Bike Sharing Demand Prediction

Vikas Kumar Manjhi

With

Team Member

Agenda

PROBLEM DESCRIPTION
3

BUSINESS UNDERSTANDING
4

EXPLORATORY DATA ANALYSIS
5

MODEL BUILDING
16

CONCLUSION
31

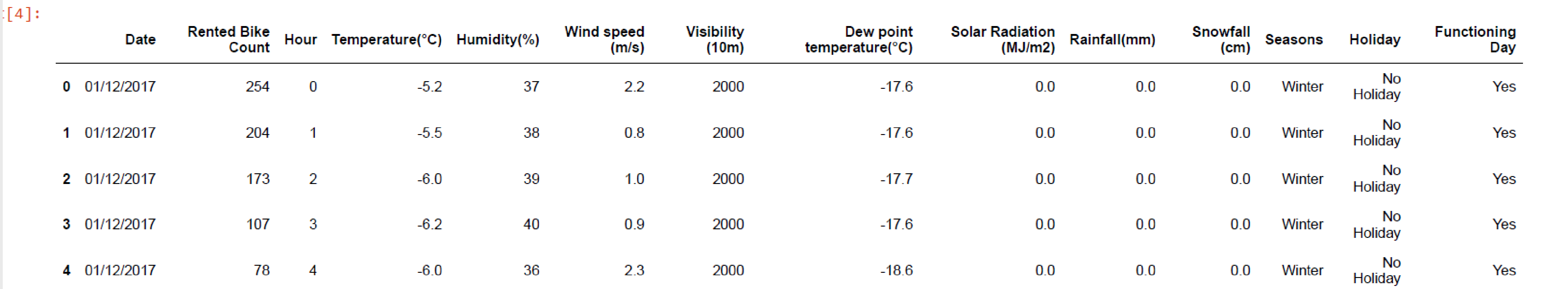
PROBLEM DESCRIPTION

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.

BUSINESS UNDERSTANDING

- Bike rentals have become a popular service in recent years and it seems people are using it more often. With relatively cheaper rates and ease of pick up and drop at own convenience is what making this business thrive.
- Mostly used by people having no personal vehicles and also to avoid congested public transport that's why they prefer rental bike.
- Therefore the business to strive and profit more, it has to be always ready and supply no. of bikes at different locations, to fulfill the demand.
- Our project goal is a pre planned set of bike count values that can be a handy solution to meet all demands.

DATA SUMMARY



	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
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

- This dataset contain 8760 rows and 14 columns
- Three categorical feature ‘Seasons’, ‘Holiday’ & ‘Functioning Day’
- One datetime column ‘Date’
- We have some numerical type variable such ad temperature, humidity, wind, visibility, dew point temp, solar radiation, rainfall, snowfall which shows the environmental conditions for that particular hour of the day.

DATA SUMMARY

- There are no missing values present.
- There are no Duplicate values present.
- There are no null values.
- The dependent variable is ‘rented bike count’ which we need to make predictions on.
- The dataset shows hourly rental data for one year(1 December 2017 to 31 November 2018)(365 days).
- We changed the feature ‘Date’ to Day, Month and Year.

FEATURE TYPES

FEATURES

NUMERICAL

- Hour
- Temp
- Humidity
- Wind
- Rainfall
- Snow
- Visibility
- Solar Radiation
- Day
- Month
- Year

CATEGORICAL

- Seasons
- Holiday
- Functioning day

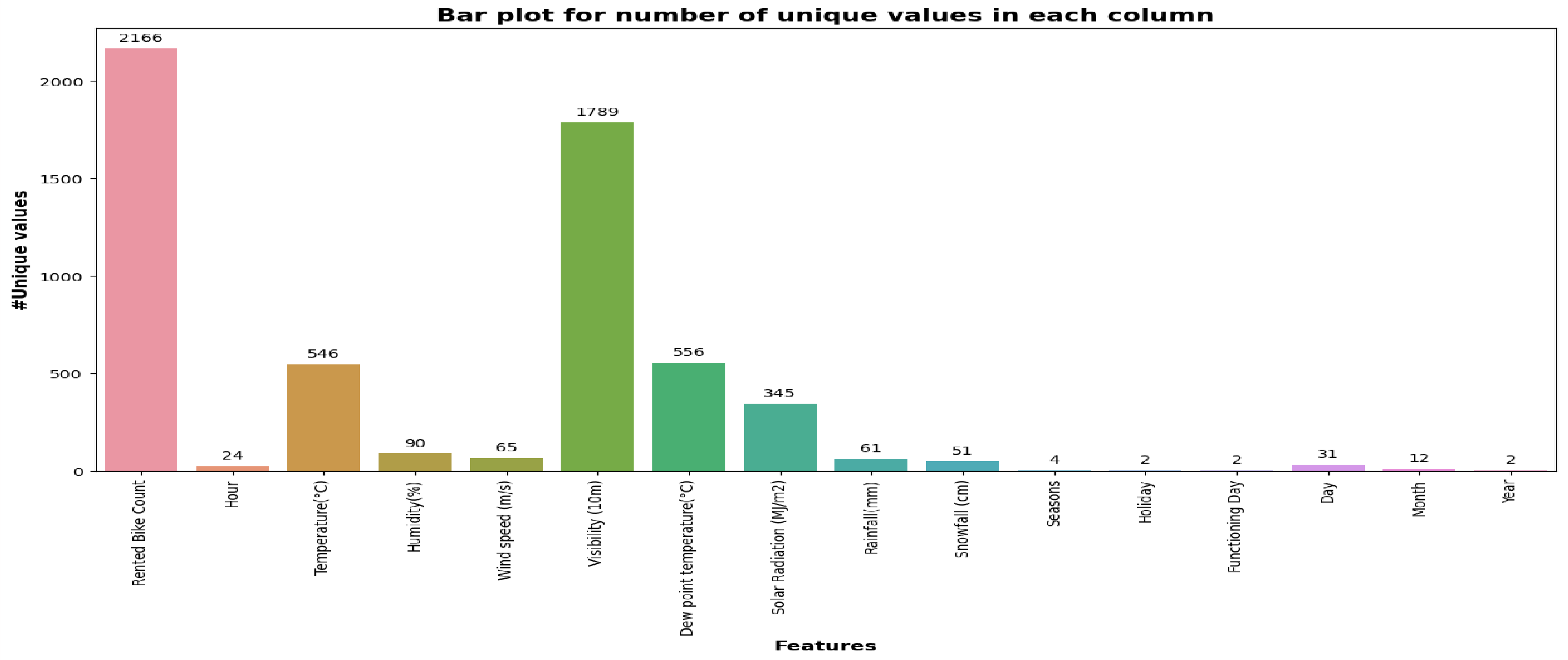
TARGET VARIABLE

Rented Bike Count

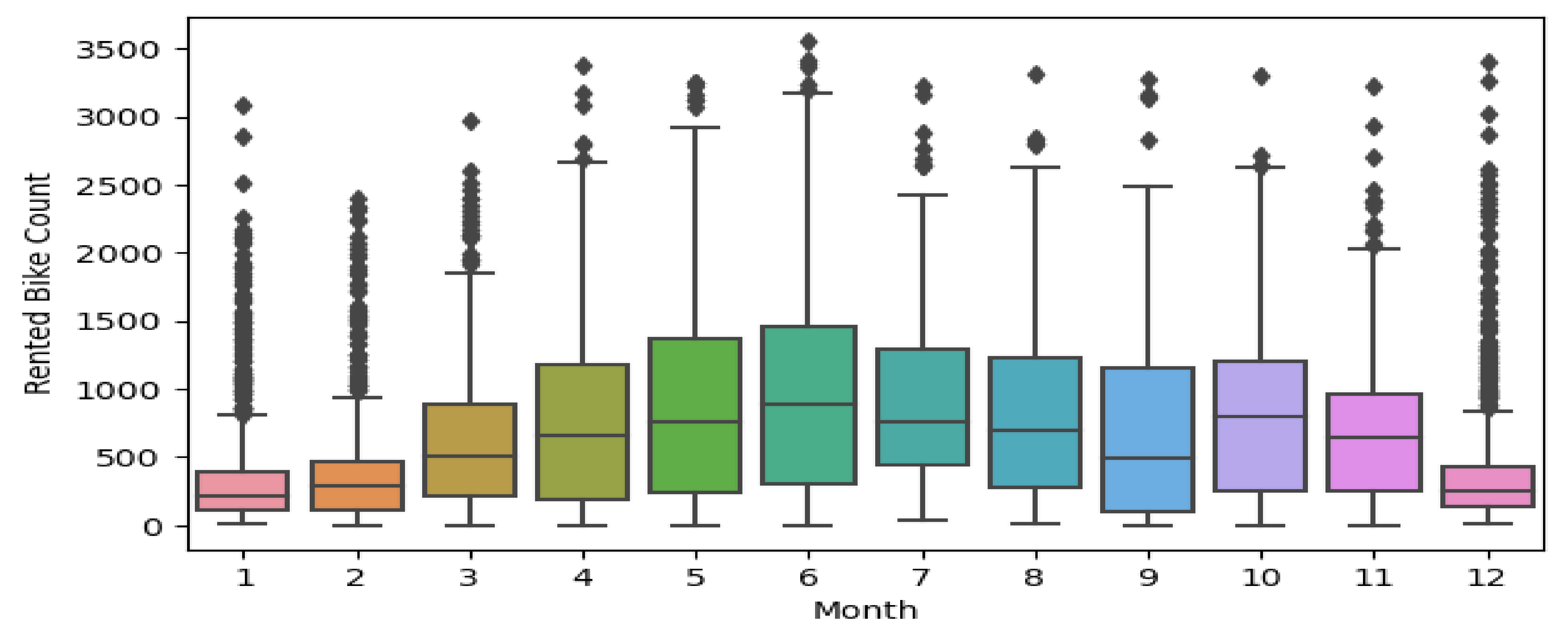
FEATURE SUMMARY

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Windspeed - m/s
- Visibility - 10m
- 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)

