Capstone Project-2

Seoul Bike Sharing Demand Prediction

(Supervised Machine Learning Regression)

By
Chand Kamble
Rupini Gandhaveeti
Kaushik Das

Points To Discuss.

- Problem Statement
- Data Summary
- Insights For Dataset
- EDA
- Feature Engineering
- Applying ML Algorithms
- Comparing Different ML Models
- Challenges
- Conclusions

Defining 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.

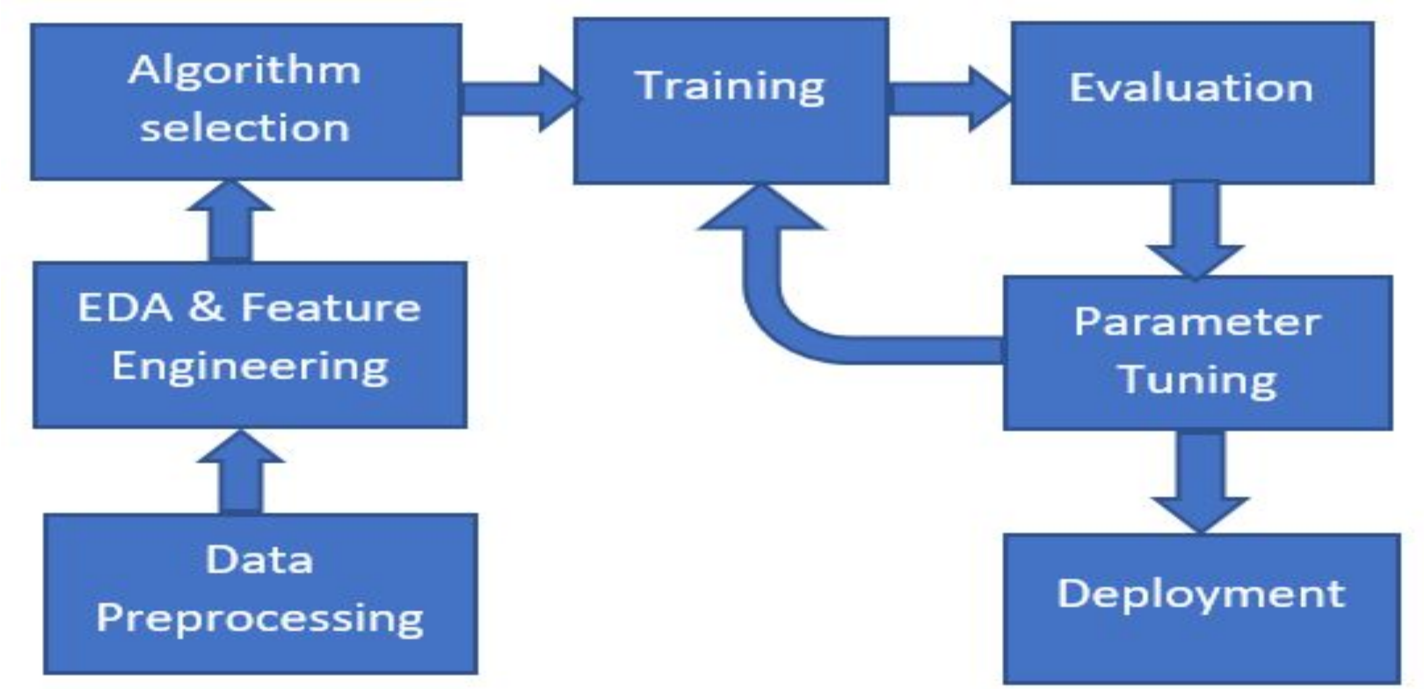
Data summary

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

Attribute Information:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind Speed - 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)

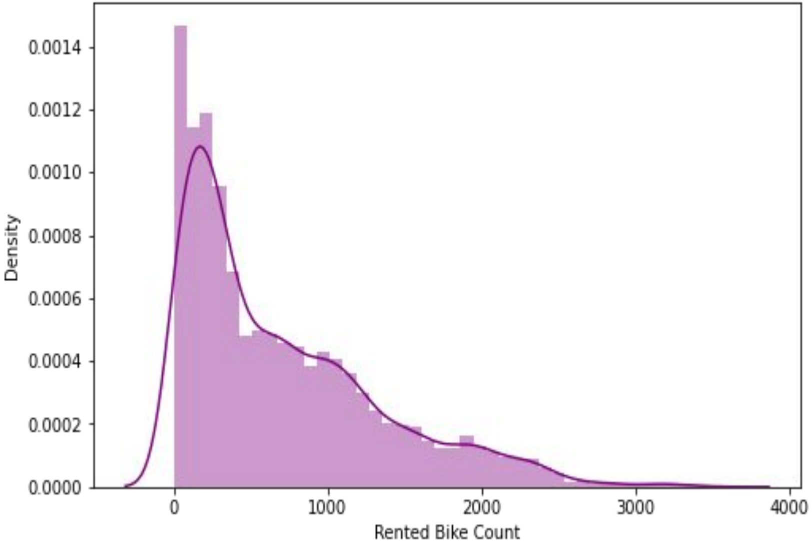
Workflow

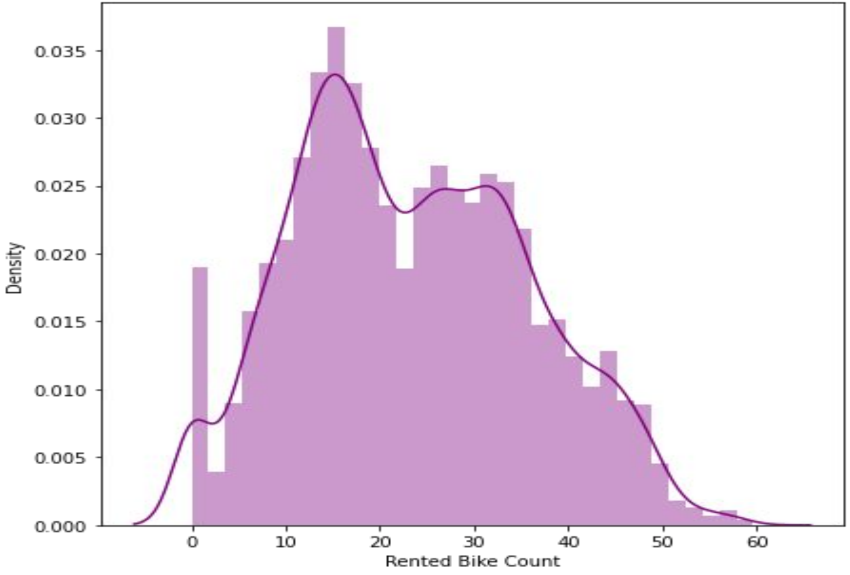

Algorithm selection
Training
Evaluation
EDA & Feature Engineering
Parameter Tuning
Data Preprocessing
Deployment

