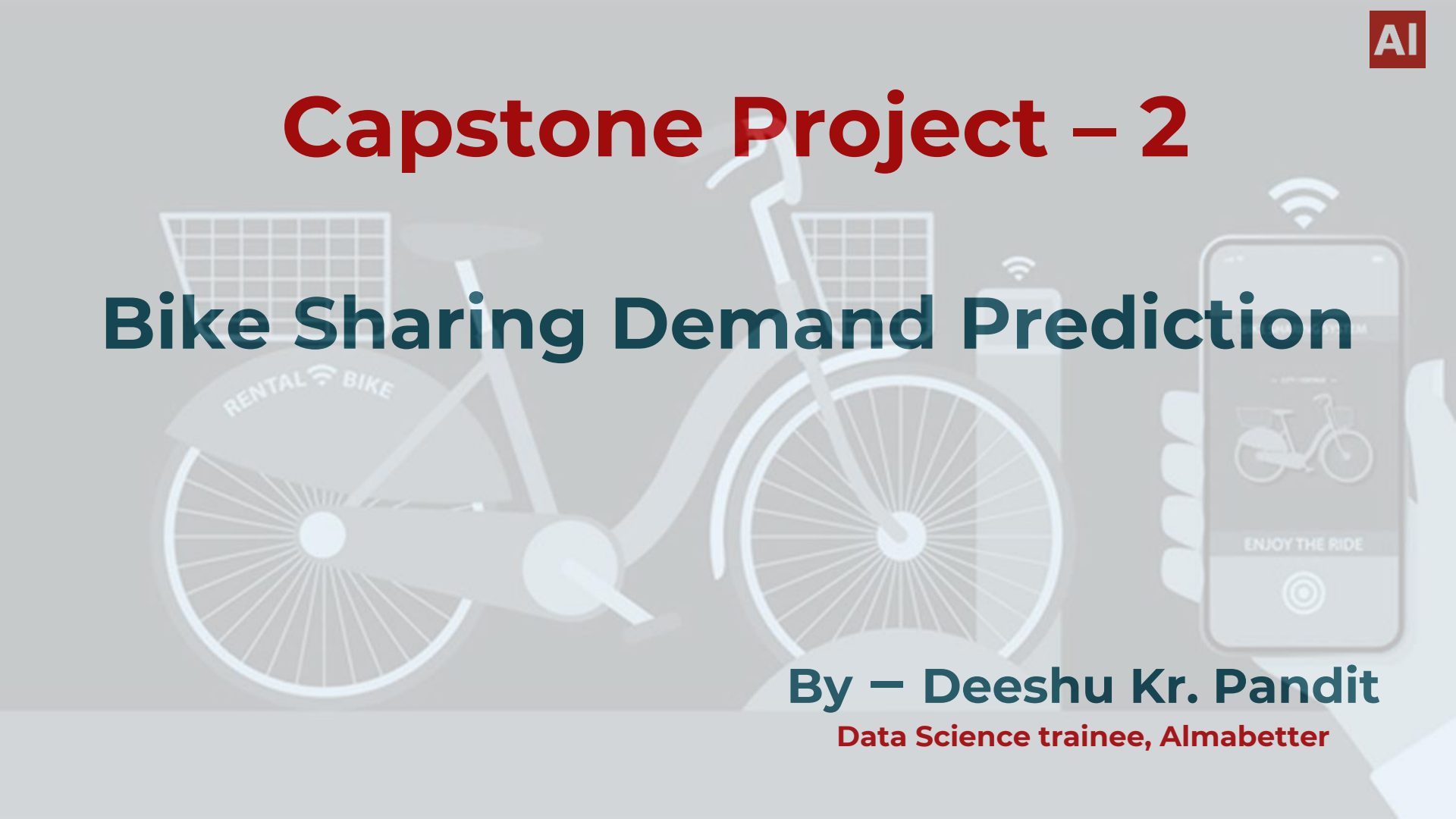
Capstone Project – 2

Bike Sharing Demand Prediction



By – Deeshu Kr. Pandit

Data Science trainee, Almabetter

Flow of the Project

1. Problem Description
2. Data Description
3. Exploratory data analysis
4. Data Preprocessing
5. Model Selection
6. Model Implementation
7. Model Explainability
8. Summary and conclusions

1. 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.

2. Data Description

- **Target Variable** :
  - Rented Bike count - Count of bikes rented at each hour
- **Independent Variables:**
  - Date : year-month-day
  - Hour - Hour of the 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)

Data Description (Cont..)

