Capstone Project-2

Bike Sharing Demand Prediction

By- Shubham Srivastava

Index

- Defining Problem Statement
- Dataset Summary
- EDA(Exploratory Data Analysis)
- Evaluation using various Regression Models
- Comparing Evaluation Metrices of all Models
- Conclusion

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.

Dataset Summary

- Bike sharing has been gaining a lot of importance over the last few decades. More people are paying attention or even turning to cities where activities like bike sharing are easily available. There are uncountable number of benefits to using bike sharing systems in cities. We can even say that it is a green way to travel.
- The given dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour for each date.

Dataset Summary

- This dataset contains the hourly and daily count of rental bikes between years 2017 and 2018 with corresponding weather and seasonal information. The dataset contains 8760 rows(every hour of each day for 2017 and 2018) and 14 columns (the features which are under consideration).

EDA(Exploratory Data Analysis)

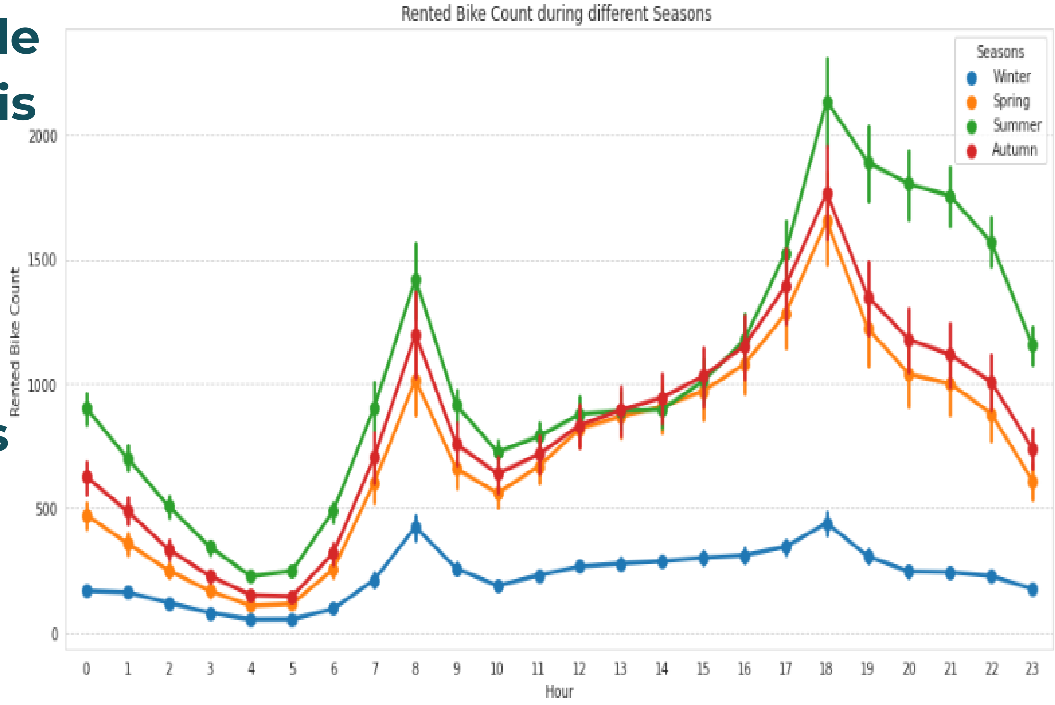
Rented Bike Count , Hour with respect to different categorical features

- **Season:**

In the season column, we are able to understand that the demand is low in the winter season.

- **Holiday:**

In the holiday column, the demand is low during holidays, but in no holidays the demand is high.



EDA(Exploratory Data Analysis)