Insights For Dataset

- Dataset Contains 8760 Rows And 14 Columns.
- Categorical Feature: Seasons ,Holiday And Functioning Day.
- One Datetime Feature ‘date’.
- Numerical Feature:Date ,Hour, Temperature, Humidity,Wind Speed, Visibility, Dew Point Temperature,Solar Radiation,Rainfall,Snowfall ,Rented Bike Count.
- There Are No Missing Values Present.
- There Are No Null Values And Duplicate Values Present.
- We Rename The Few Features , Temperature, Humidity ,Wind Speed, Visibility, Dew Point Temperature,Solar Radiation,Rainfall,Snowfall ,Rented Bike Count.
- we changed the datatype of date column from ‘object’ to ‘datetime’ and Replace The Date Column With New Columns “Year”,”Month”,”Day”.

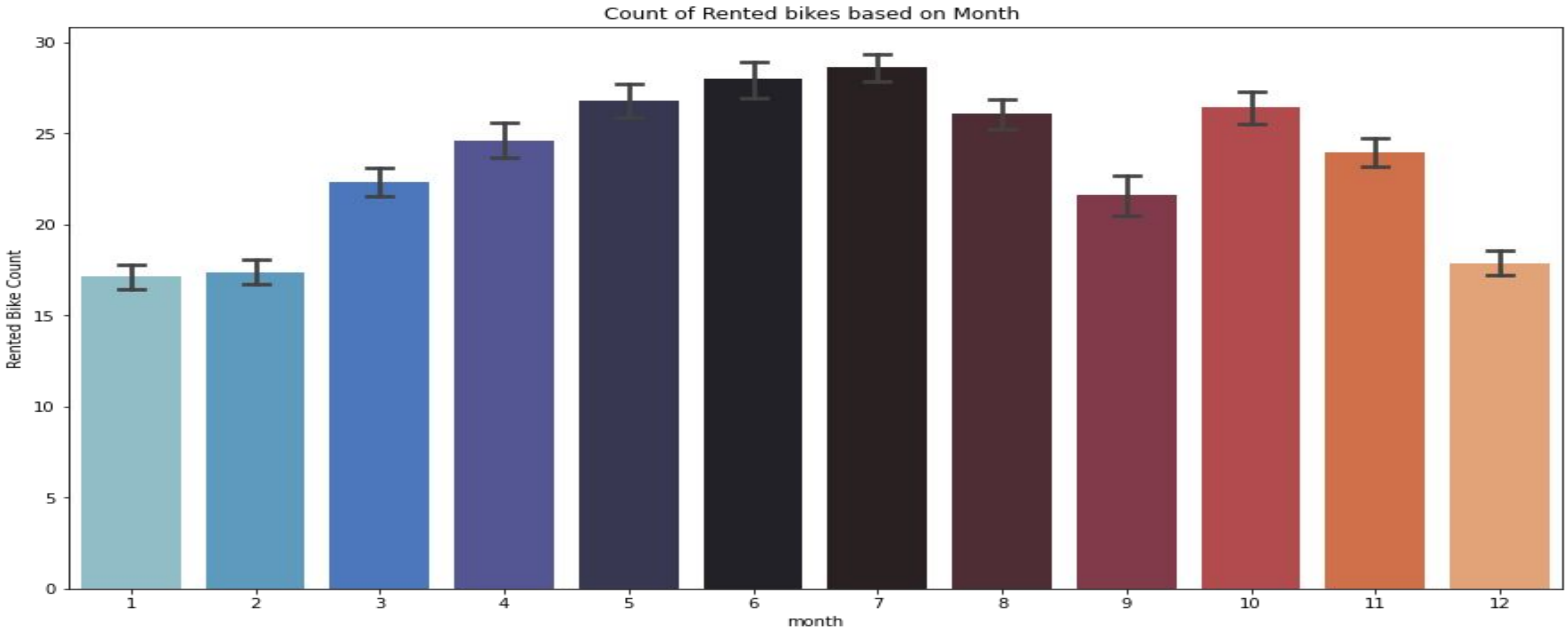
EDA





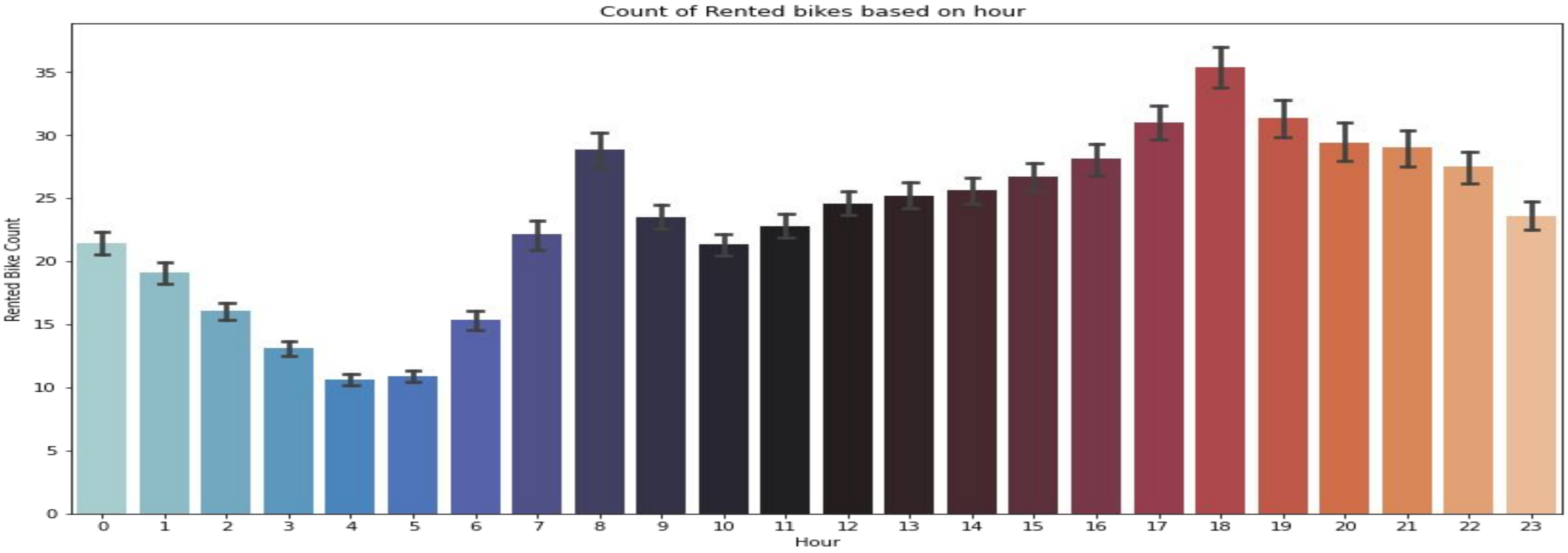
- Above Graph Shows That Rented Bike Count Has Moderate Right Or Positive Skewness . The Distribution Of Dependent Variable Has To Be Normal So We Should Perform Square Root Operation To Make It Normal.
- After Applying Square Root Operation We Get Normal Distribution.