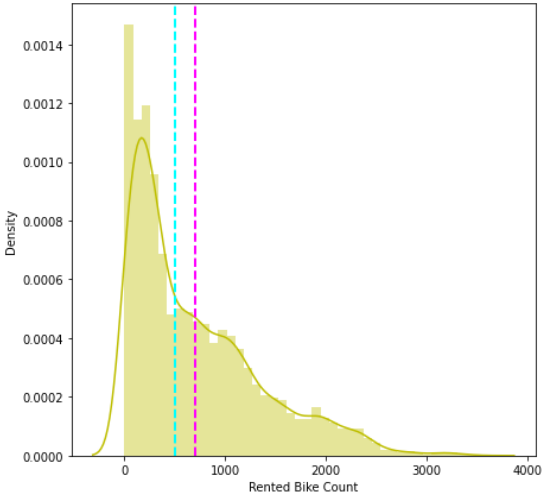
- The shape of the data is found to be (8760,4), that means a total of 8760 entries of data with 14 attributes.
- There are no duplicates neither any null values present in the data.
- The duration of the data collected is-

  `Start date: 2017-01-12 , End date: 2018-12-11`
- The data contains both numerical and categorical attributes with datatypes as integer, float and object.
- Converted date column to Datetime format which was earlier object type.
- Created three new features – WeekDay, Month and Year from date feature.
- Dropped date column.

3. Exploratory Data Analysis (EDA)

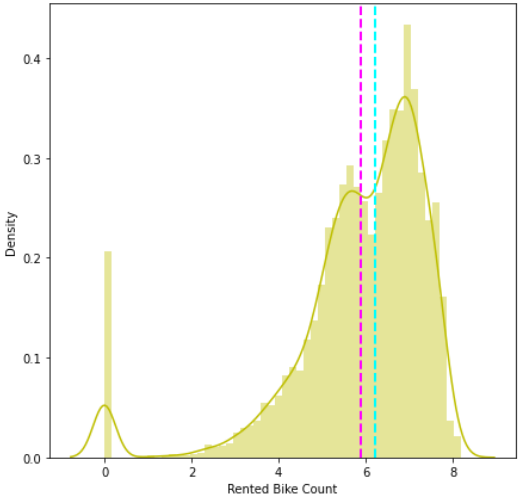
- **Distribution of Dependent variable:**