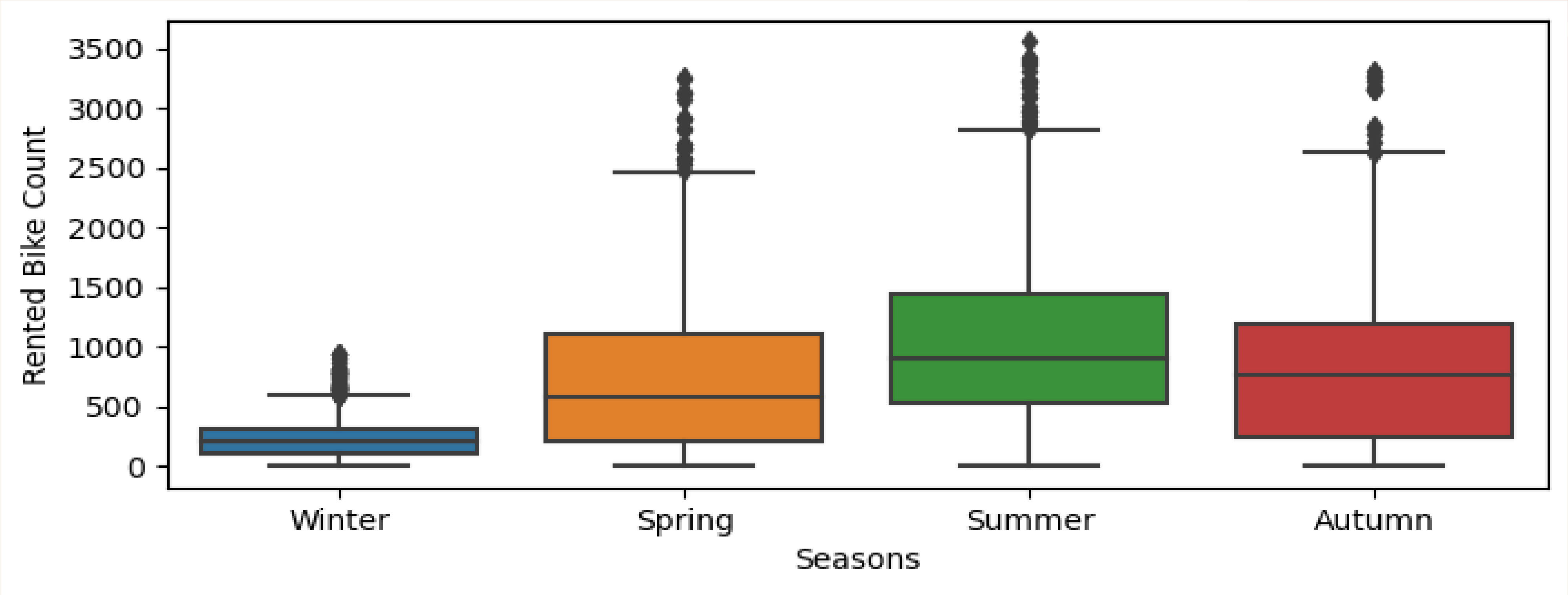
UNIQUE FEATURE OF DATASET



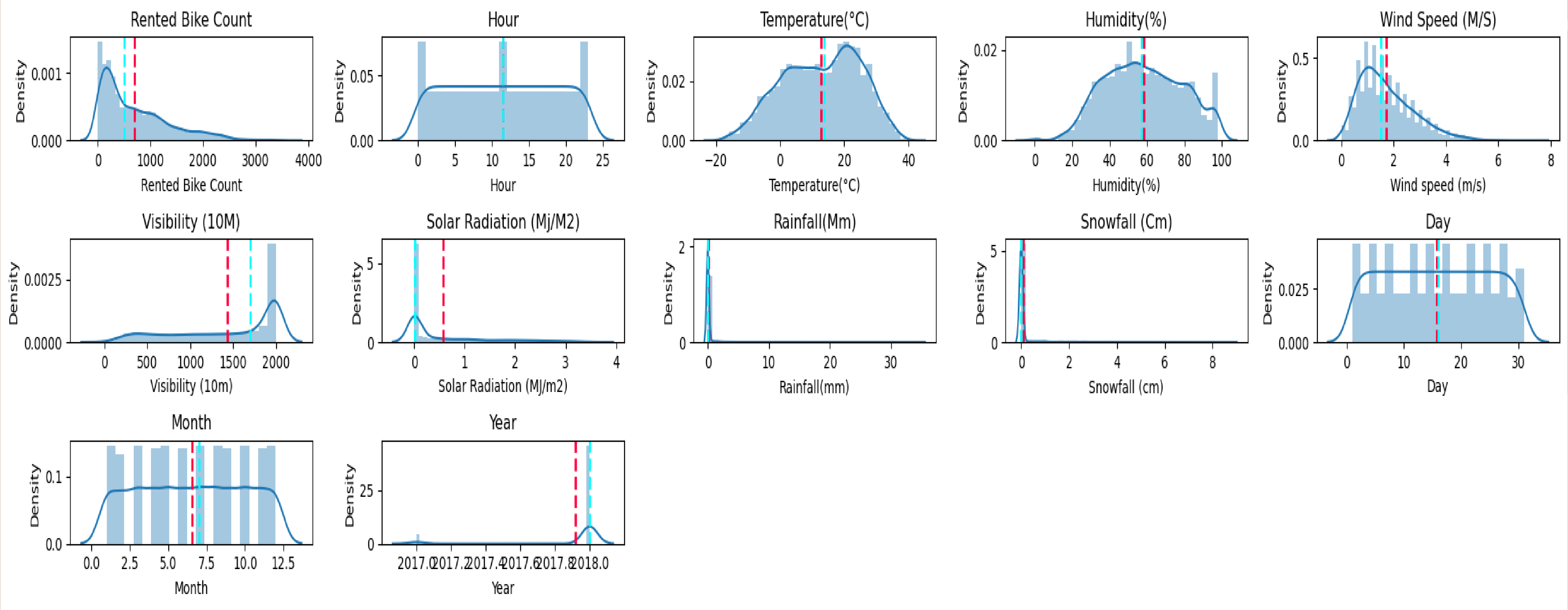
Month wise analysis



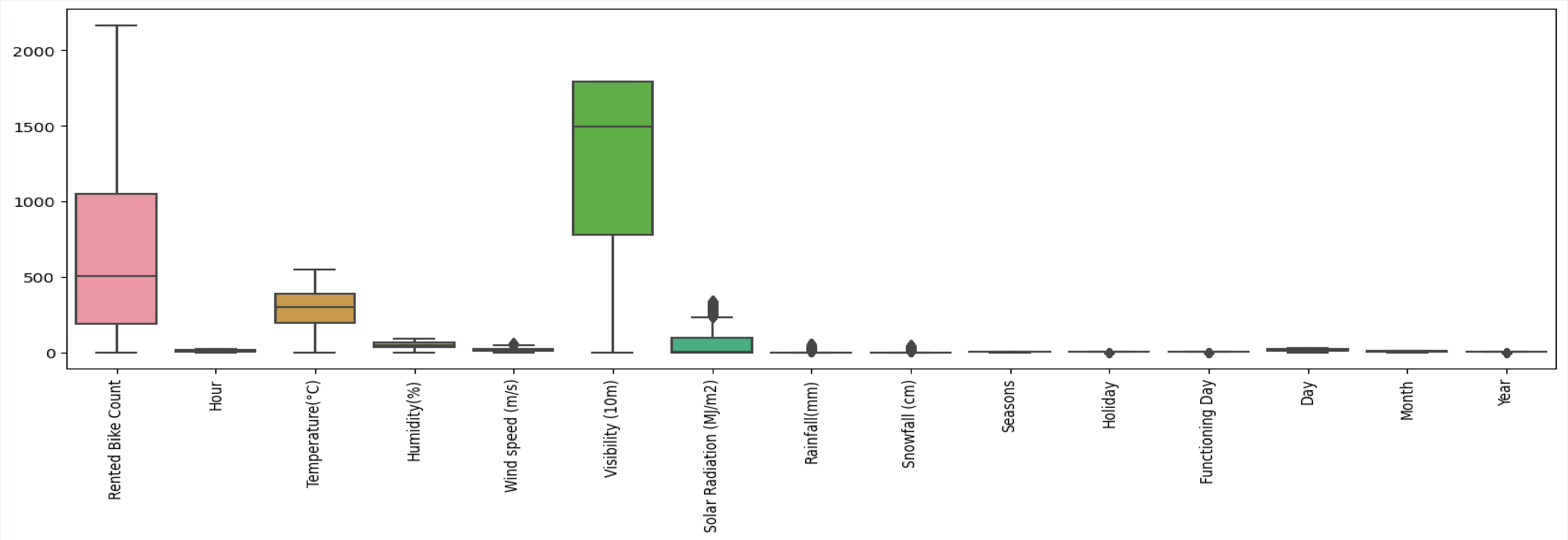
Season by analysis



VISUALIZING DISTRIBUTION

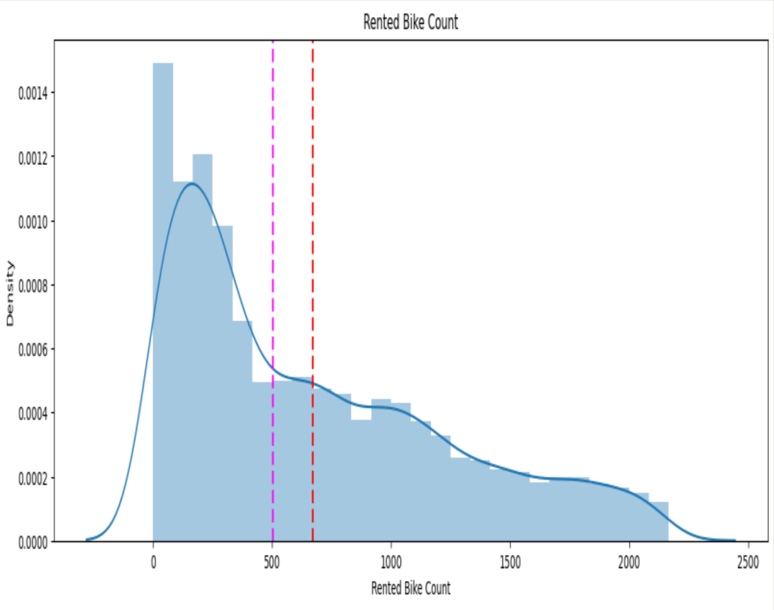


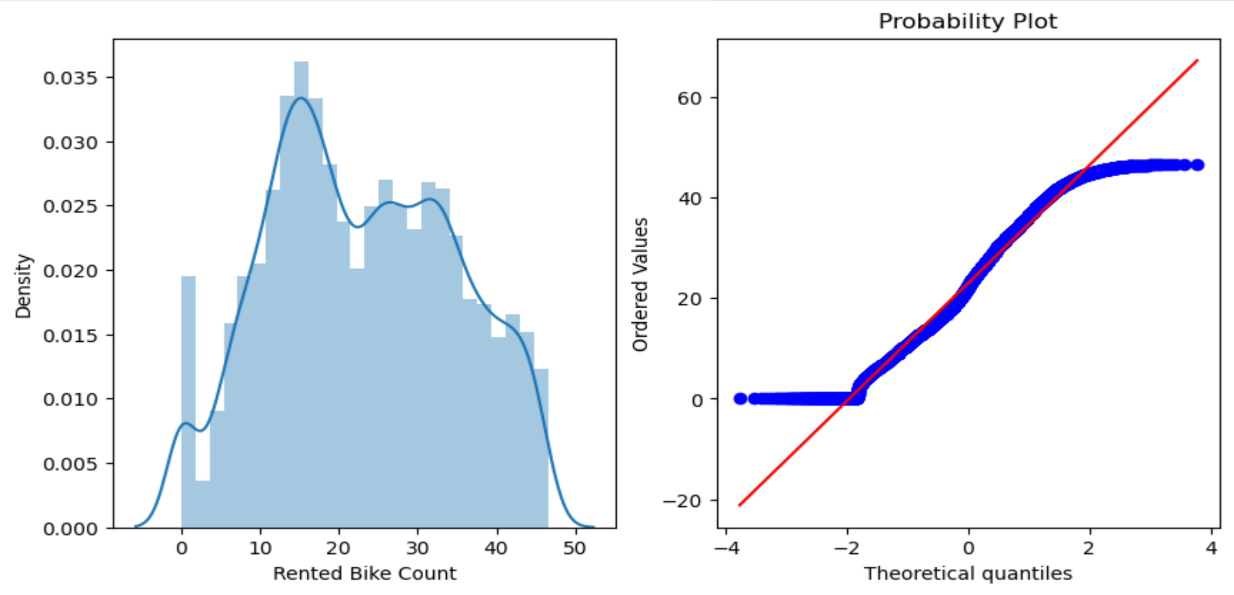
CHECKING OUTLIERS



- We see outliers in some columns like Solar radiation, Wind, Rainfall, and Snowfall but lets not treat them because they may not be outliers as snowfall, rainfall etc. themselves are rare event in some countries.

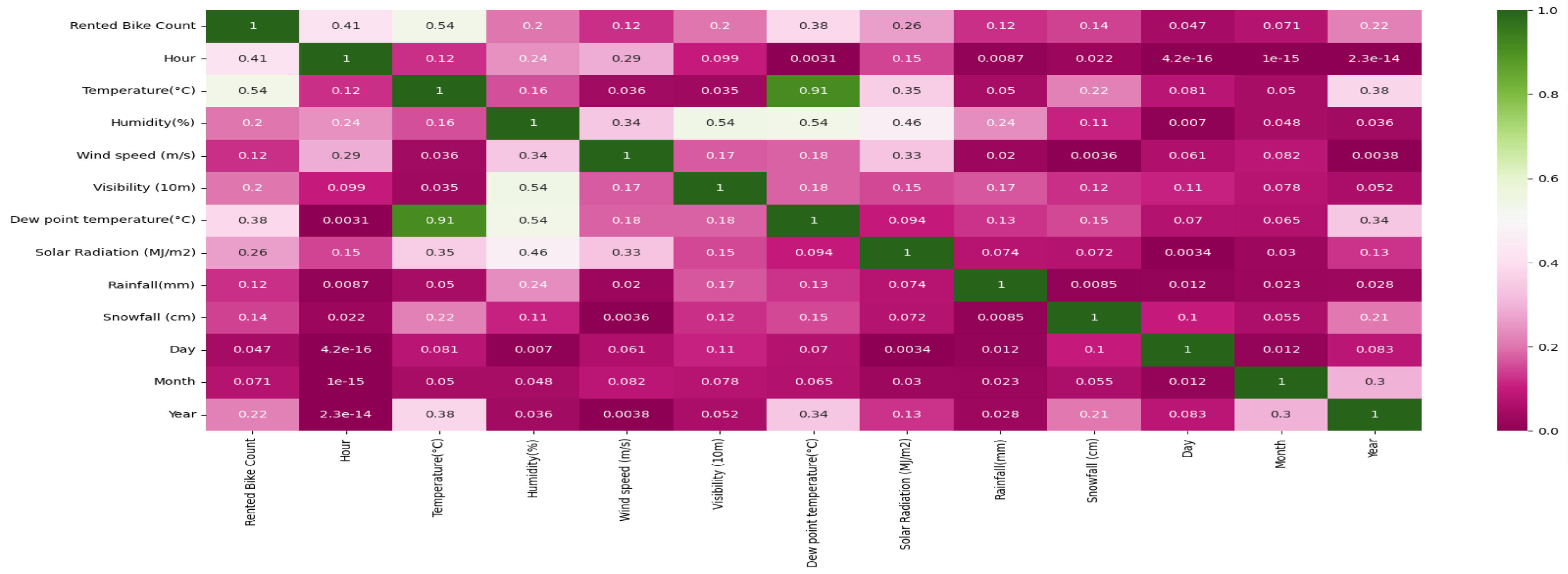
DEPENDENT VARIABLE





- Earlier the distribution of the target variable was positively skewed. We tried to make this distribution somewhat close to normal distribution.
- First we apply log transform but it did not give desired result, er finally applied square root transformation. We got the favorable results, the skewness value was dropped, which is comparatively closer to the normal distribution.

MULTICOLLINEARITY ANALYSIS



	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)	Day	Month	Year
Rented Bike Count	1	0.41	0.54	0.2	0.12	0.2	0.38	0.26	0.12	0.14	0.047	0.071	0.22
Hour	0.41	1	0.12	0.24	0.29	0.099	0.0031	0.15	0.0087	0.022	4.2e-16	1e-15	2.3e-14
Temperature(°C)	0.54	0.12	1	0.16	0.036	0.035	0.91	0.35	0.05	0.22	0.081	0.05	0.38
Humidity(%)	0.2	0.24	0.16	1	0.34	0.54	0.54	0.46	0.24	0.11	0.007	0.048	0.036