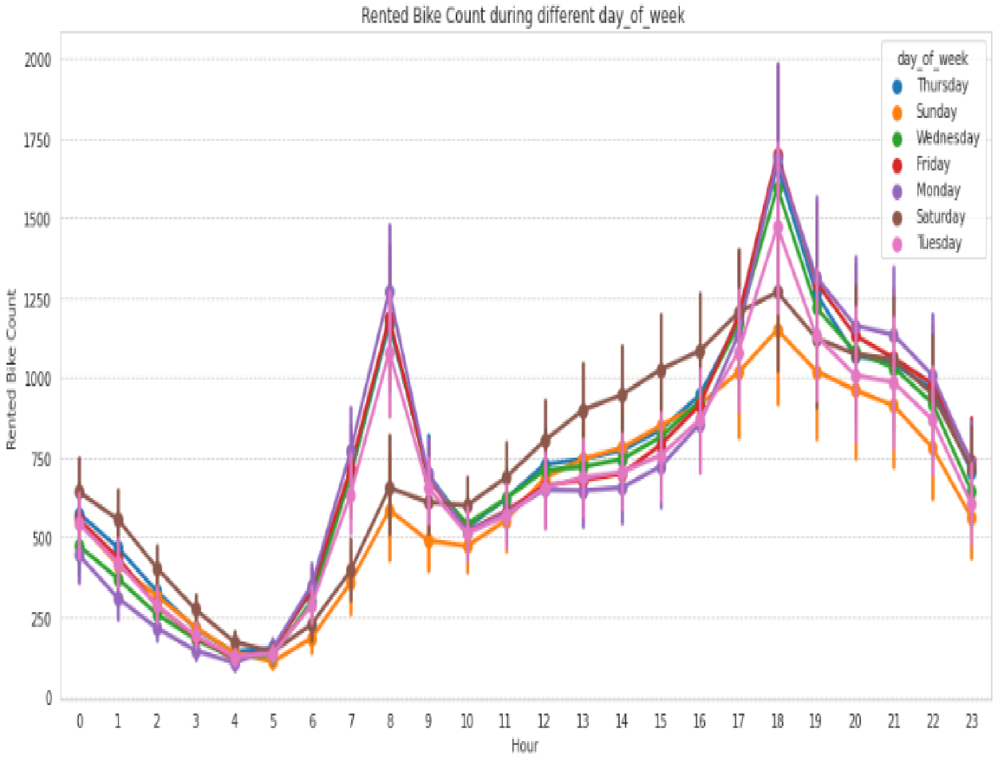
Rented Bike Count , Hour with respect to different categorical features

- **Functioning Day:**

In the functioning day column, if there is no functioning day then there is no demand.

- **Days of week:**

In the days of week column, we can observe from this column that the pattern of weekdays and weekends is different, in the weekend.



EDA(Exploratory Data Analysis)

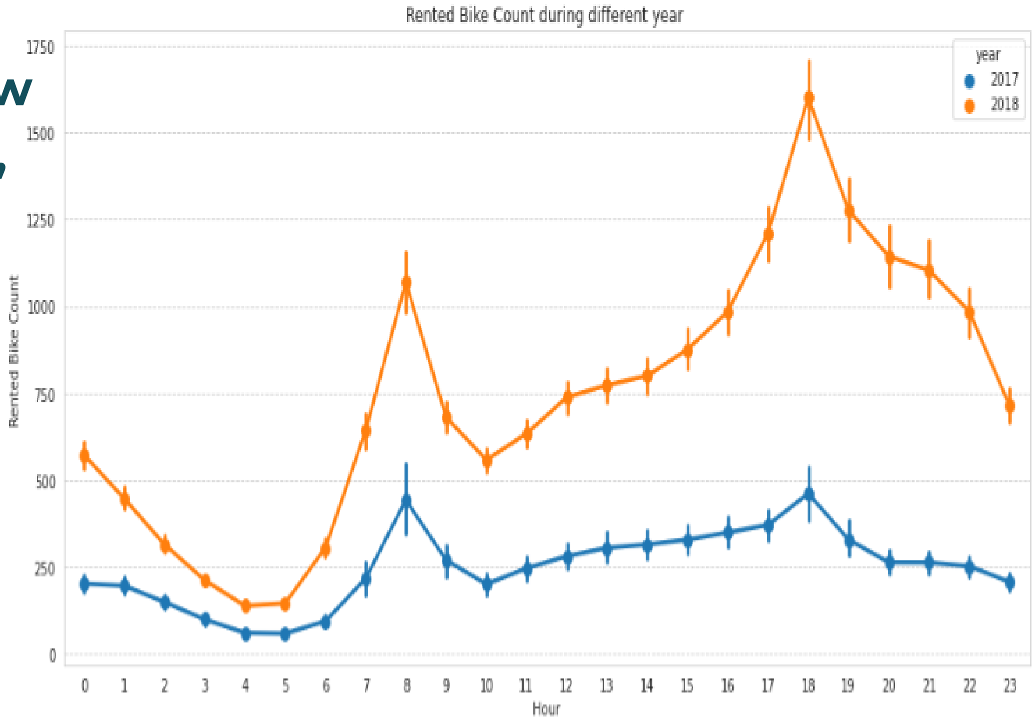
Rented Bike Count , Hour with respect to different categorical features

- **Month:**

In the month column, we can clearly see that the demand is low in December, January &February, it is cold in these months and we have already seen in season column that demand is less in winters.