Original



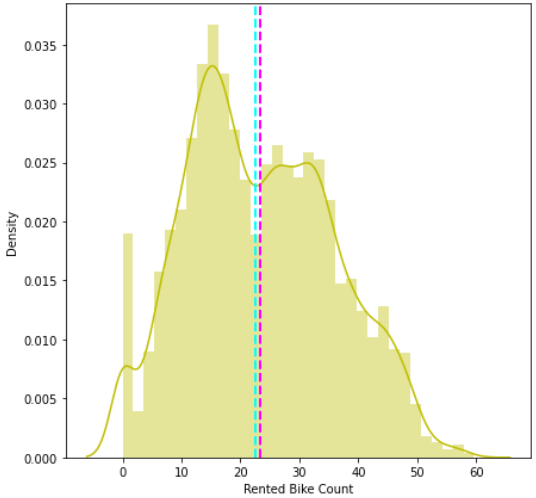
Skewness value: 1.153

Log Transformed



Skewness value: -1.832

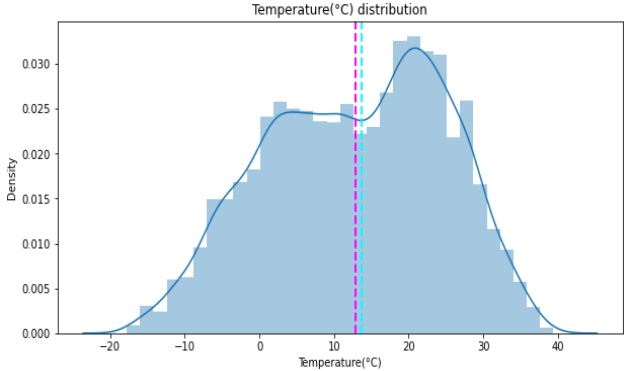
Sq. root transformed

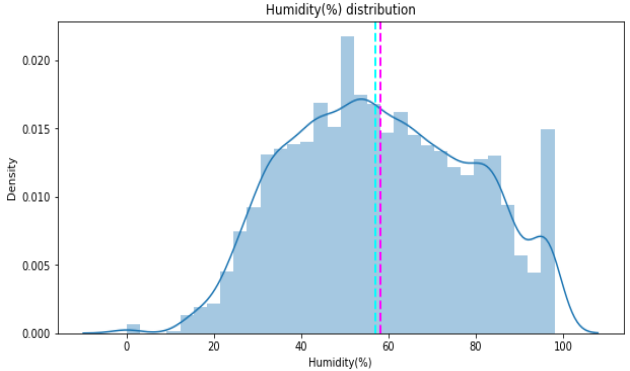


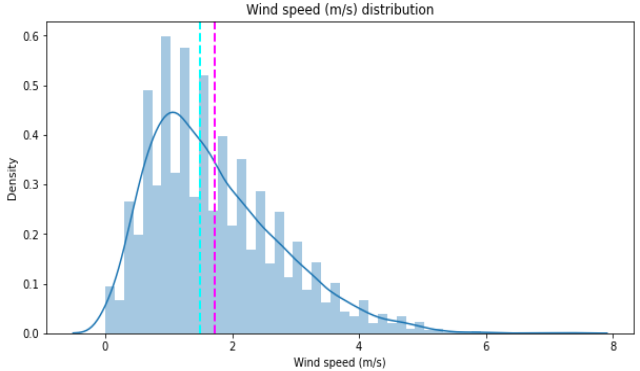
Skewness value: 0.237

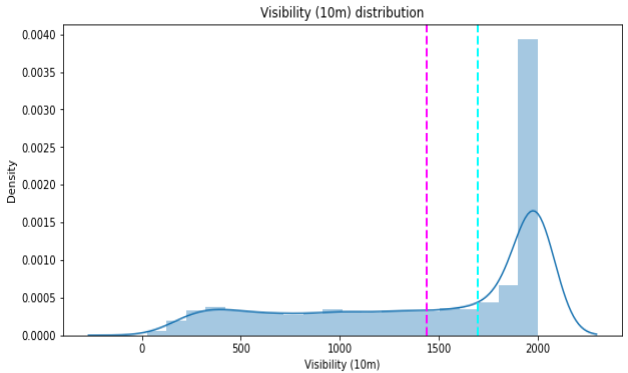
EDA (Cont..)

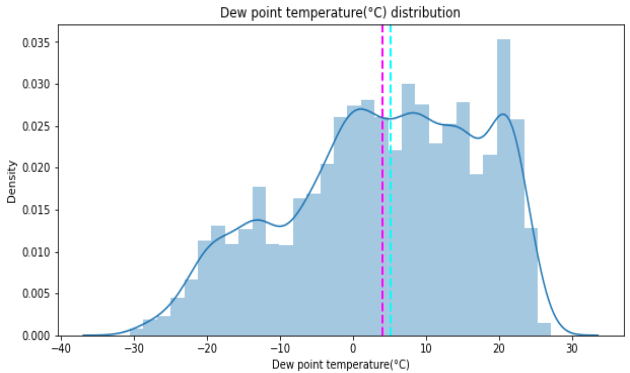
- **Distribution of Independent variables(continuous):**

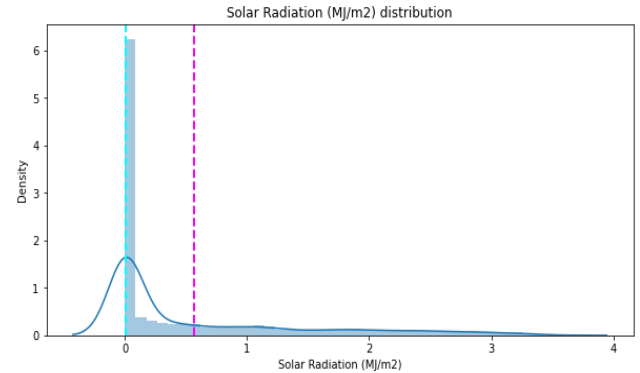






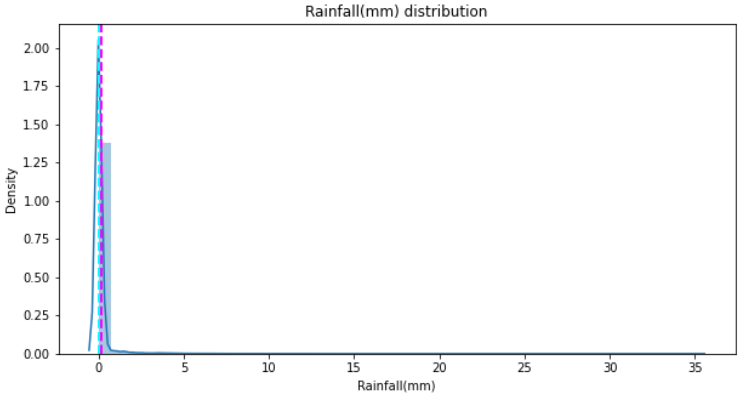


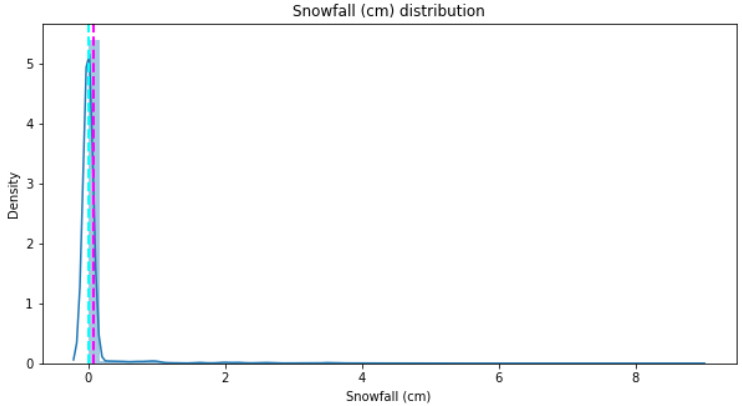




EDA (Cont..)