Wind speed (m/s)	0.12	0.29	0.036	0.34	1	0.17	0.18	0.33	0.02	0.0036	0.061	0.082	0.0038
Visibility (10m)	0.2	0.099	0.035	0.54	0.17	1	0.18	0.15	0.17	0.12	0.11	0.078	0.052
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	0.18	0.18	1	0.094	0.13	0.15	0.07	0.065	0.34
Solar Radiation (MJ/m2)	0.26	0.15	0.35	0.46	0.33	0.15	0.094	1	0.074	0.072	0.0034	0.03	0.13
Rainfall(mm)	0.12	0.0087	0.05	0.24	0.02	0.17	0.13	0.074	1	0.0085	0.012	0.023	0.028
Snowfall (cm)	0.14	0.022	0.22	0.11	0.0036	0.12	0.15	0.072	0.0085	1	0.1	0.055	0.21
Day	0.047	4.2e-16	0.081	0.007	0.061	0.11	0.07	0.0034	0.012	0.1	1	0.012	0.083
Month	0.071	1e-15	0.05	0.048	0.082	0.078	0.065	0.03	0.023	0.055	0.012	1	0.3
Year	0.22	2.3e-14	0.38	0.036	0.0038	0.052	0.34	0.13	0.028	0.21	0.083	0.3	1

- Temperature and Dew point temperature are almost 0.91 correlated, So it's generate multicollinearity issue. so we drop Dew point temperature feature.

MODEL BUILDING PREREQUISITE

Feature Scaling :-

Standardization:

- It is a step of data pre processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.
- Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$X\_new = (X - mean)/Std$$

MODEL BUILDING PREREQUISITE

Normalization:-

- Normalization scales our feature to a predefined range (normally the (0-1) or (-1 to 1) range), independently of the statistical distribution they follow. It does this using **minimum and maximum values** of each feature in the dataset.