- **Year:**

The demand was less in 2017 and higher in 2018.



EDA(Exploratory Data Analysis)

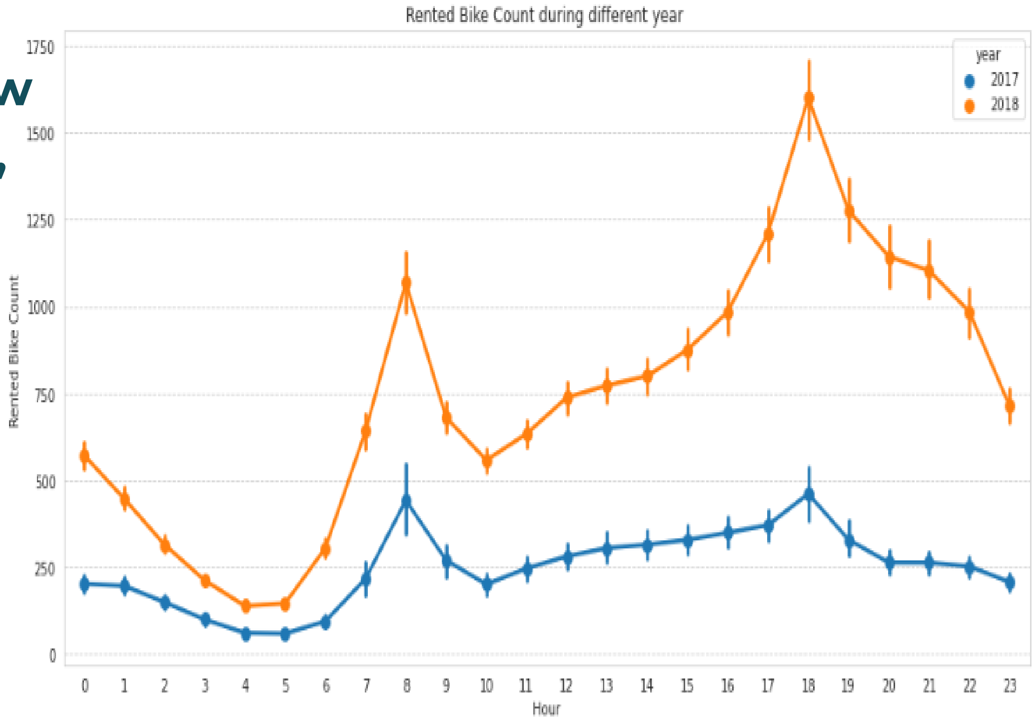
Rented Bike Count , Hour with respect to different categorical features

- **Month:**

In the month column, we can clearly see that the demand is low in December, January &February, it is cold in these months and we have already seen in season column that demand is less in winters.

- **Year:**

The demand was less in 2017 and higher in 2018.



EDA(Exploratory Data Analysis)