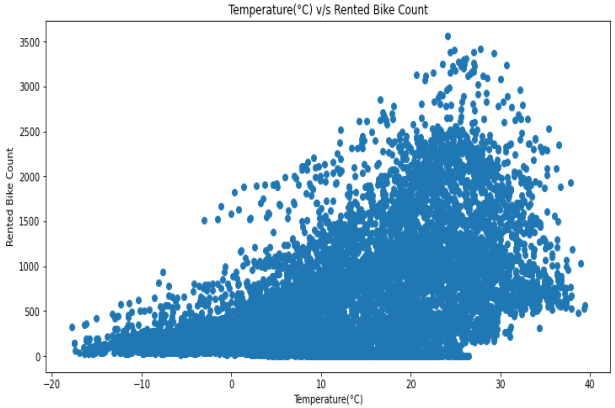
- **Distribution of Independent variables(continuous):**

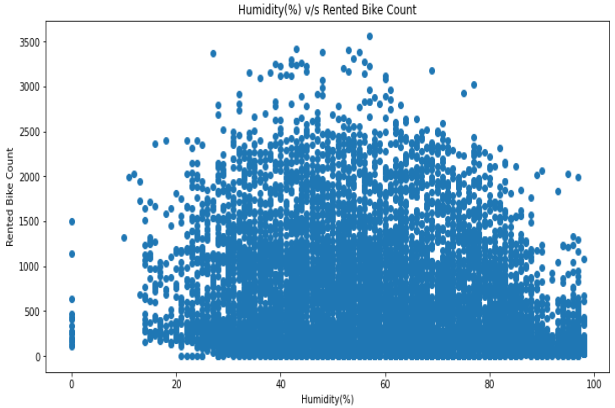


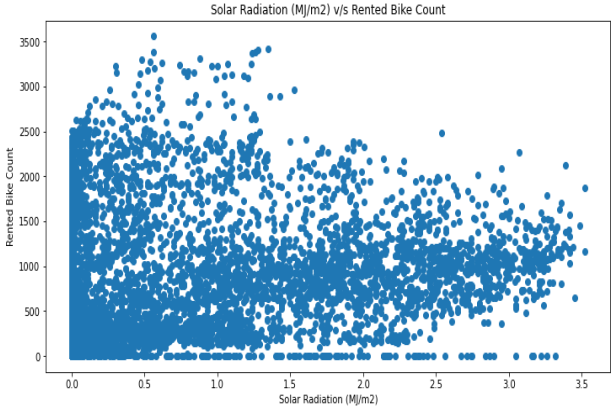


- **Summary:**
  - **Normally distributed variables:** Temperature(°C),Dew Point Temperature(°C), Humidity(%)
  - **Positively skewed variables:** Wind speed, Solar Radiation, Snowfall, Rainfall.
  - **Negatively skewed variables:** Visibility.

EDA (Cont..)

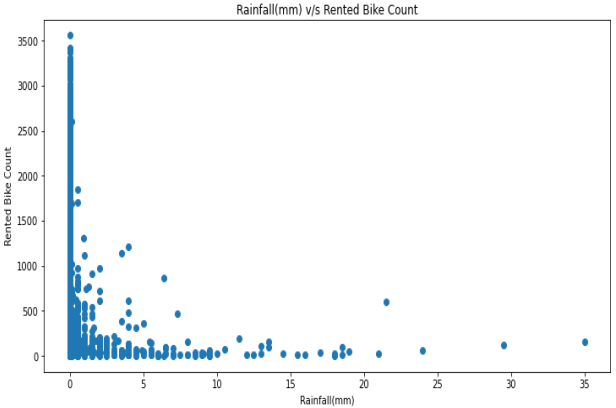




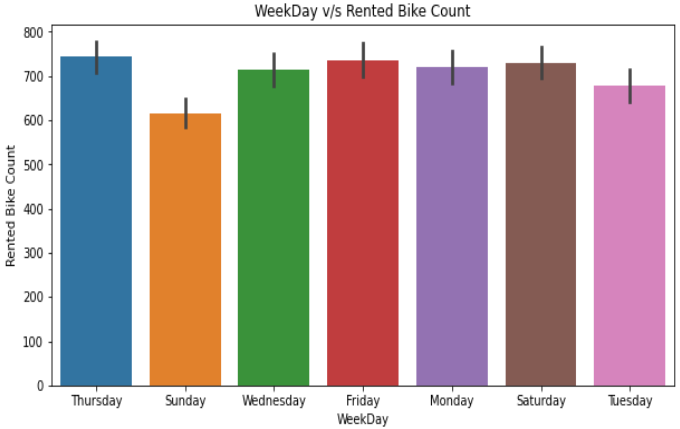


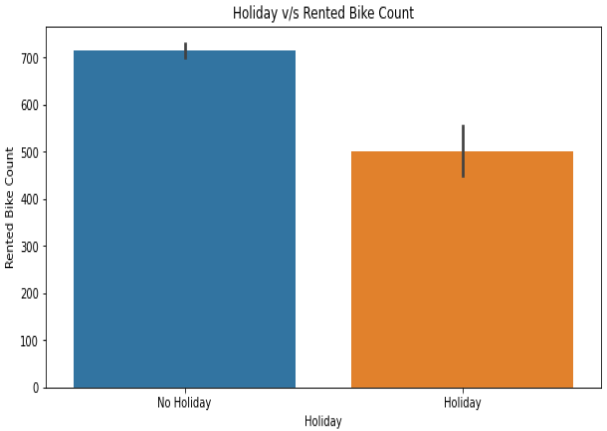
- The demand for rental bike is **high** as the **temperature** increases while it is **low** with the increase of **solar radiation**.