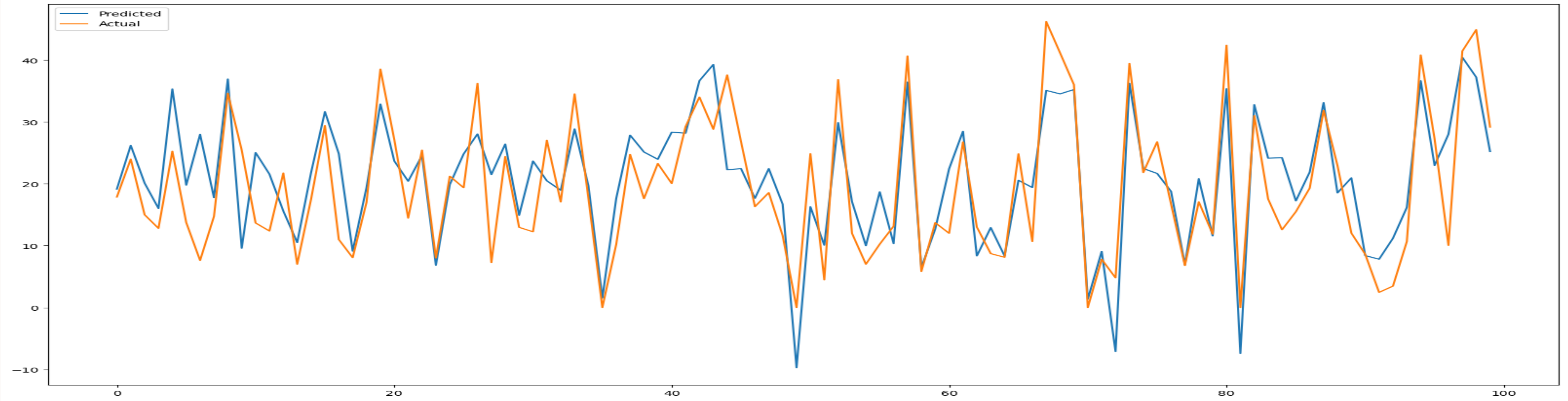
$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

MODEL BUILDING PREREQUISITE

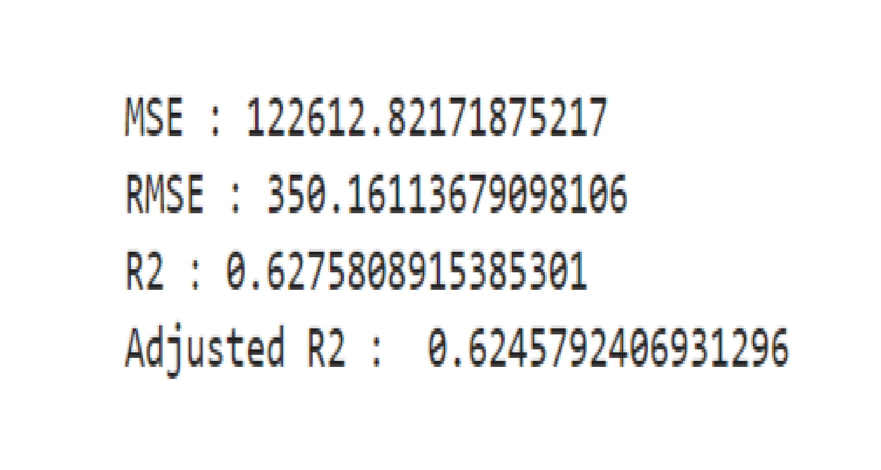
- Defining a new function called **tnt_model** which takes model, X_train, y_train, X_test, y_test and print evaluation matrix like MSE, RMSE, R2, Adejusted R2. Also plots the feature importance based on the algorithm used.
- We also defined range of values for hyperparameters such as:

1. Number of trees: n_estimators =[80,100,150]
2. Maximum depth of trees: [15,20,30]
3. Min no of sample required for split a node: [40,60]

LINEAR REGRESSION

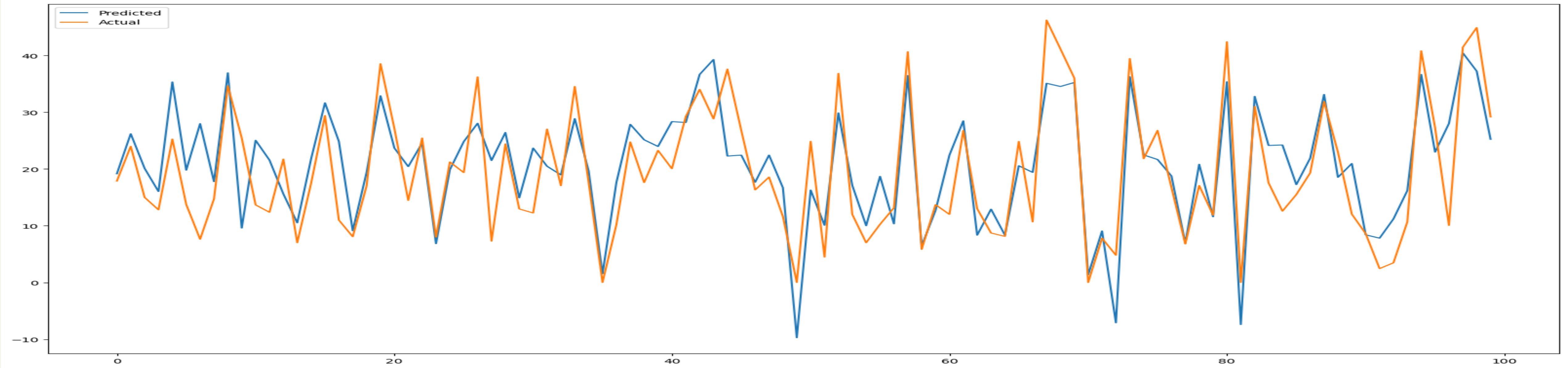


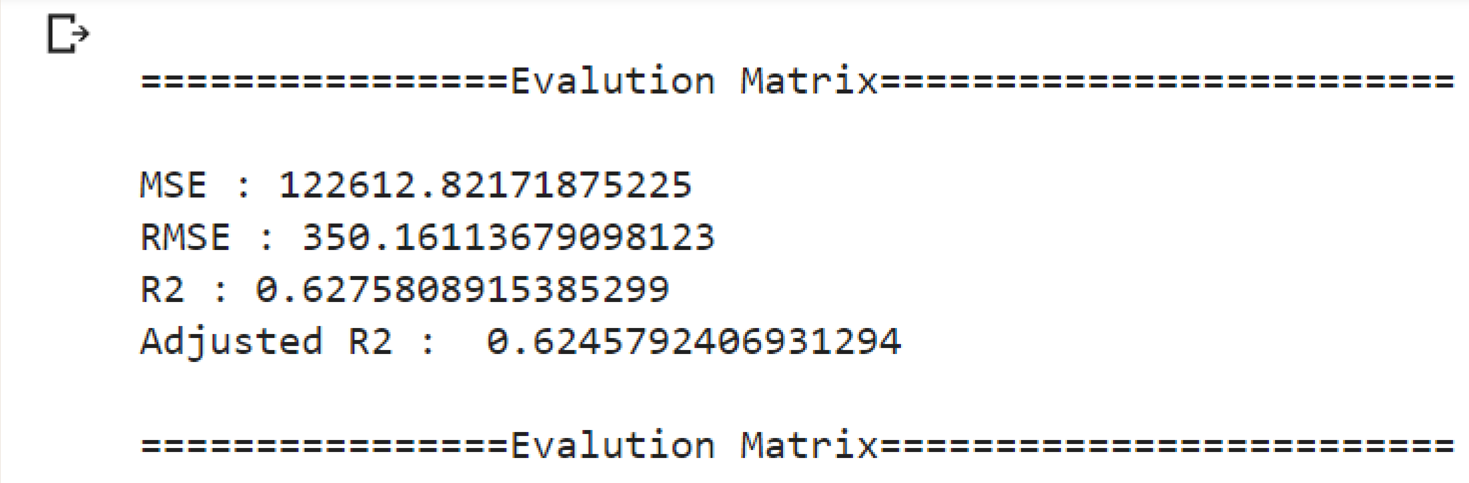
- We plotted the graph of actual and predicted dependent variable ‘Rented bike count’.
- Since the performance of simple linear model is not so good. We experienced with some complex models.



```
MSE : 122612.82171875217
RMSE : 350.16113679098106
R2 : 0.6275808915385301
Adjusted R2 :  0.6245792406931296
```