Analysis Of Months Variables.



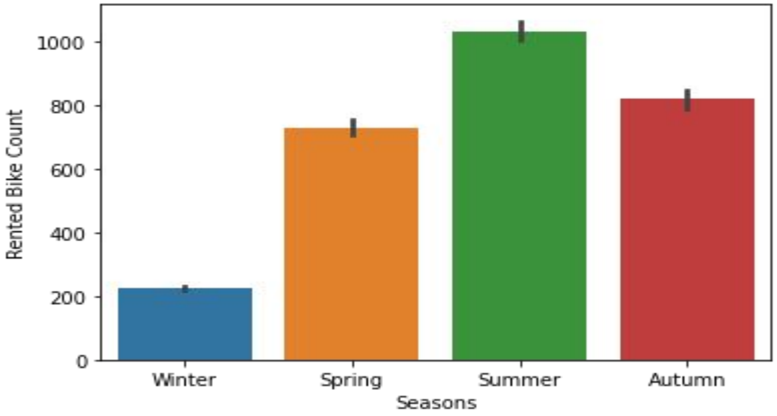
- From The Above Graph We Can Clearly Seen That From Month 5 To 10 Has High Demand Of Rented Bikes As Compared To Other Months These Months Come Under The Summer Season.

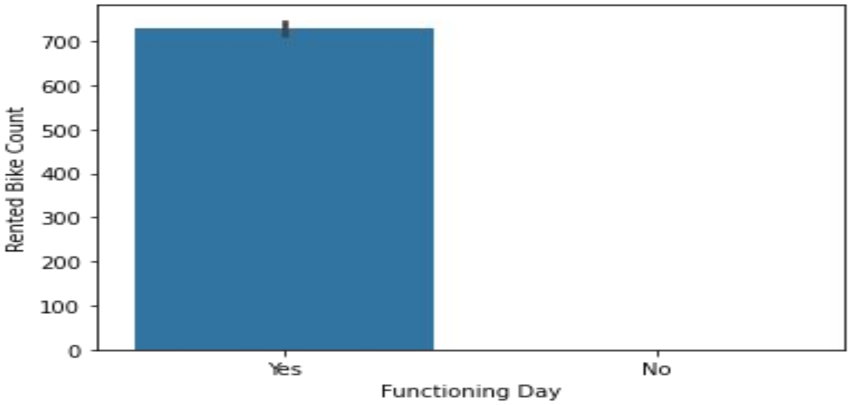
Analysis Of Hour Variable.

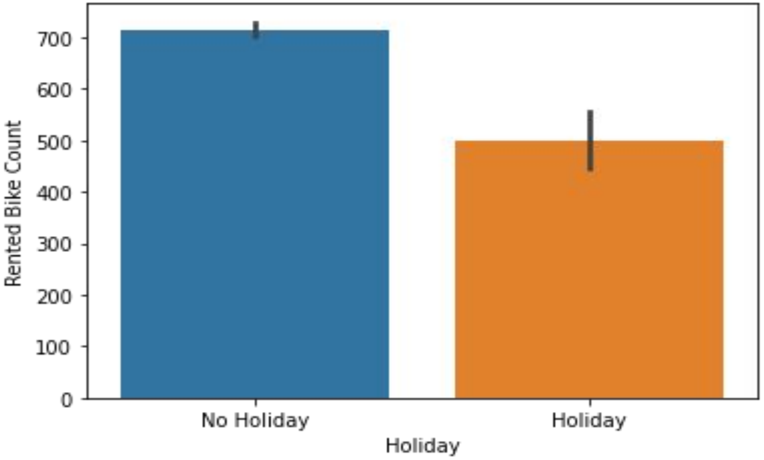


- High Rise Of Rental Bike In 8:00am To 9:00pm Means People Prefer Rented Bike In Rush Hour.
- We Can Clearly See The Demand Rise Most At 8:00am To 9:00 Pm So We Can Say That During Office Opening And Closing Time There Is Much High Demand .

