802107
Adjusted R2 : 0.7672118874358922
```

Scores on Test set

```
MSE : 33.27678426818438
RMSE : 5.768603320404722
MAE : 4.410414932539515
R2 : 0.7893426477812578
Adjusted R2 : 0.7847203809491939
```





❖ Model Selection and Evaluation :

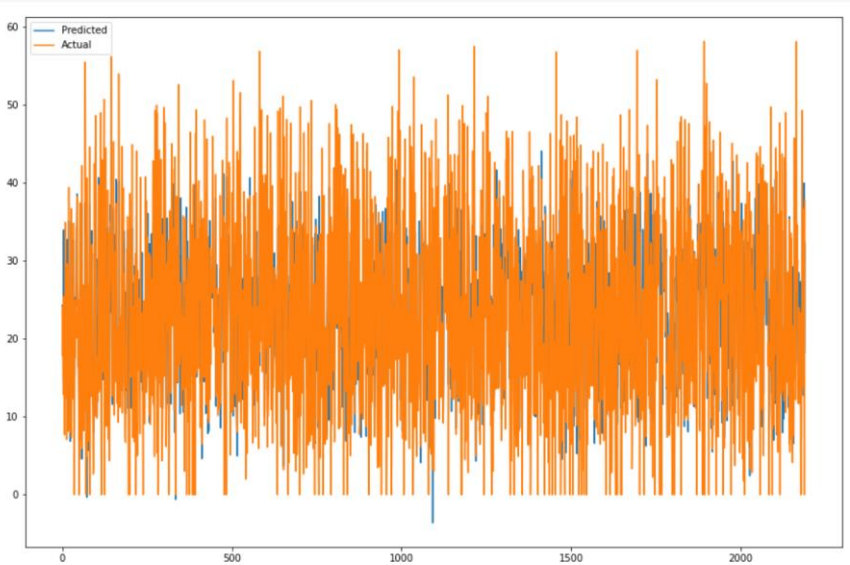
Elastic Net (Hyper-parameter tuned- alpha=0.001,l1_ratio=0.5)

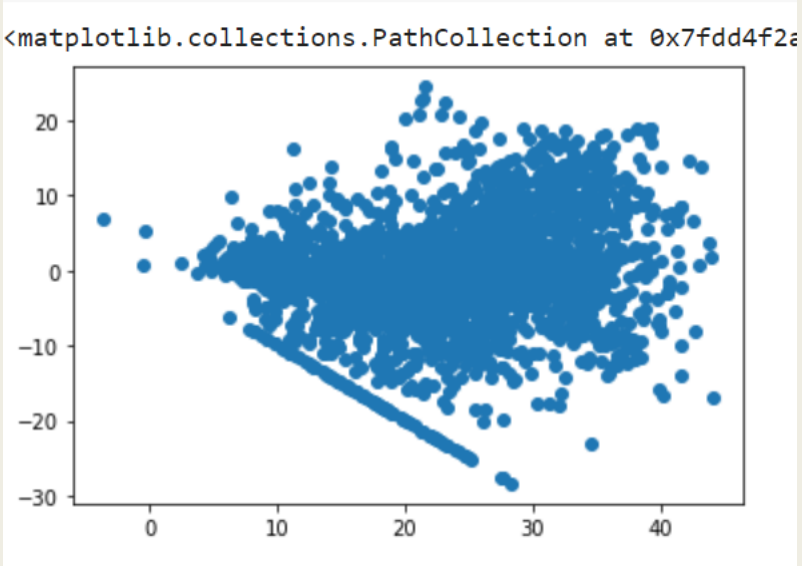
Scores on Train set

```
MSE : 57.5742035398887
RMSE : 7.587766703048315
MAE : 5.792276538970546
R2 : 0.6261189054494012
Adjusted R2 : 0.6179151652795234
```

Scores on Test set

```
MSE : 59.45120536350042
RMSE : 7.710460775044538
MAE : 5.873612334800099
R2 : 0.6236465216363589
Adjusted R2 : 0.6153885321484546
```





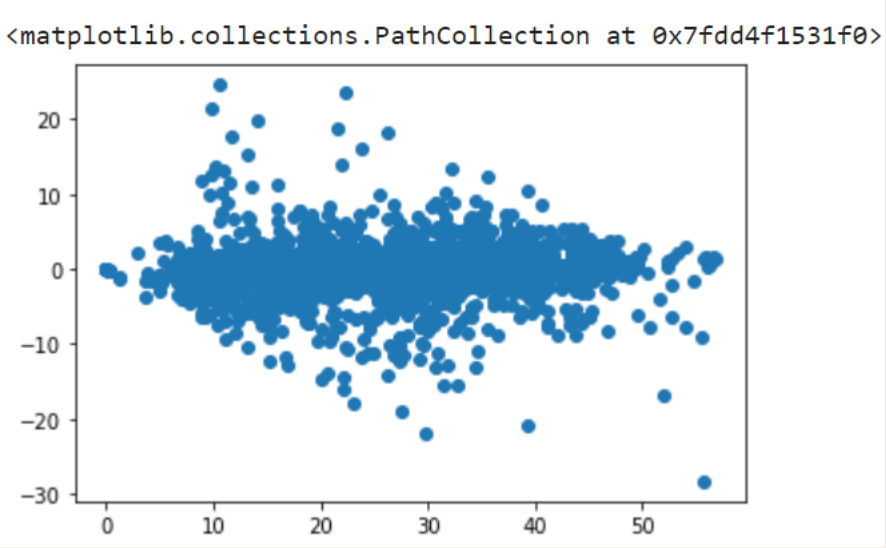
❖ Model Selection and Evaluation

- Decision Tree regression(Hyper-parameter tuned- max_depth=9,max_features='auto')
- Scores on Train set

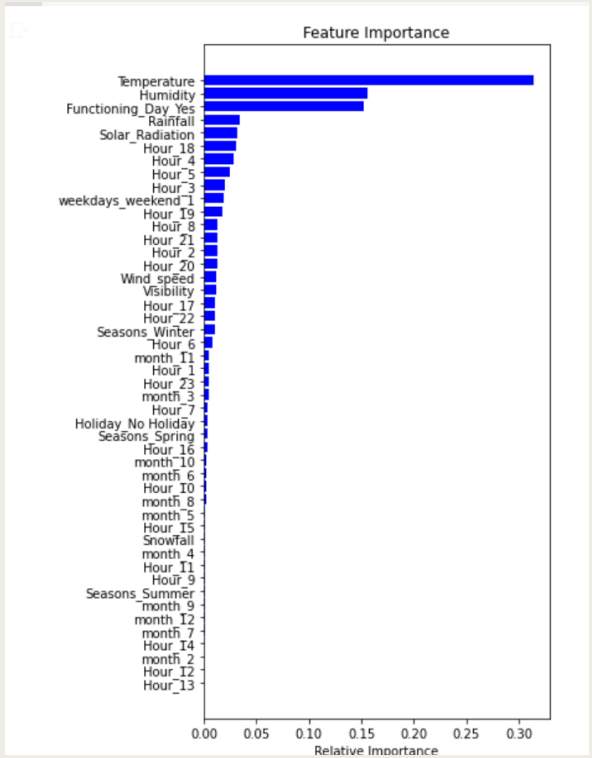
```
Model Score: 0.989723996094852
MSE : 1.5824088166929668
RMSE : 1.257938319908002
MAE : 0.8020998531252489
R2 : 0.989723996094852
Adjusted R2 : 0.9894985188849817
```

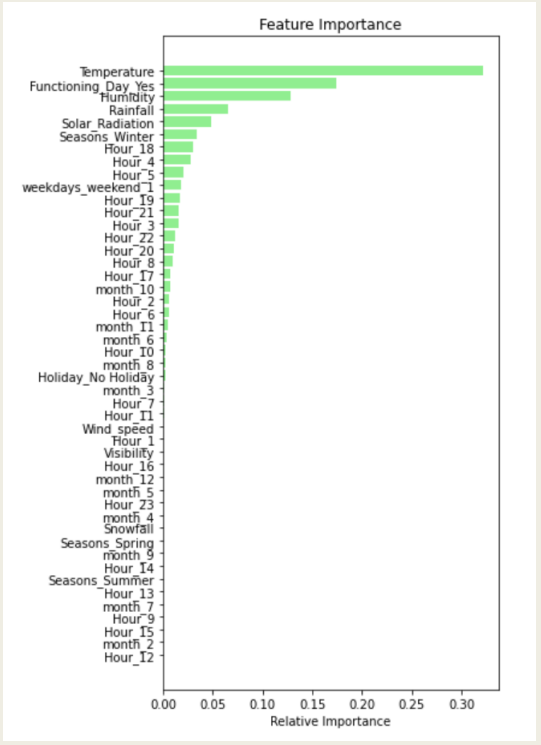
Scores on Test set

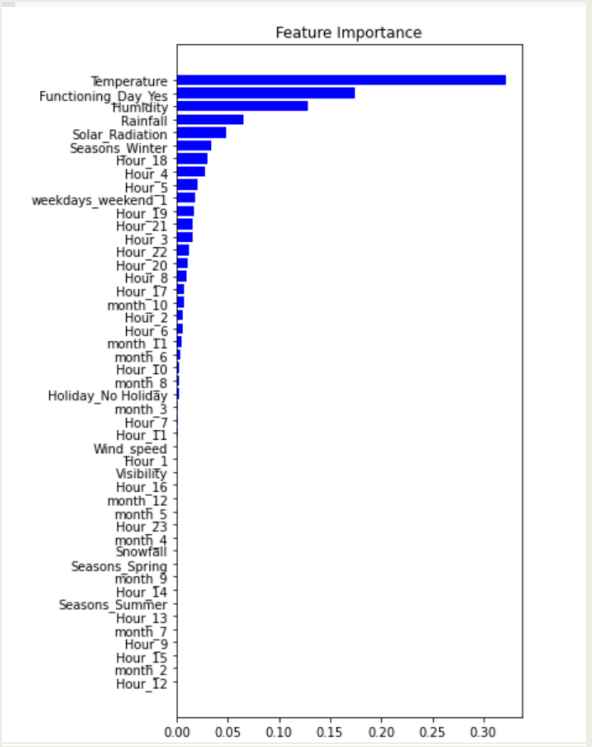
```
MSE : 12.814377206449244
RMSE : 3.5797174757862167
MAE : 2.2228597529191703