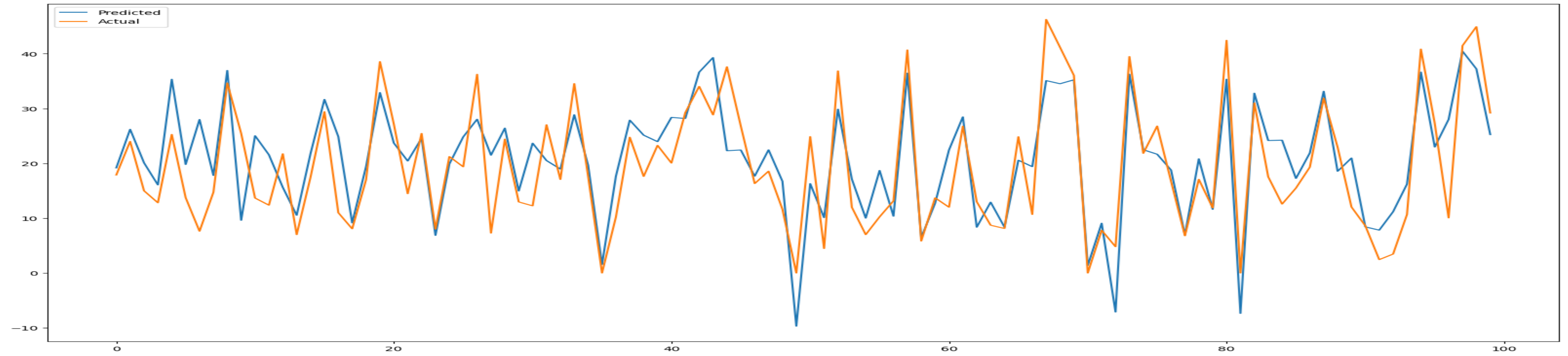
LASSO REGRESSION

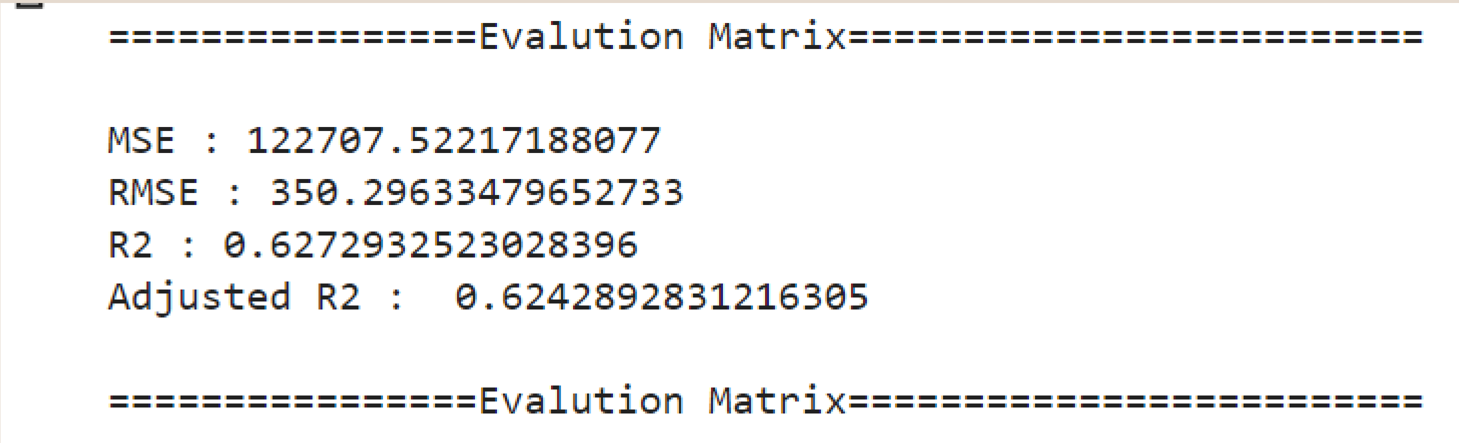




```
================Evalution Matrix==========================

MSE : 122612.82171875225
RMSE : 350.16113679098123
R2 : 0.6275808915385299
Adjusted R2 :  0.6245792406931294

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

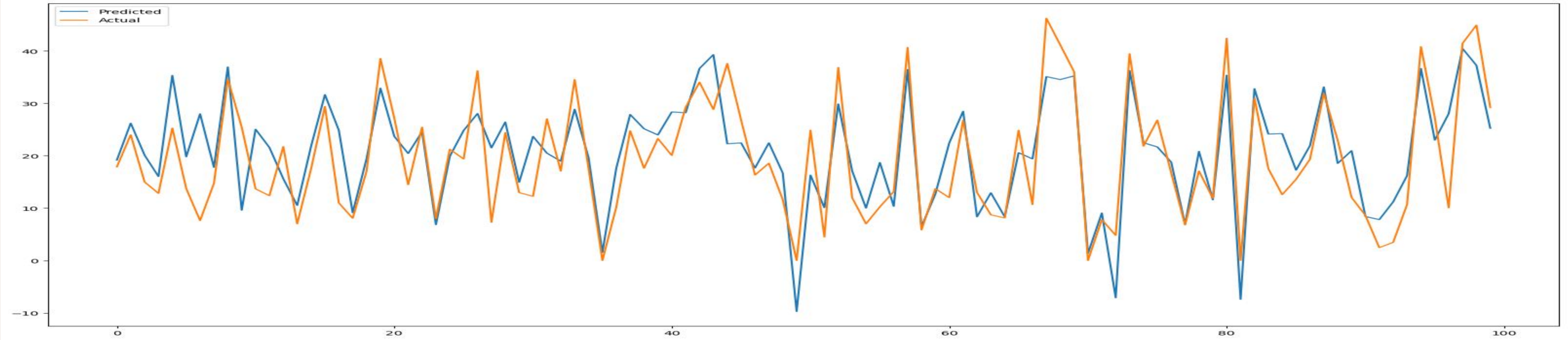
RIDGE REGRESSION

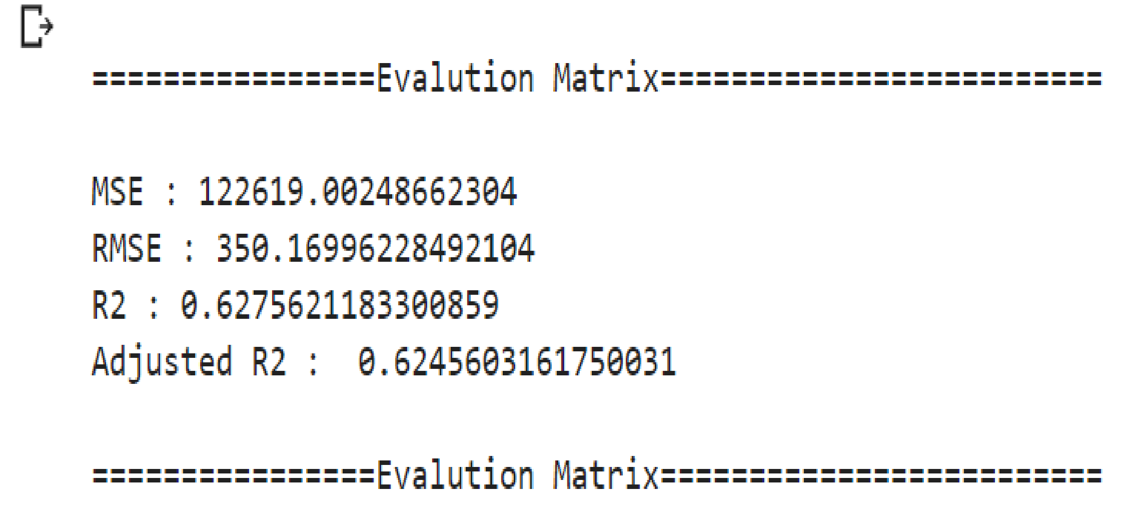




```
================Evalution Matrix===========================

MSE : 122707.52217188077
RMSE : 350.29633479652733
R2 : 0.6272932523028396
Adjusted R2 :  0.6242892831216305