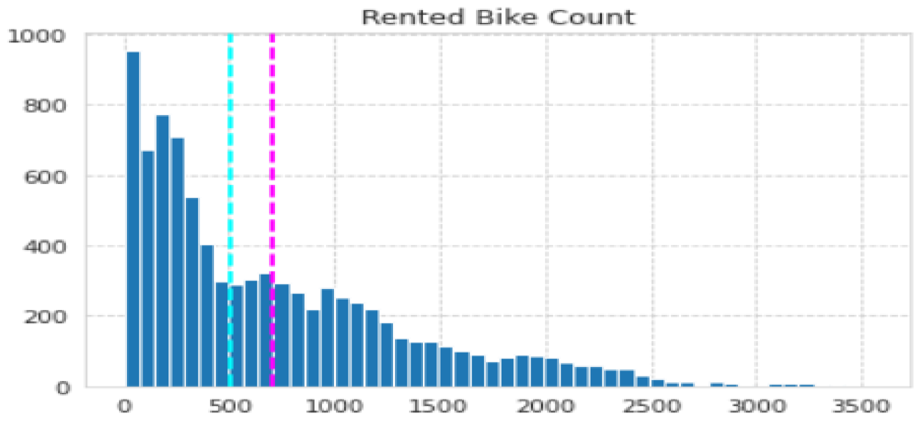
Distribution of Numerical features

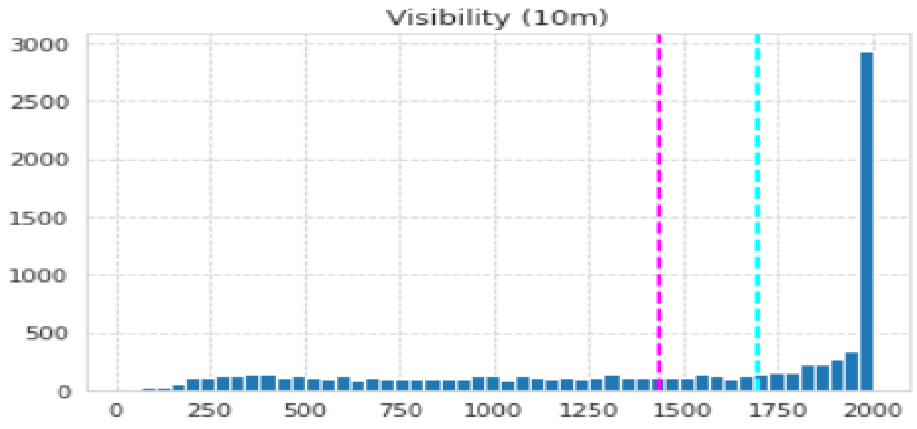
- **Right Skewed Columns:**

Rented Bike Count(Dependent variable), Wind Speed(m/s), Solar Radiation(MJ/m2), Rainfall(mm), Snowfall(cm).

- **Left Skewed Columns:**

Visibility(10m), dew point temperature(Celsius).

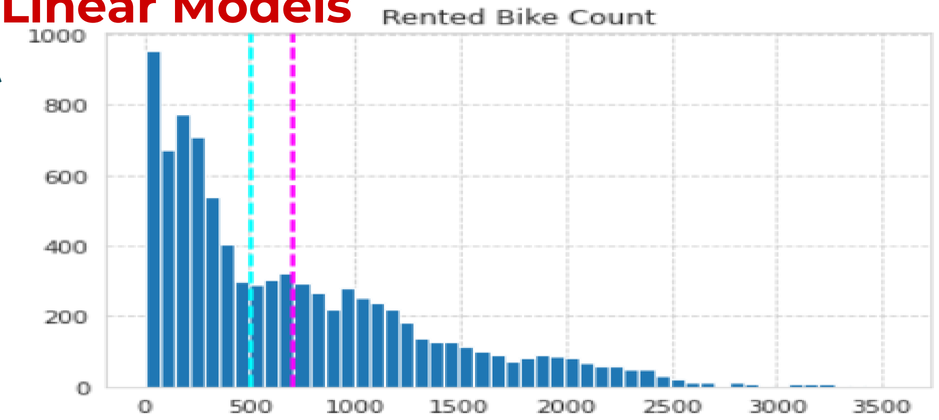


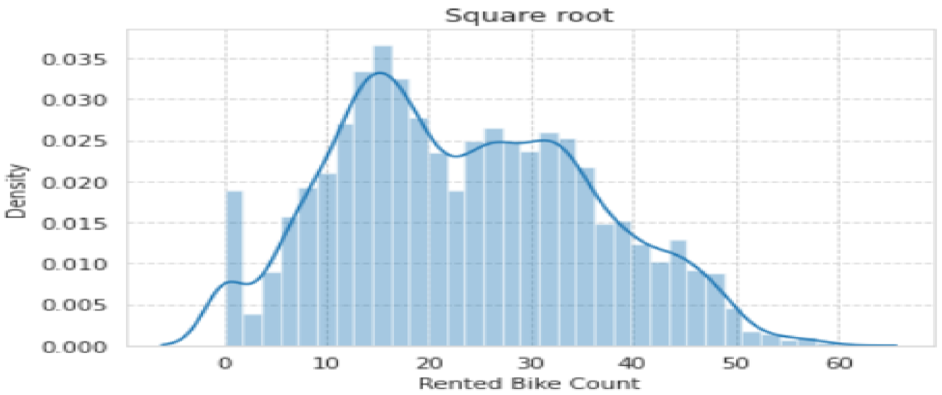


EDA(Exploratory Data Analysis)

Normalize Dependent variable for Linear Models

- Our Dependent variable is right skewed .
- To normalize our dependent variable we tried log10, square, square root of dependent variable.
- As we can see that square root was helpful in normalizing our Dependent variable. So, we will take square root of dependent variable.