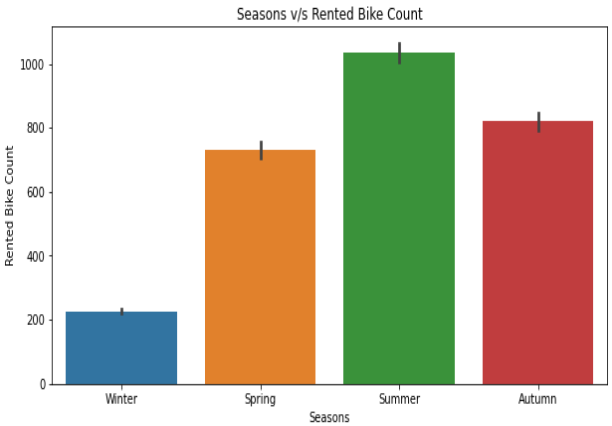
- The demand for rental bike is very **low** when there is a significant **Rainfall**.



EDA (Cont..)





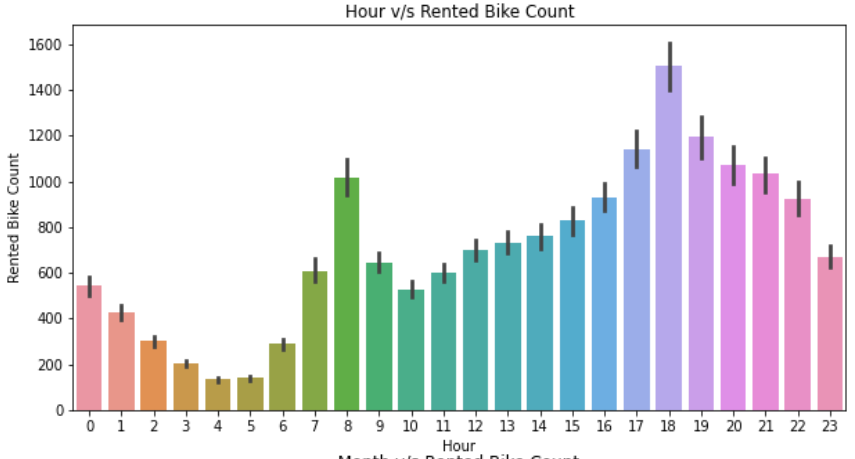


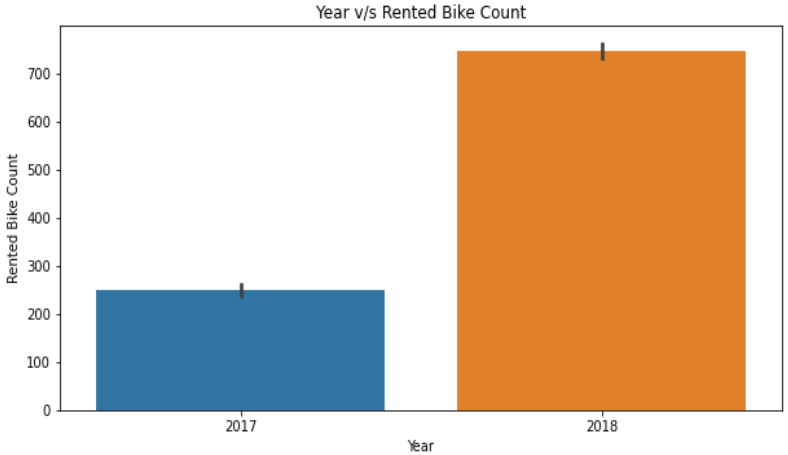
- The number of bikes rented based on weekday is hard to understand but it is for sure that it is least on Sunday.
- The number of bikes rented is higher on working days than on holidays, which indicates that most of the users are employees or students.
- People prefer rental bikes most in Summer season and least in Winter.

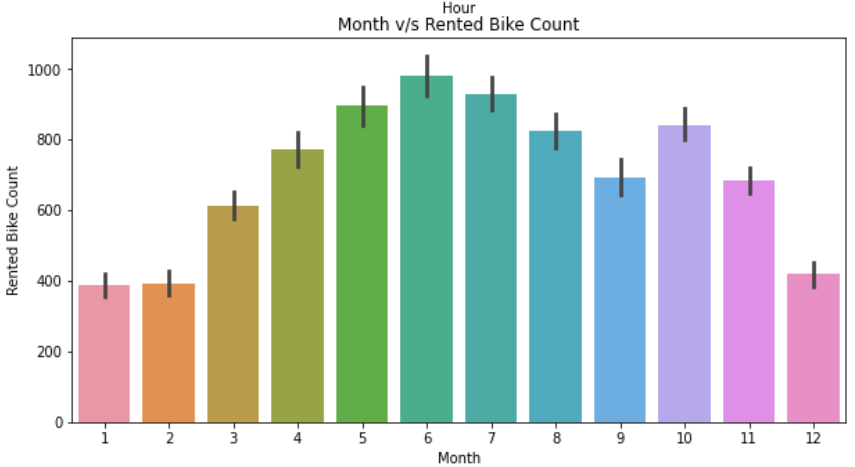
EDA (Cont..)