Categorical Vs Rented Bike Count

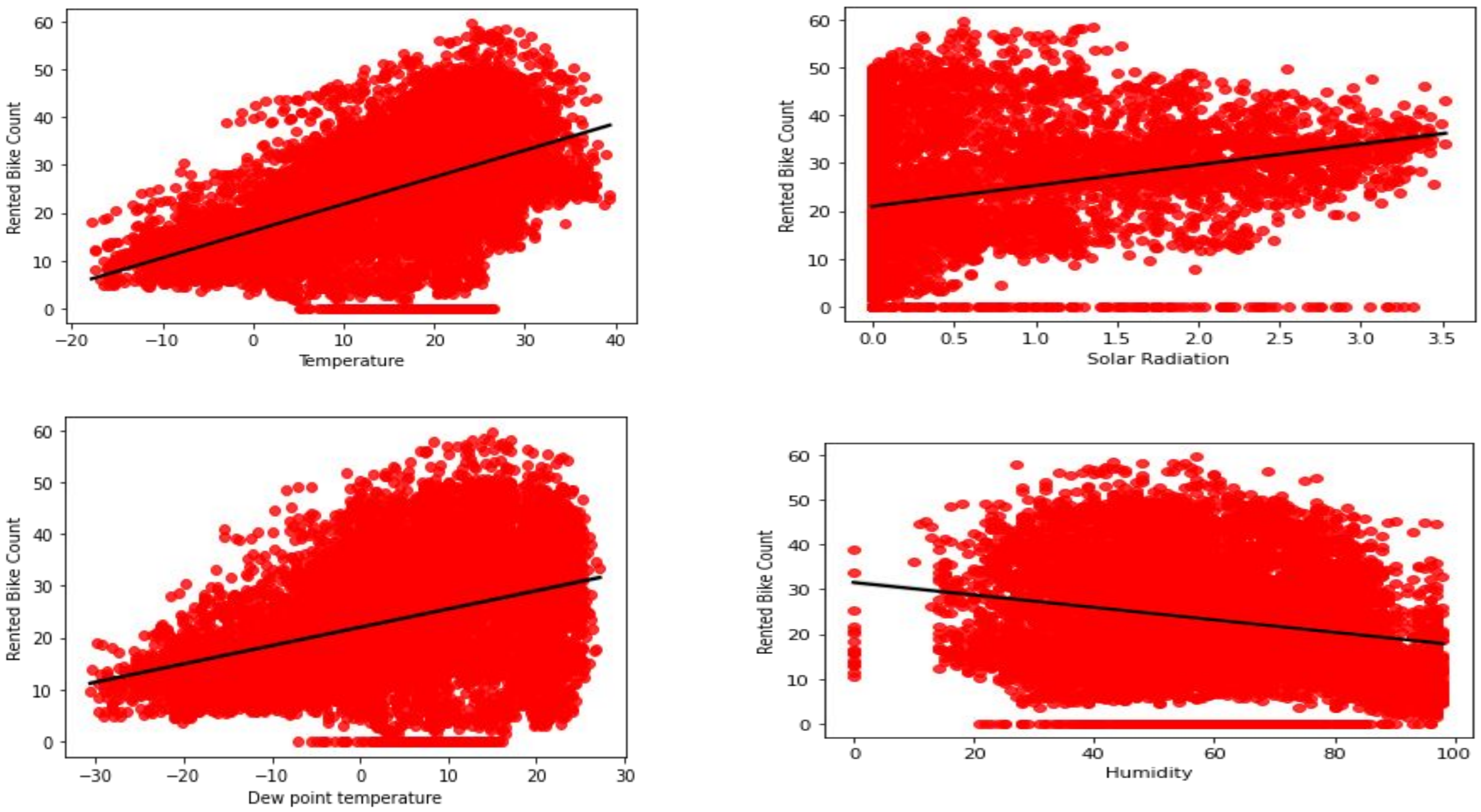




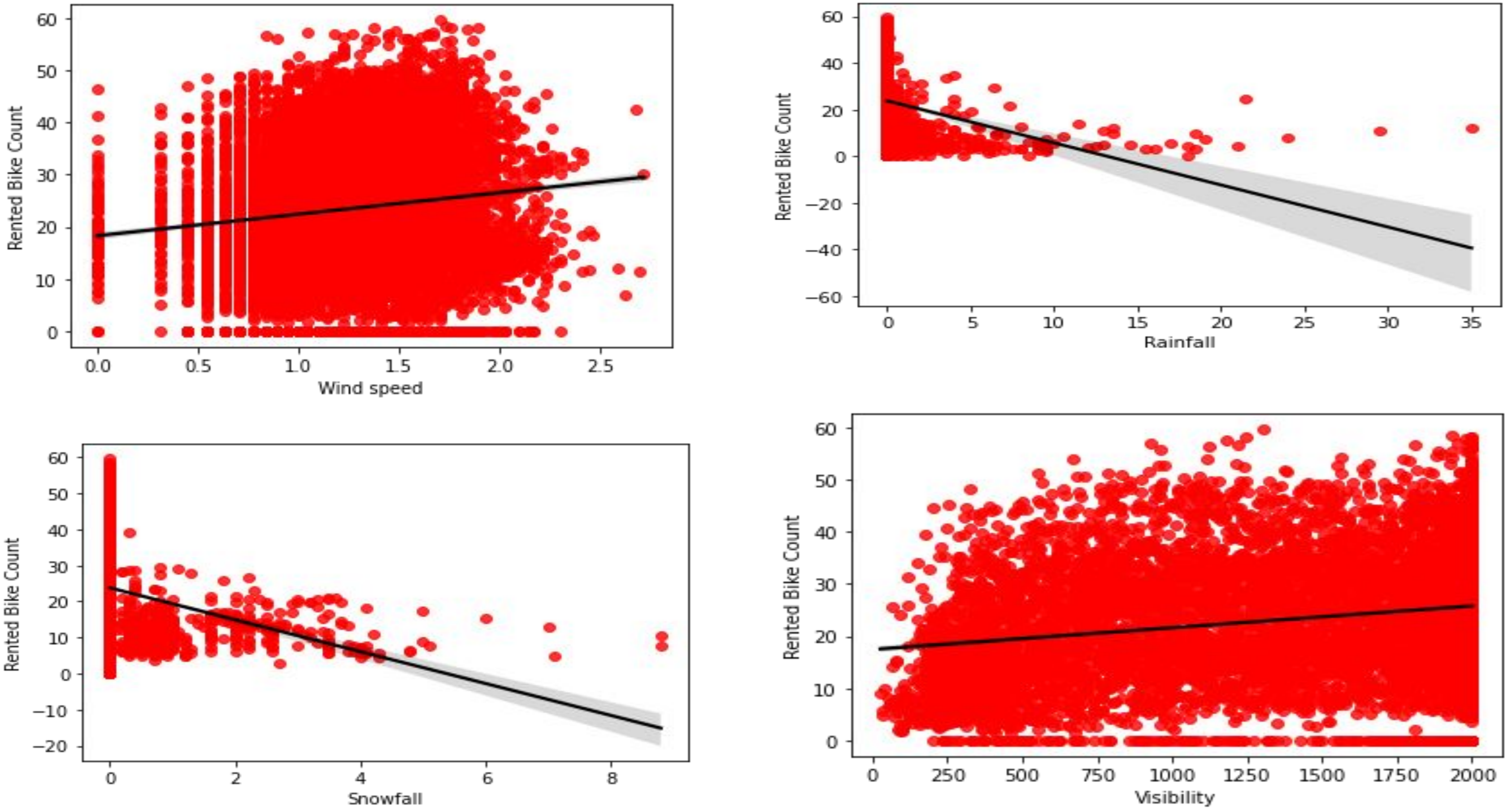


- We Can Clearly See The People Loves To Ride Bike During Summer Season And Autumn Season.
- But People Don’t Prefer Rented Bike In Winter Season Due To Snowfall.
- People Rented Bike In Non Holidays Most Compared To Holidays.
- Zero Bikes Were Rented On Non Functioning Day And 700 Bikes Were Rented On Functioning Day.

Regression Plot for Numerical Variable.