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

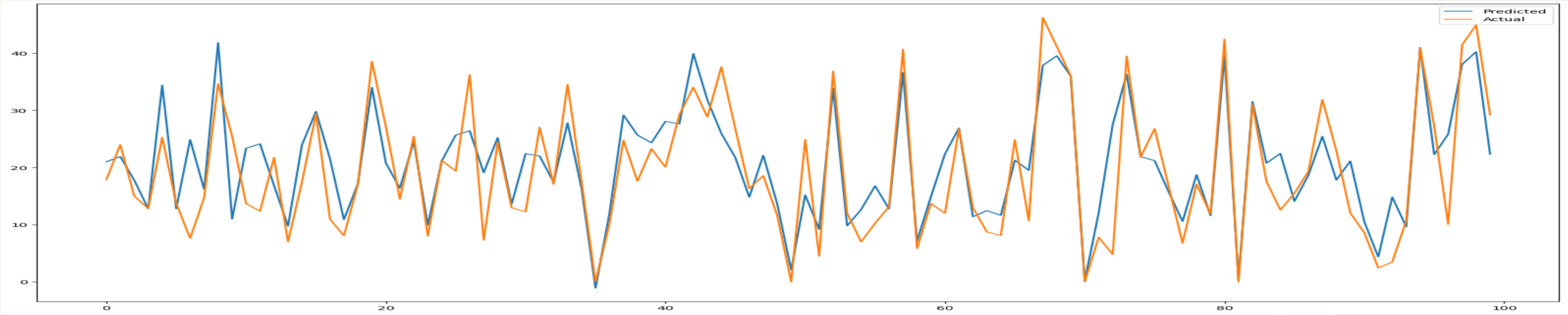
ELASTICNET REGRESSION

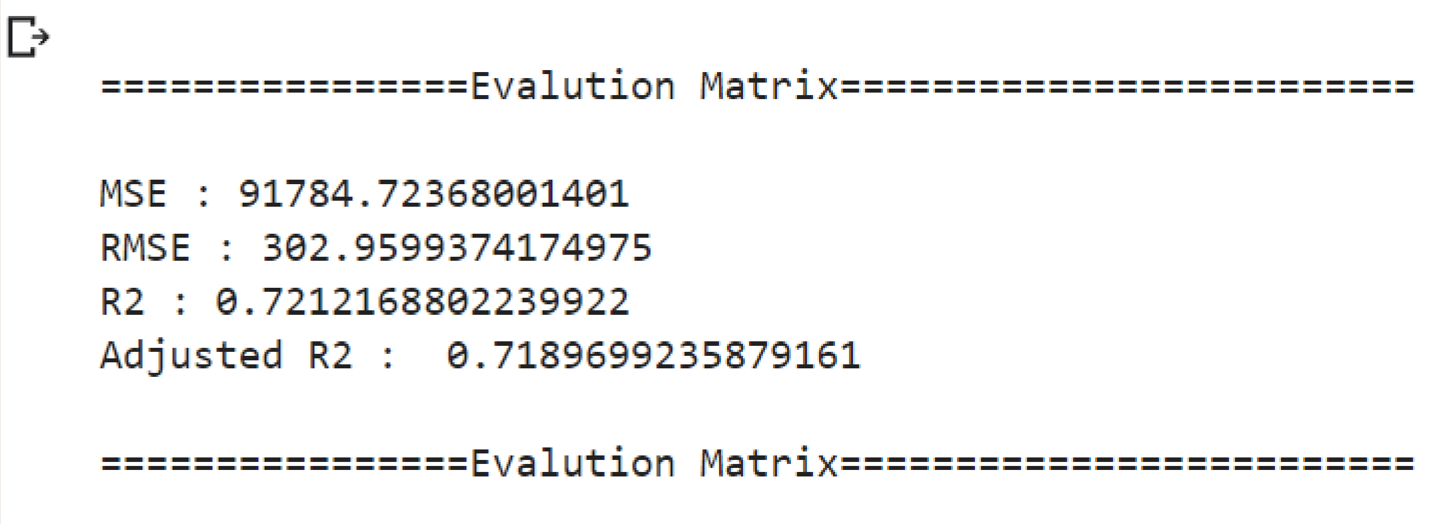




```
================Evalution Matrix=========================

MSE : 122619.00248662304
RMSE : 350.16996228492104
R2 : 0.6275621183300859
Adjusted R2 :  0.6245603161750031

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

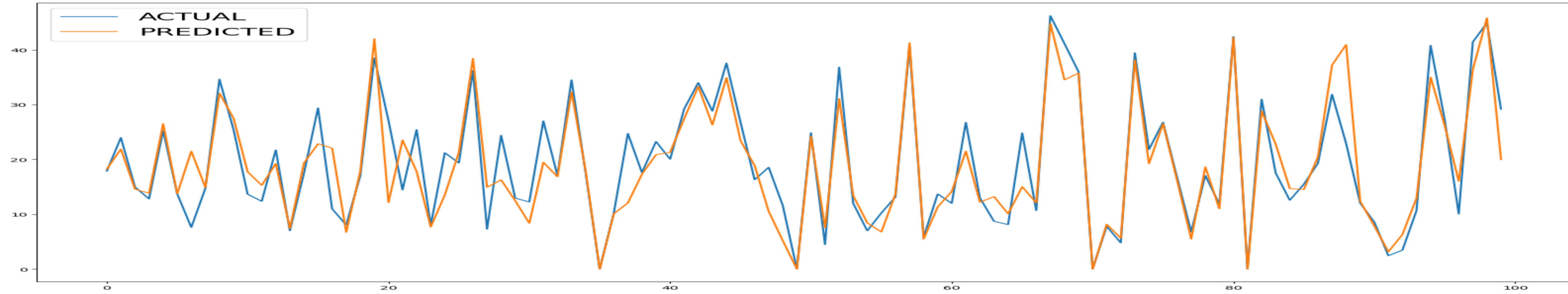
POLYNOMIAL REGRESSION



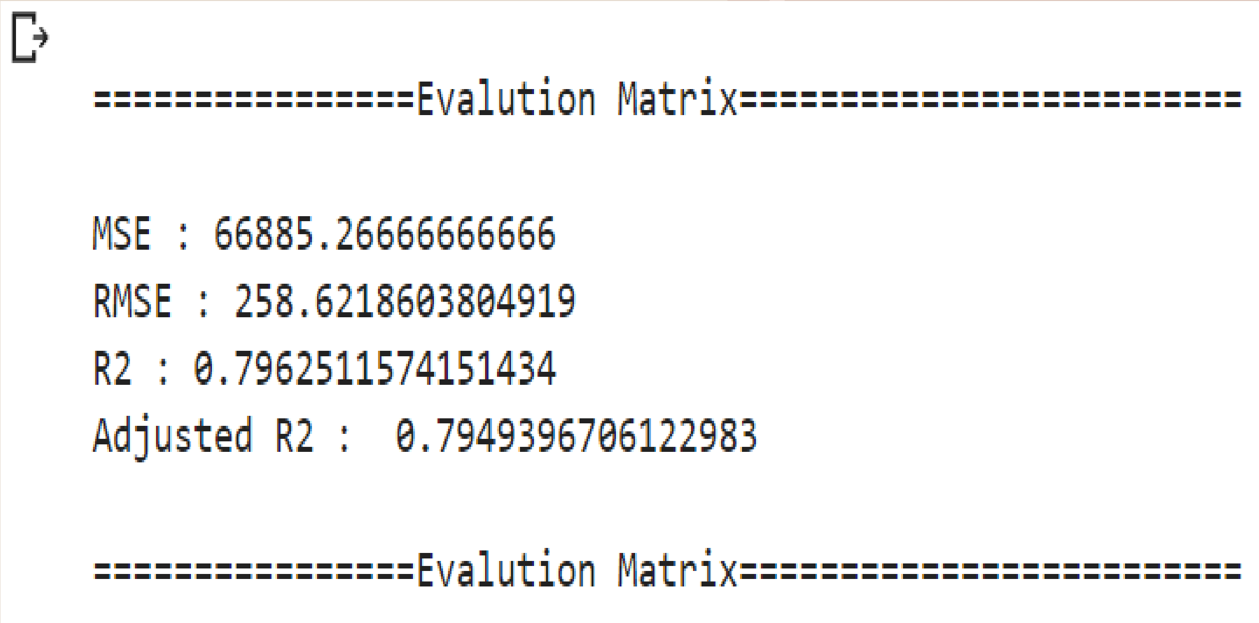


```
===============Evalution Matrix========================

MSE : 91784.72368001401
RMSE : 302.9599374174975
R2 : 0.7212168802239922
Adjusted R2 :  0.7189699235879161

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

DECISION TREE REGRESSOR

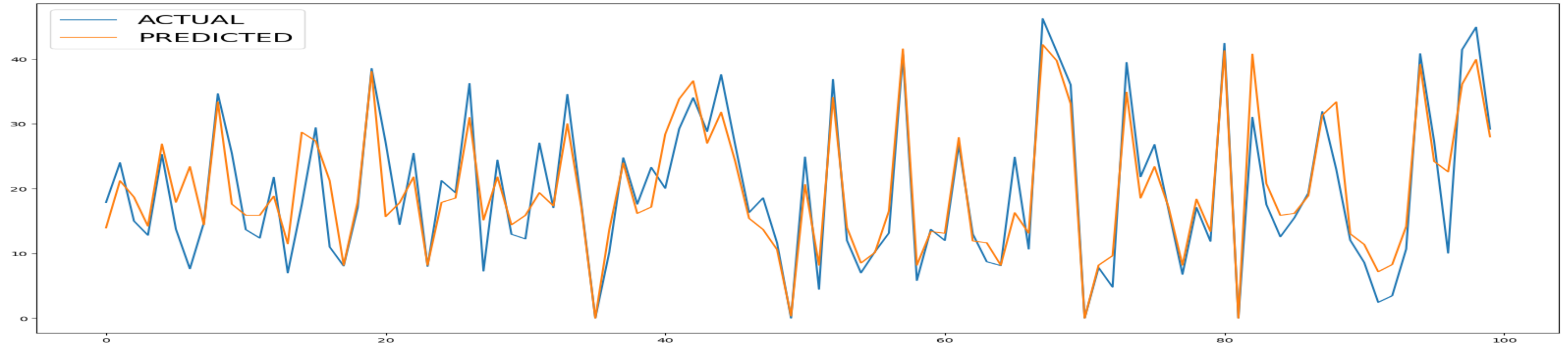


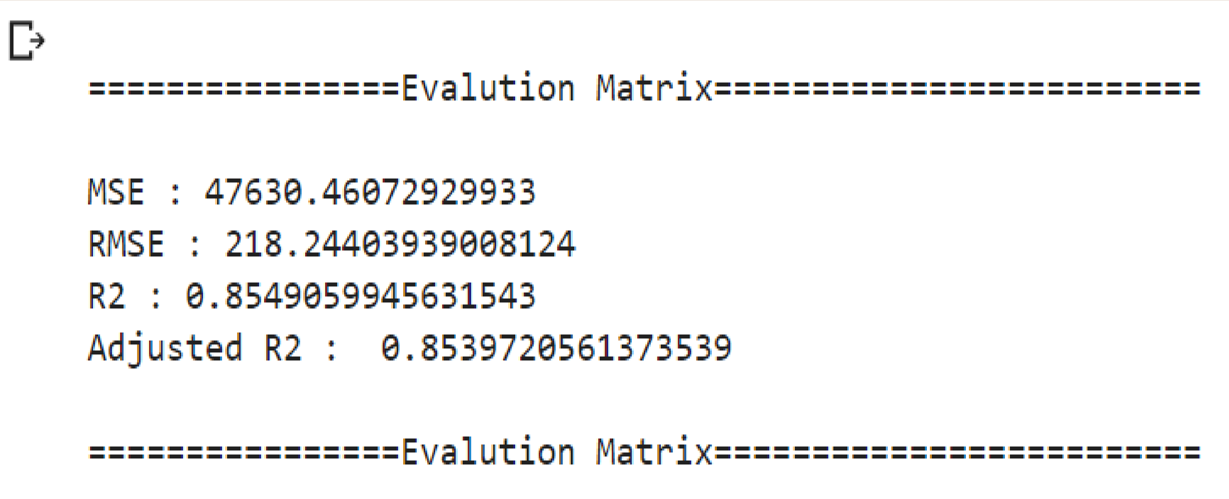
DecisionTreeRegressor performs well better than linear and regularization reg with a test R2 score is more than 75%



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


MSE : 66885.26666666666
RMSE : 258.6218603804919
R2 : 0.7962511574151434
Adjusted R2 :  0.7949396706122983