- The number of bikes rented is **higher** during the **rush hours** - which might be due to the offices, schools and colleges timings.

- The demand is **highest** in the months around **May, June and July** followed by **Oct. and Nov**.
- As compared to 2017, the demand is much higher for the year 2018.

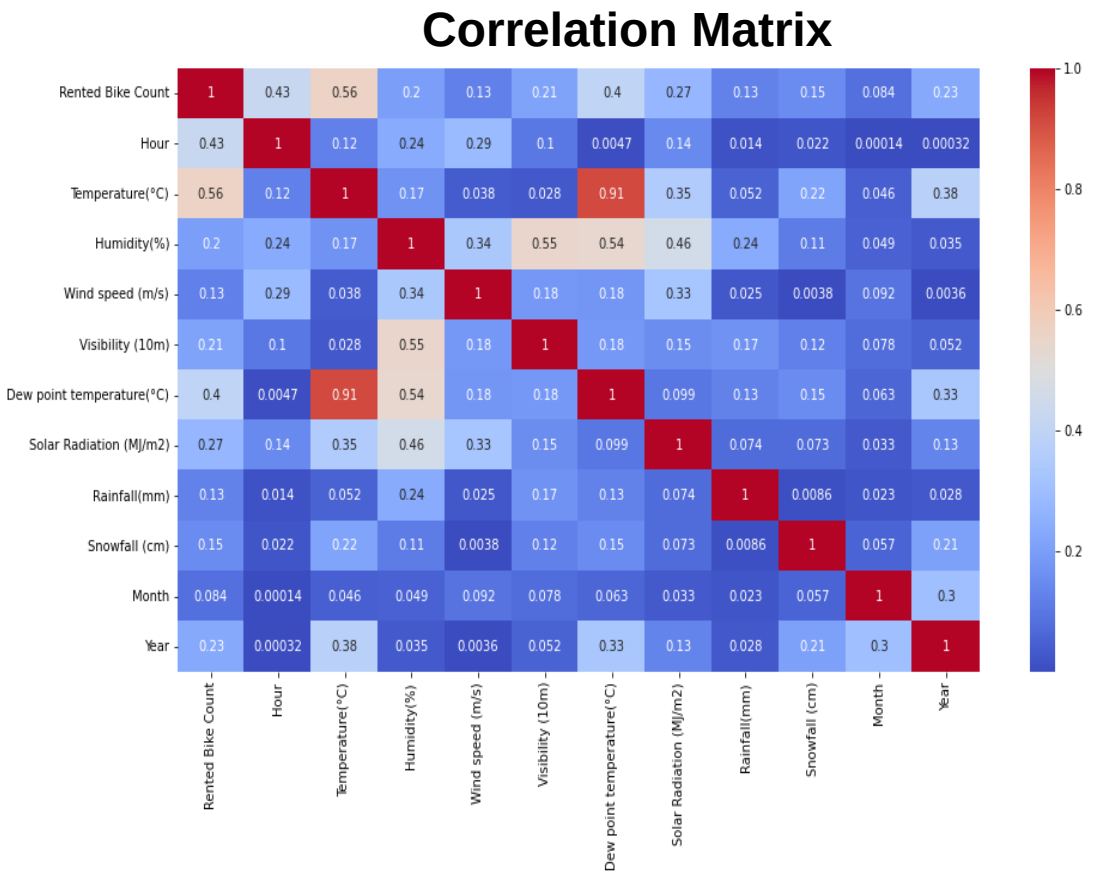






EDA (Cont..)

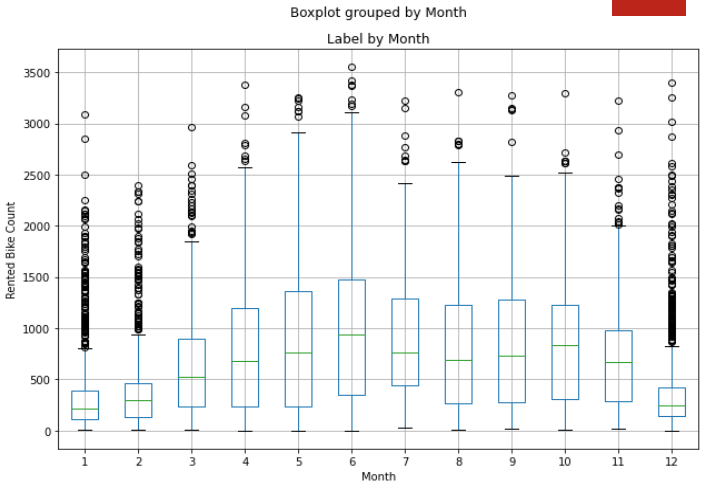
- The only **significant correlation** can be observed between **Temperature** and **Dew point temperature**, which is obvious as there is a relation exist between them.
- The **highest correlated feature** with the target variable is the **Temperature** followed by **Hour**.