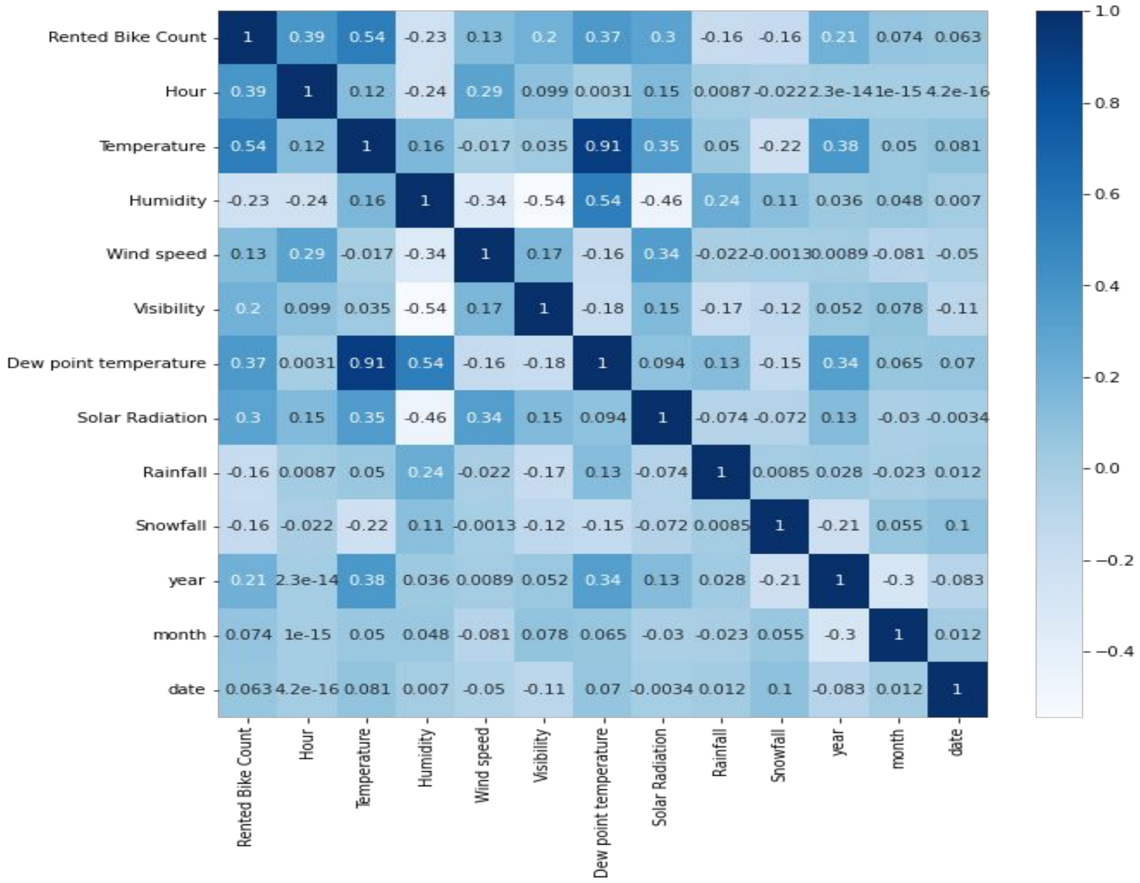


Regression Plot for Numerical Variable.



- From The Regression Plot We Can See That Numerical Features ‘temperature’, ‘dew Point Temperature’ ,’solar Radiation’ ,’ Wind Speed ‘, ‘visibility’ Have Positive Relation With The ‘rented Bike Count’.
- Which Means ‘rented Bike Count’ Increases With These Features.
- Also We Can See That ‘humidity’ ,’snowfall’, ‘rainfall’ Features Have Negative Relation With The ‘rented Bike Count’ .
- Which Means ‘rented Bike Count ‘ Decreases With These Features.

Correlation Heatmap.



- From The Heatmap We Can See That ‘dew Point Temperature’ Is Highly Correlated With ‘temperature’.

- So We Dropped The ‘dew Point Temperature ‘ Column Because It Has Less Correlation With Target Variable Compared To ‘temperature’ Variable to solve the multicollinearity problem.

Model Selection .

- Linear Regression.
- Lasso Regression.
- Ridge Regression.
- Decision Tree Regression.
- Random Forest Regression.

Linear Regression

MSE :
36.18670436018153

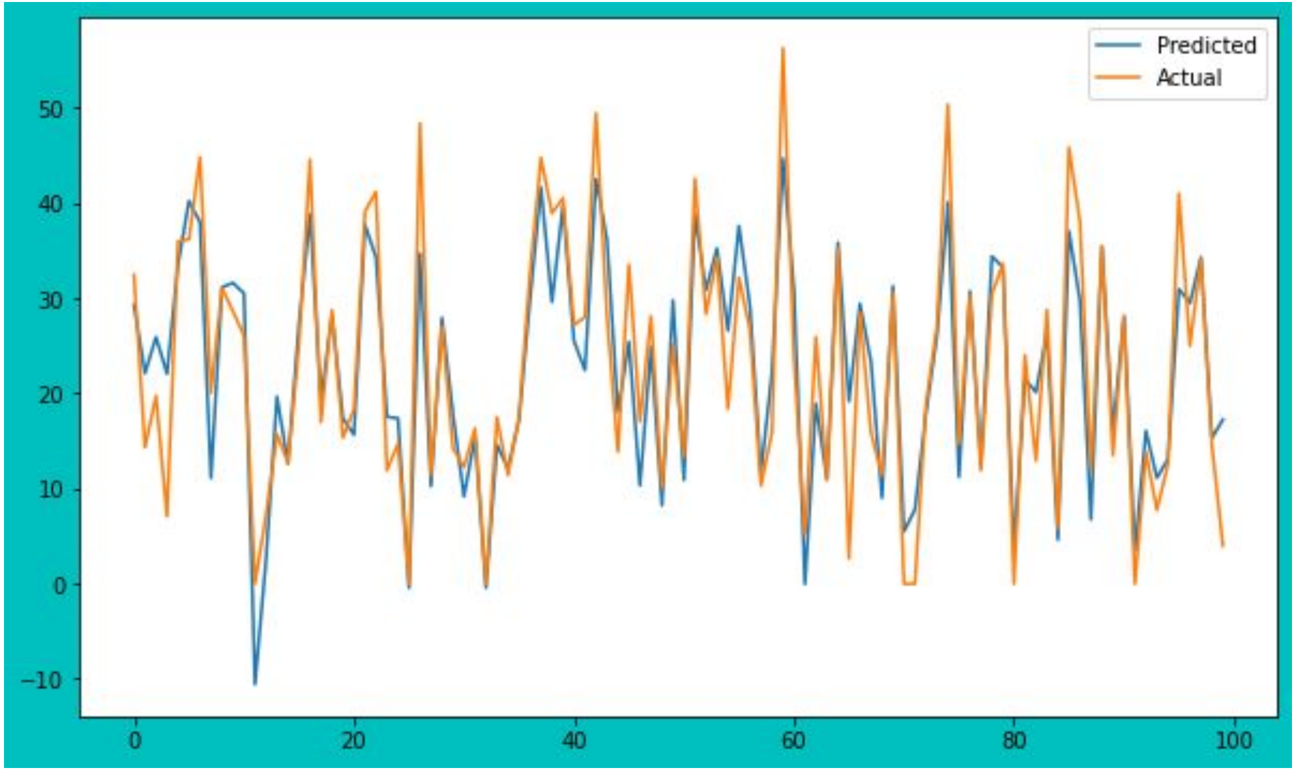
RMSE :
6.015538576069605

MAE :
4.598304269592001

R2 :
0.7715504021535617

Adjusted R2 :