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

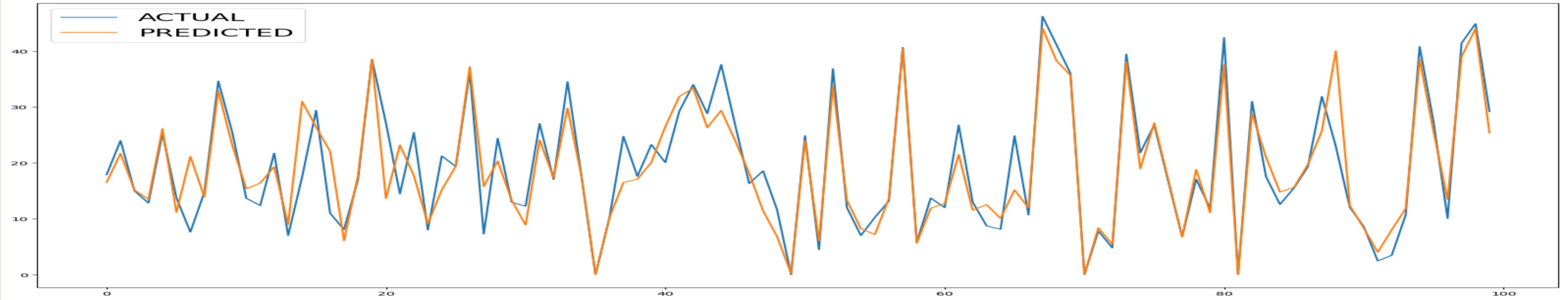
RANDOM FOREST REGRESSOR

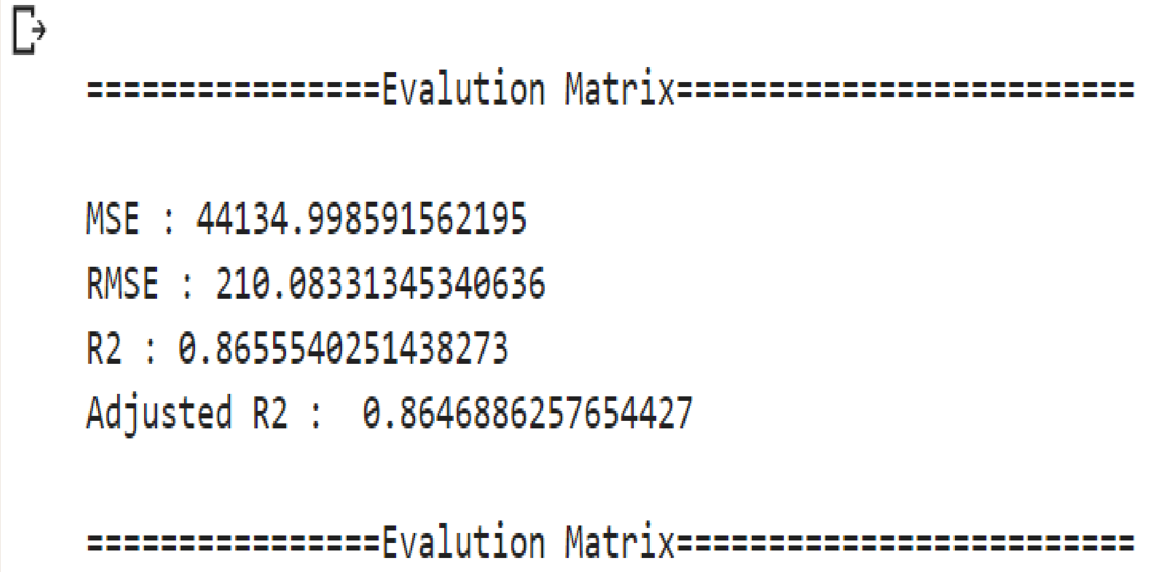




```
================Evalution Matrix=========================

MSE : 47630.46072929933
RMSE : 218.24403939008124
R2 : 0.8549059945631543
Adjusted R2 :  0.8539720561373539

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

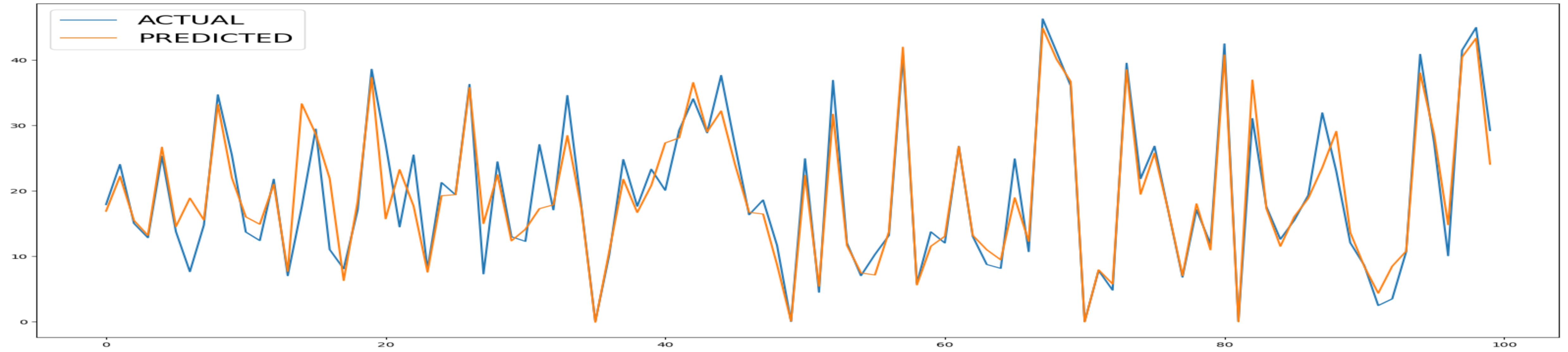
GRADIENT BOOSTING



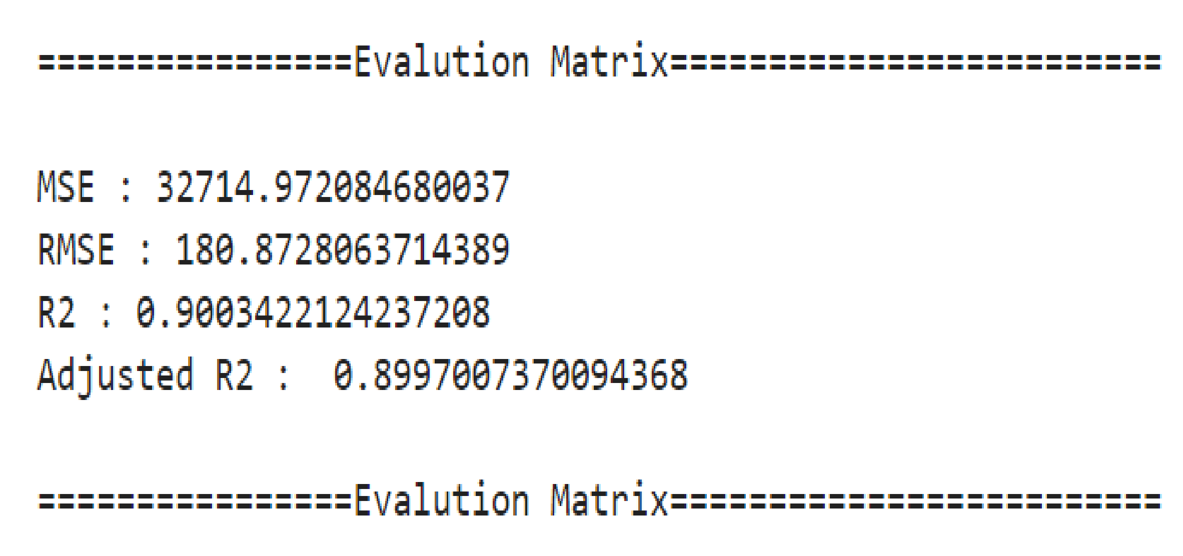


```
===============Evalution Matrix========================


MSE : 44134.998591562195
RMSE : 210.08331345340636
R2 : 0.8655540251438273
Adjusted R2 :  0.8646886257654427


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

EXTREME GRADIENT BOOSTING

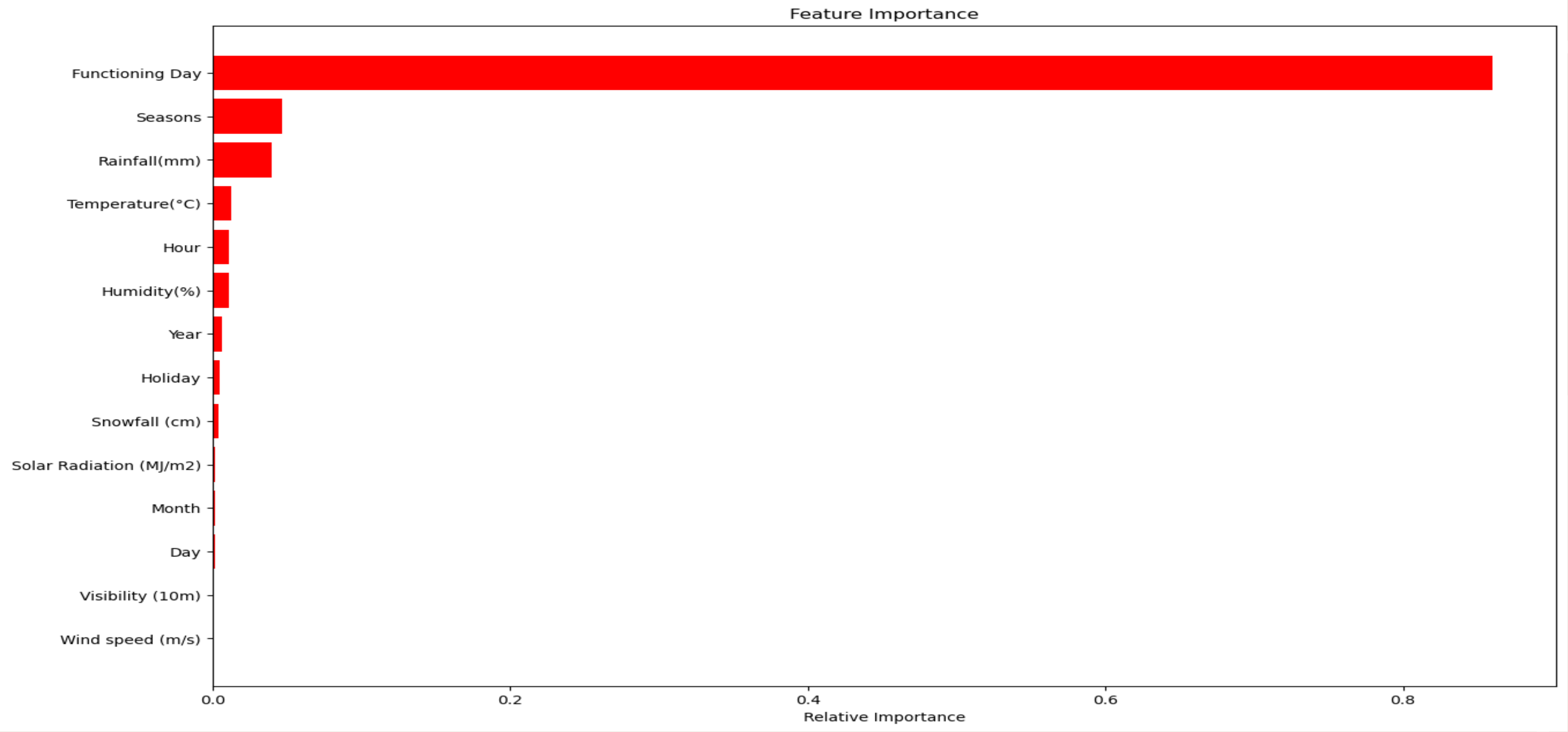


- XGBoost regressor emerges as the best model according to the evolution matrix score.