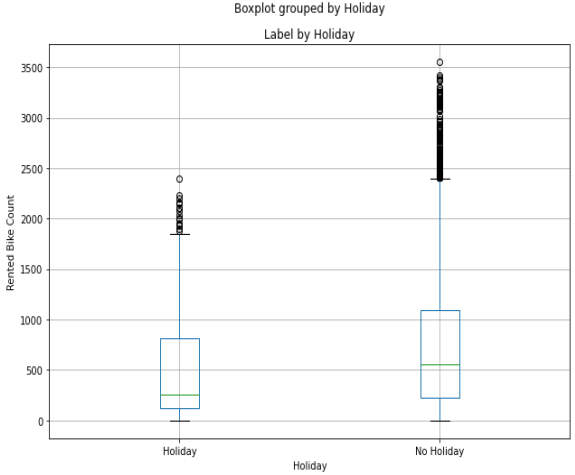


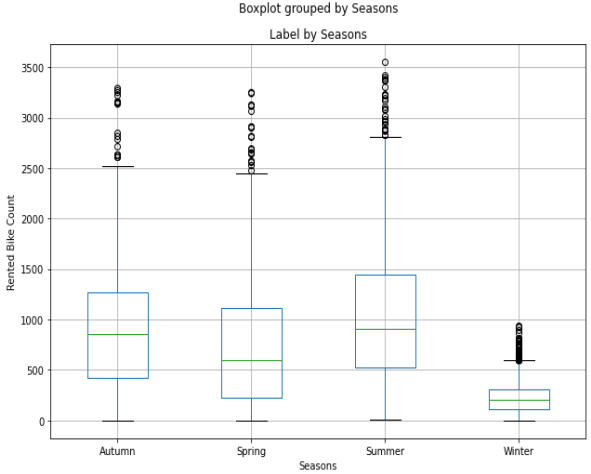
EDA (Cont..)

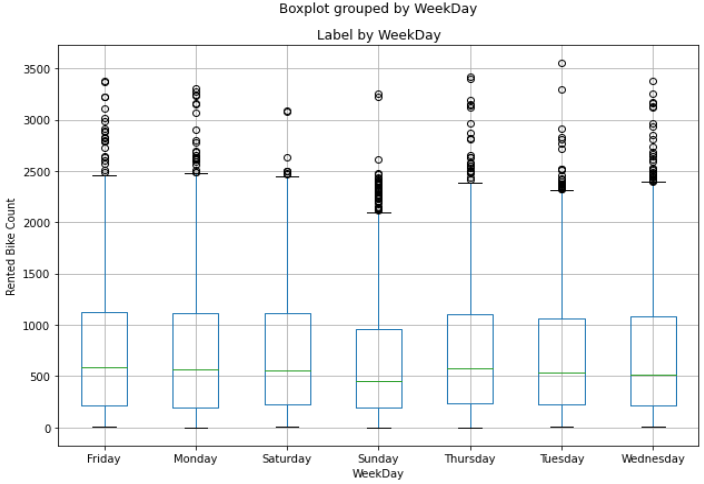
- The different box plots here shows the **range of values** present for our dataset along with the **outliers**.

- It can be easily observed that our dataset has a lot of outliers present. Since removing a lot of outliers, we may loose our most of the data and the patterns we have found. So, we will go with same data.