0.7632215105654103



Lasso Regression

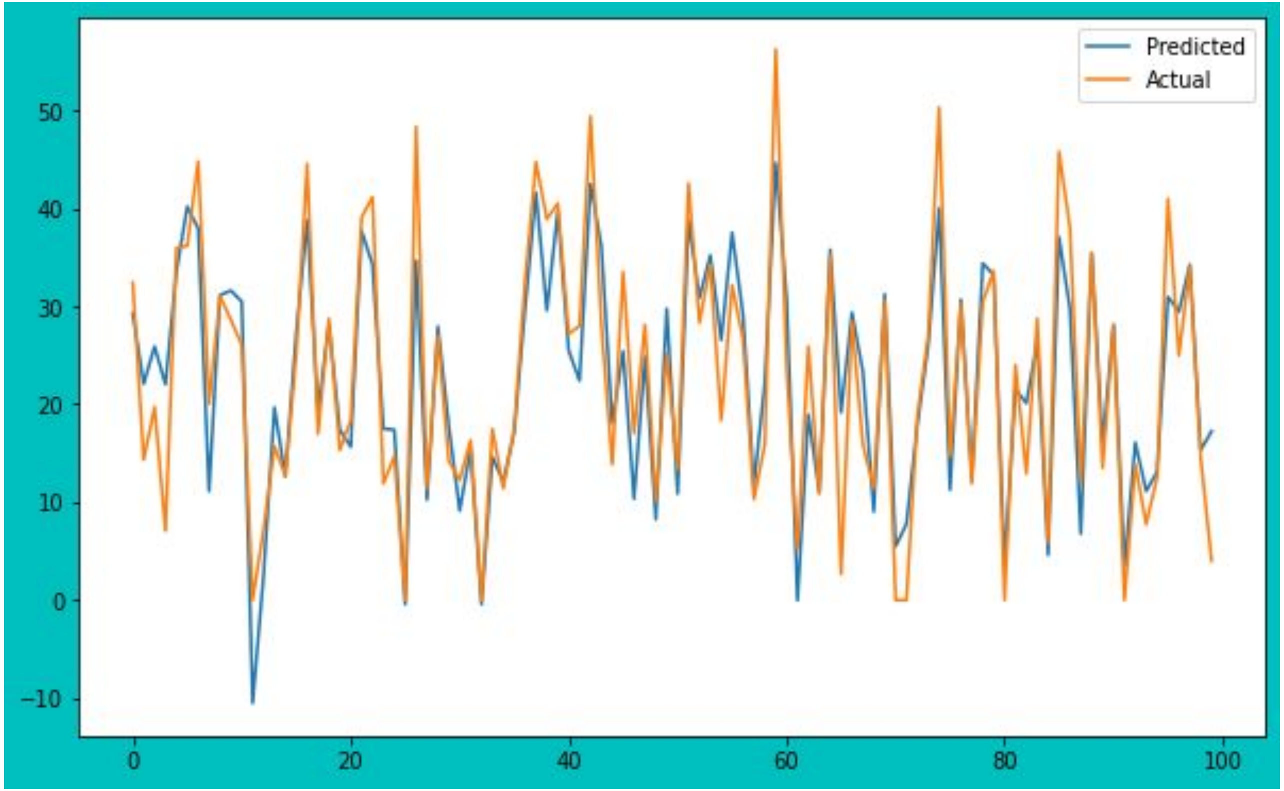
MSE :
36.18670436018153

RMSE :
6.015538576069605

MAE :
4.598304269592001

R2 : 0.7715504021535617

Adjusted R2
:0.7632215105654103



Ridge Regression

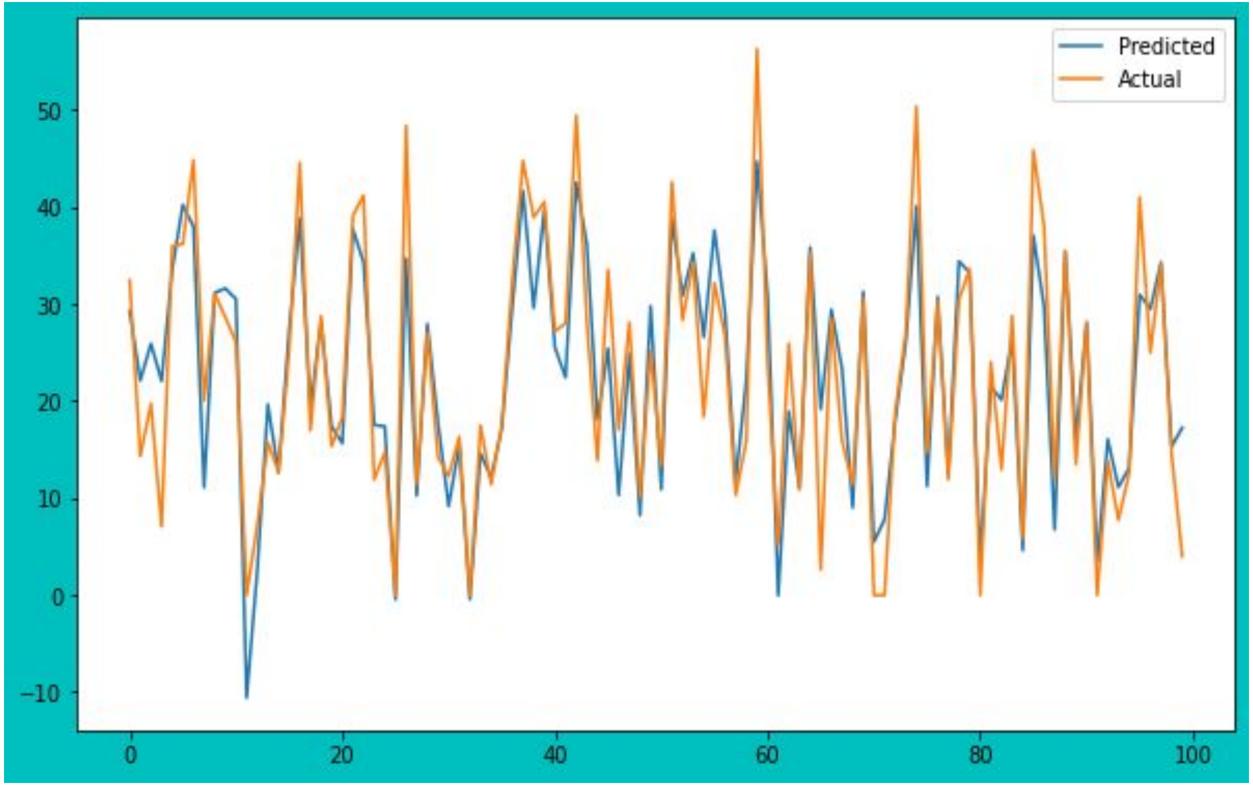
MSE :
36.186847345062894

RMSE :
6.015550460686279