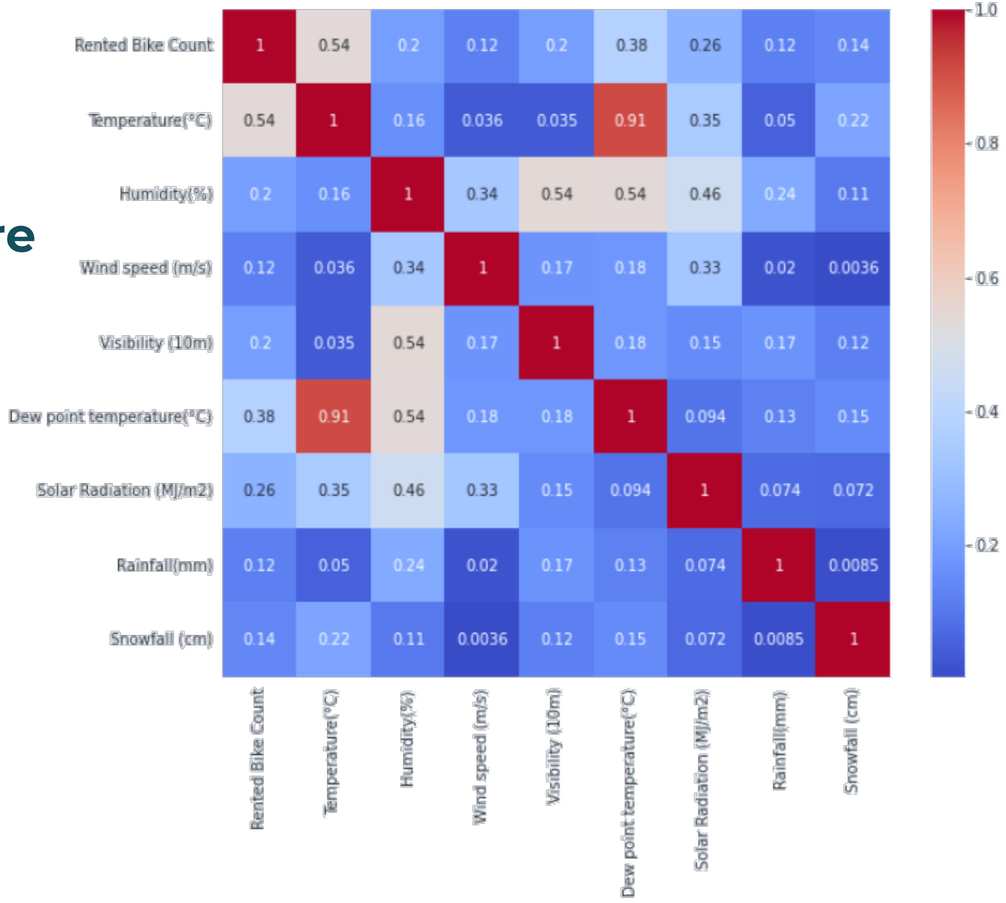




EDA(Exploratory Data Analysis)

Correlation Analysis

We can see in the Correlation graph that dew point temperature and temperature are highly correlated. Then we also found they affect VIF score. So, we will drop one feature which has least Correlation with Dependent variable. Therefore, we dropped Dew point temperature.



List of Models Performed

- Linear Regression
- Lasso Regression
- Lasso Regression using Cross Validation
- Ridge Regression
- Polynomial Regression
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosting Regression
- XGBoost
- LightGBM

All Models Evaluation Metrices

	Models_Name	Mean Absolute Error(MAE)	Mean Squared Error(MSE)	Root Mean Squared Error(MSE)	R2 Score	Adjusted R2 Score
0	LinearRegression	4.617596	38.727948	6.223178	0.744109	0.738372
1	LassoRegression	4.617071	38.722468	6.222738	0.744145	0.738409
2	LassoRegression(cv)	4.614588	38.692002	6.220290	0.744347	0.738615
3	RidgeRegression	4.617492	38.710778	6.221799	0.744223	0.738488
4	PolynomialRegression	3.092010	45.707569	6.760737	0.697992	0.691221
5	Decision_Tree	174.881735	90626.733790	301.042744	0.778528	0.773456
6	Random_Forest	135.311406	49150.736511	221.699654	0.879886	0.877136
7	Gradient_Boosting	190.688314	75729.050757	275.189118	0.814934	0.810697
8	XGBoost	174.255238	60439.712500	245.844895	0.852298	0.848916
9	lightGBM	127.612972	42131.042795	205.258478	0.897041	0.894683

Top 3 best performing models: 1. LightGBM
2. Random Forest
3. XGBoost

Conclusion

1. In holiday or non-working days there is high demands for bike.
2. People preferred more rented bikes in the morning than the evening.
3. When the rainfall was less, people have booked more bikes.
4. The Temperature, Hour & Humidity are the most important features that positively drive the total rented bikes count.
5. After performing the various models the lightGBM 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 lightGBM! We can use either lightGBM for the bike rental stations.