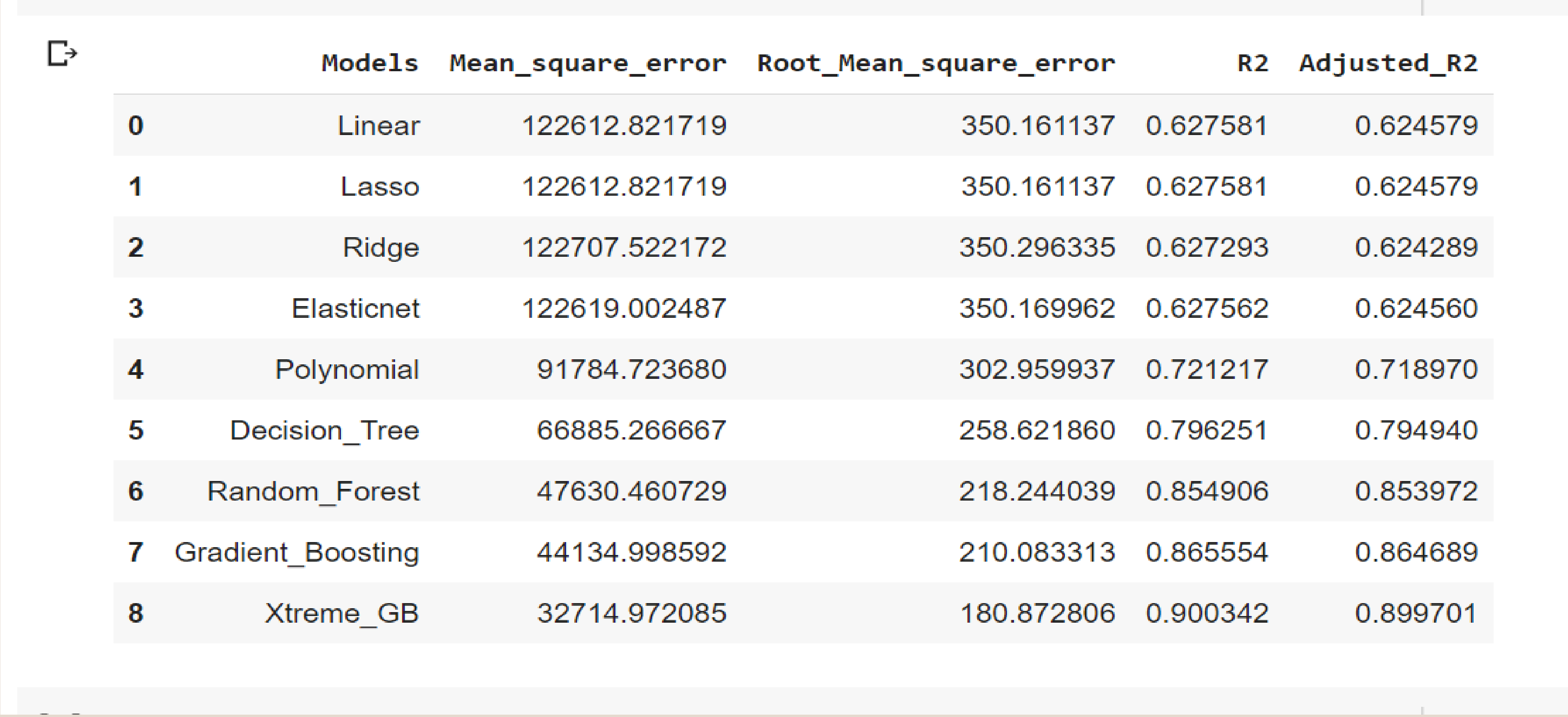


```
================Evalution Matrix=========================


MSE : 32714.972084680037
RMSE : 180.8728063714389
R2 : 0.9003422124237208
Adjusted R2 :  0.8997007370094368


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

EXTREME GRADIENT BOOSTING

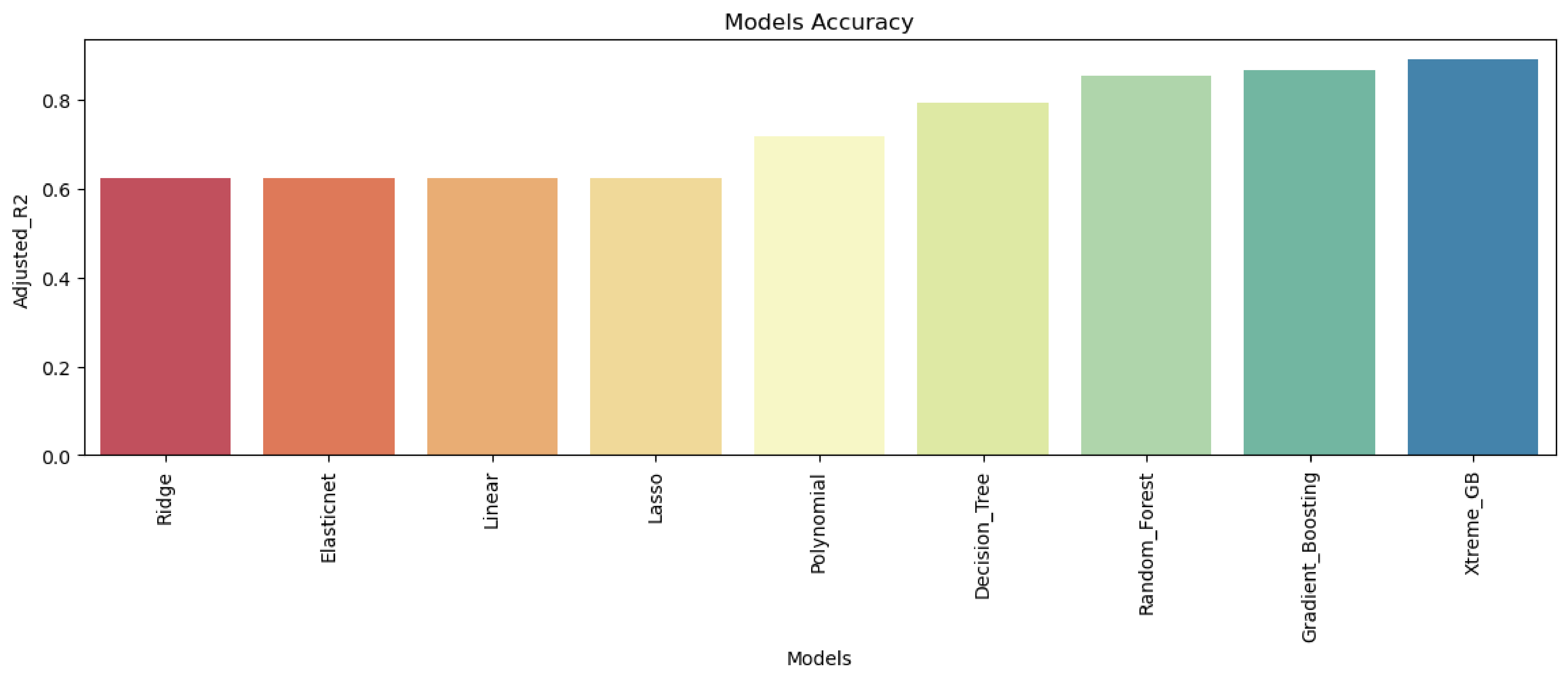


Matrics of different Models



	Models	Mean_square_error	Root_Mean_square_error	R2	Adjusted_R2
0	Linear	122612.821719	350.161137	0.627581	0.624579
1	Lasso	122612.821719	350.161137	0.627581	0.624579
2	Ridge	122707.522172	350.296335	0.627293	0.624289
3	Elasticnet	122619.002487	350.169962	0.627562	0.624560
4	Polynomial	91784.723680	302.959937	0.721217	0.718970
5	Decision_Tree	66885.266667	258.621860	0.796251	0.794940
6	Random_Forest	47630.460729	218.244039	0.854906	0.853972
7	Gradient_Boosting	44134.998592	210.083313	0.865554	0.864689
8	Xtreme_GB	32714.972085	180.872806	0.900342	0.899701

Model Accuracy



Conclusion

- The independent variable in the data does not have a good linear relation with the target variable so the simple linear model was not performing good on this data. Tree based algorithm perform well in this case.
- 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 returning from their work at the evening 6PM.
- After performing the various models the Gradient Boosting and Extreme Gradient Boosting 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 Gradient Boosting and Extreme Gradient Boosting models !
- We can use either Gradient Boosting and Extreme Gradient Boosting model for the bike rental stations.

Thank You

Vikas Kumar Manjhi
vikasmanjhi.it@gmail.com