MAE :
4.5983151961113915

R2 : 0.7715494994784157

Adjusted R2 :
0.763220574980233



Decision Tree Regressor

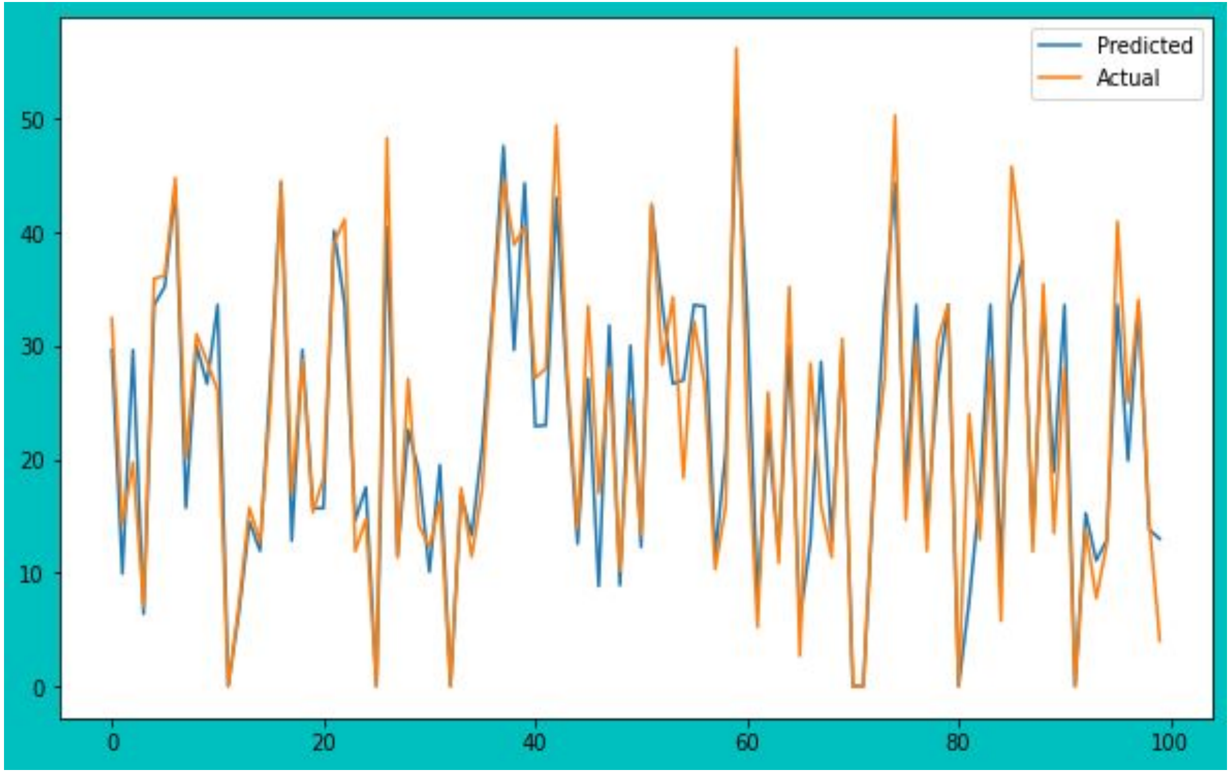
MSE :
26.523932513742043

RMSE :
5.150139077126174

MAE :
3.5211534750607054

R2 :
0.8325522640647556

Adjusted R2 :
0.82644739869



Random Forest Regressor.