R2 : 0.9188790974846776
Adjusted R2 : 0.9170991337039959
```



❖ Feature importance's:







From all 3 models we can say that temperature, hour, functioning day are the top three important features.

❖ Conclusion :

		Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Training set	0	Linear regression	4.410	33.275	5.768	0.789	0.78
	1	Lasso regression	7.255	91.594	9.570	0.405	0.39
	2	Ridge regression	4.474	35.078	5.923	0.772	0.77
	3	Elastic net regression	5.792	57.574	7.588	0.626	0.62
	4	Dicision tree regression	5.146	54.662	7.393	0.645	0.64
	5	Random forest regression	0.802	1.582	1.258	0.990	0.99
	6	Gradient boosting regression	3.269	18.648	4.318	0.879	0.88
	7	Gradient Boosting gridsearchcv	1.849	7.455	2.730	0.952	0.95
Test set	0	Linear regression	4.410	33.275	5.768	0.789	0.78
	1	Lasso regression	7.456	96.775	9.837	0.387	0.37
	2	Ridge regression	4.410	33.277	5.769	0.789	0.78
	3	Elastic net regression Test	5.874	59.451	7.710	0.624	0.62
	4	Dicision tree regression	5.728	69.764	8.352	0.558	0.55
	5	Random forest regression	2.223	12.814	3.580	0.919	0.92
	6	Gradient boosting regression	3.493	21.289	4.614	0.865	0.86
	7	Gradient Boosting gridsearchcv	2.401	12.393	3.520	0.922	0.92

As we have calculated MAE,MSE,RMSE and R2 score for each model. Based on r2 score will decide our model performance.
Our assumption: if the difference of R2 score between Train data and Test is more than 5 % we will consider it as over fitting.
Linear, Lasso, Ridge and Elastic Net: Linear, Lasso, Ridge and Elastic regression models have almost similar R2 scores(61%) on both training and test data.(Even after using GridserachCV we have got similar results as of base models). Decision Tree **Regression:** On Decision tree regressor model, without hyper -parameter tuning, we got r2 score as 100% on training data and on test data it was very less. Thus our model memorized the data. So it was a over fitted model. After hyper -parameter tuning we got r2 score as 88% on training data and 83% on test data which is quite good for us.
Random Forest: On Random Forest regressor model, without hyper -parameter tuning we got r2 score as 98% on training data and 90% on test data. Thus our model memorized the data. So it was a over fitted model, as per our assumption After hyper -parameter tuning we got r2 score as 90% on training data and 87% on test data which is very good for us

❖ Conclusion:

Thus Gradient Boosting Regression(GridSearchCV) and Random forest(GridSearchCv) gives good r2 scores. We can deploy this models.


THANK YOU