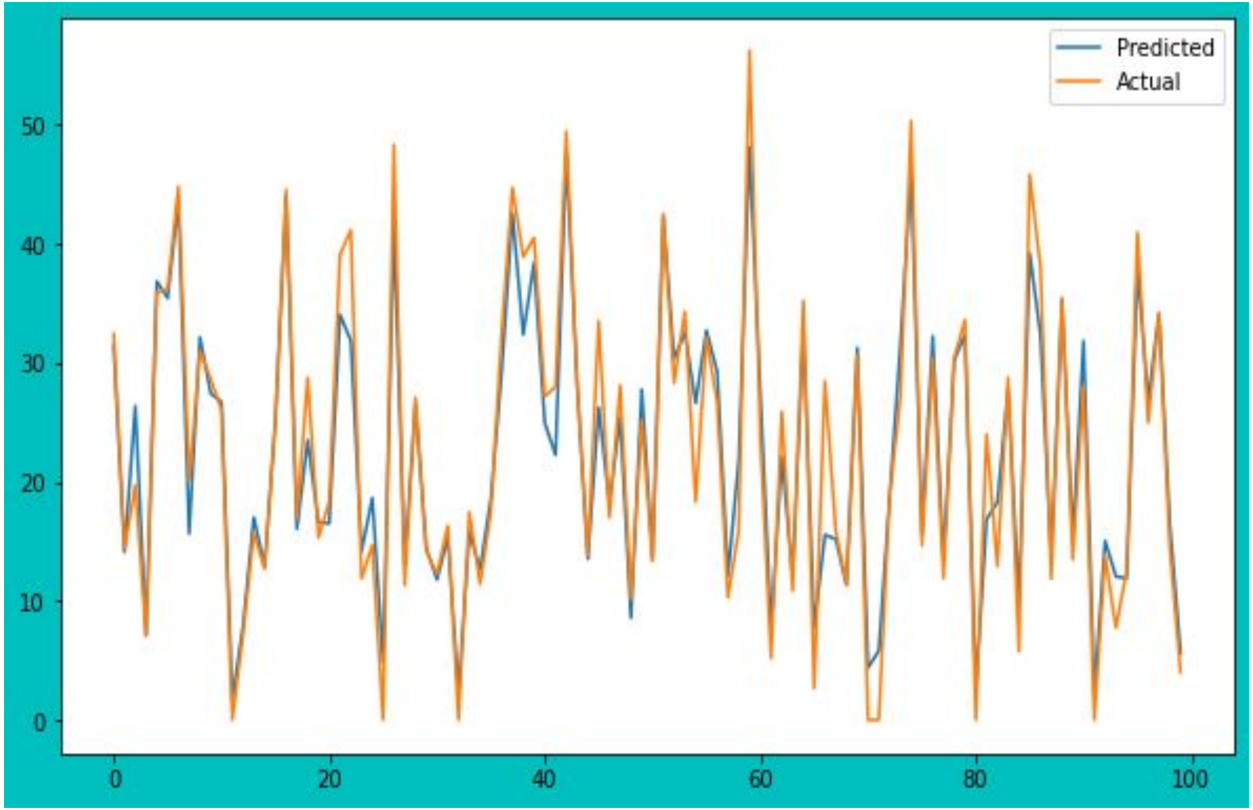
MSE :
15.861890575399553

RMSE :
3.982698905942998

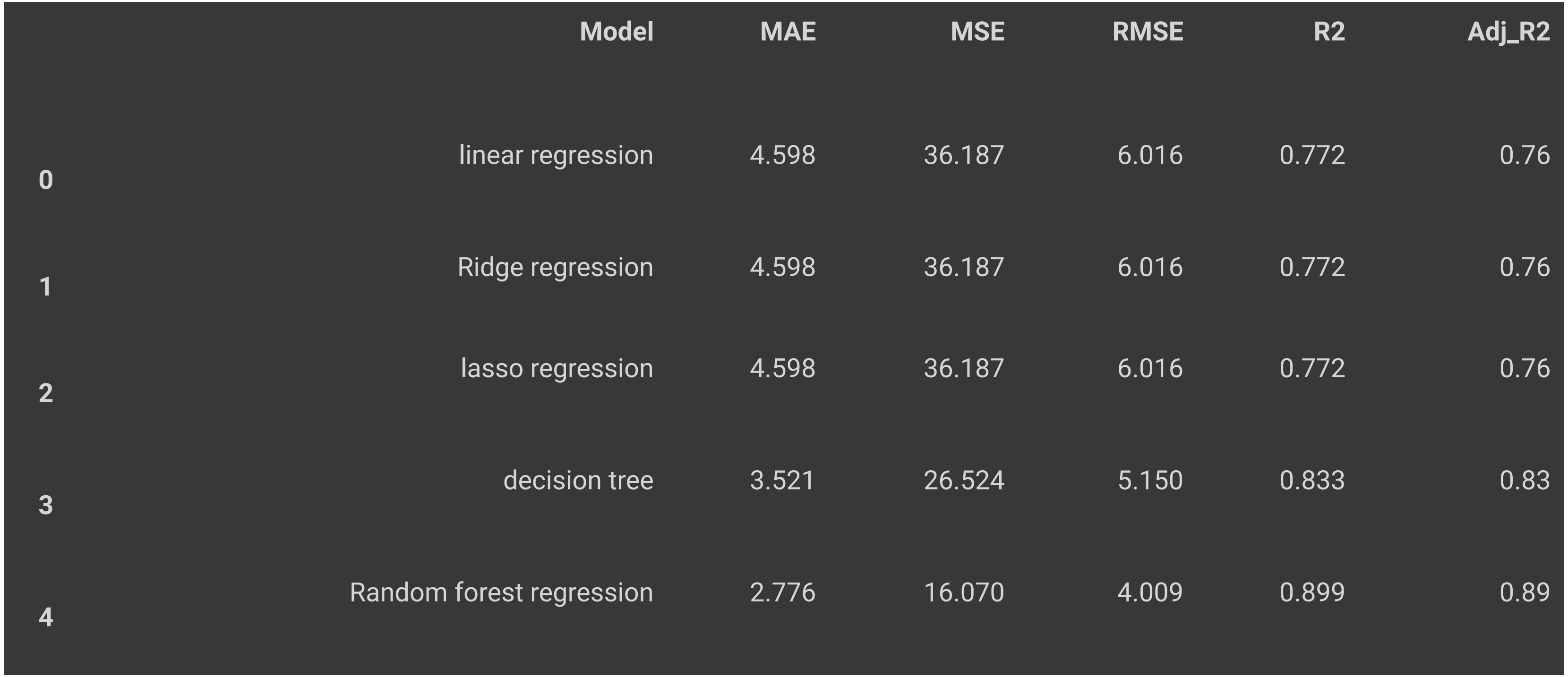
MAE :
2.654448482430134

R2 :
0.8998625990649329

Adjusted R2 :
0.8962117563225086



Comparison Chart.



	Model	MAE	MSE	RMSE	R2	Adj_R2
0	linear regression	4.598	36.187	6.016	0.772	0.76
1	Ridge regression	4.598	36.187	6.016	0.772	0.76
2	lasso regression	4.598	36.187	6.016	0.772	0.76
3	decision tree	3.521	26.524	5.150	0.833	0.83
4	Random forest regression	2.776	16.070	4.009	0.899	0.89

Conclusion:

- Most number of bikes are rented in the Summer season and the lowest in the winter season.
- Over 96% of the bikes are rented on days that are considered as No Holiday.
- Most number of bikes are rented when there is no snowfall or rainfall.
- The highest number of bike rentals have been done in the 18th hour, i.e 6pm, and lowest in the 4th hour, i.e 4am.
- Most of the bike rentals have been made when there is high visibility.

Challenges:

- changing date column into datetime format .
- Encoding the categorical columns.
- Removing Multicollinearity from the dataset.
- Choosing Model explainability technique.

Conclusion:

- We Implemented Five Machine Learning Algorithms Linear Regression,Lasso Regression, Ridge Regression , Decision Tree Regressor , Random Forest Regressor , We Did Some Hyperparameter Tuning To Improve Our Model Performance.

- Random Forest Regressor Is The Best Model That Can Be Used For The Bike Sharing Demand Prediction Since The Performance Matrices Shows Lower Value Of (RMSE,MSE) And Higher Value Of R Square ,Adjusted R Square For Random Forest Regressor.

Thank you