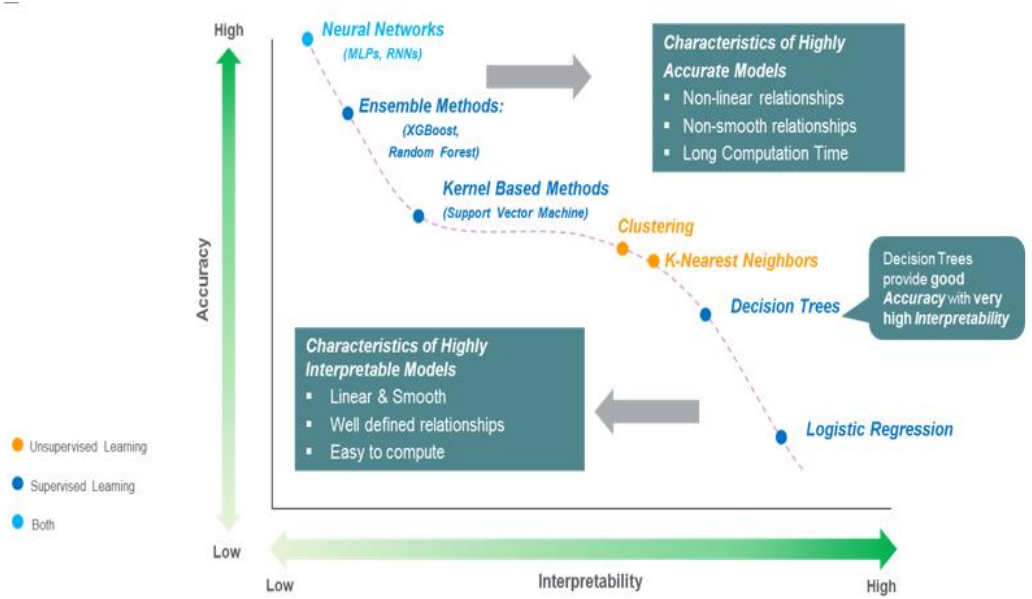


Model Selection:

- Our task is to find the count of rental bikes required per hour, which can be predicted using any regression model.

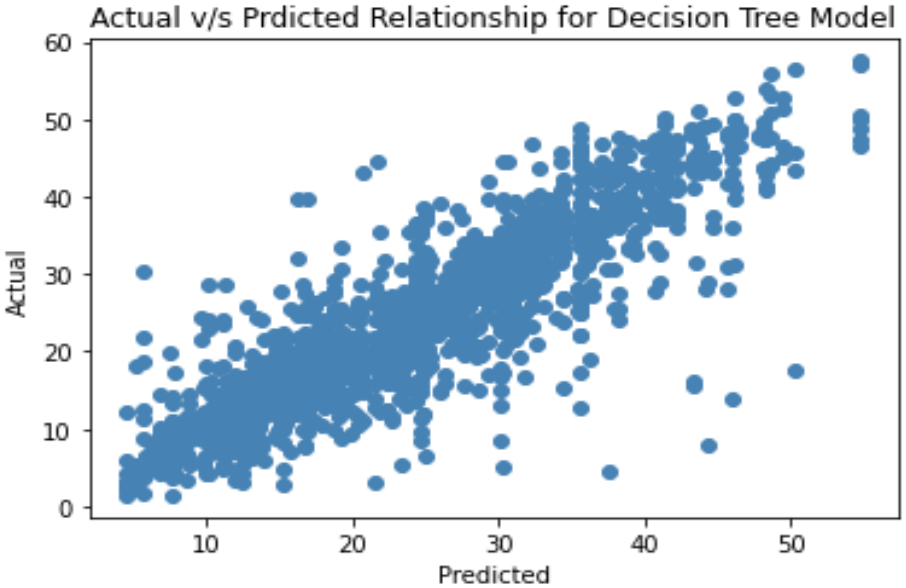
- Since our data contains many outliers, linear regression is not a good choice. Hence, we will go with regression models based on decision tree:
  - **Decision Tree Regressor**
  - **Random Forest Regressor**
  - **Gradient Boosting Regressor**
  - **XG Boost Regressor**

- While selecting the best model, the demand from **stakeholders** should be kept in mind because at the cost of getting high **accuracy** we loose the **model explainability**.
- **Evaluation metrics** used – **r2_score** and **RMSE**(Root mean square error)



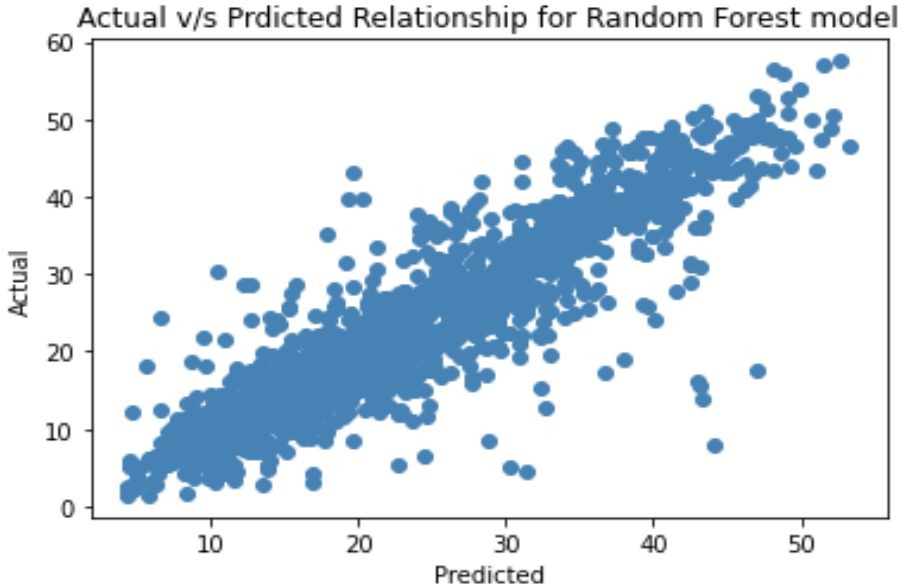
Model Implementation and Results:

Decision Tree Regressor



`R^2_score: 0.772, RMSE:299.36`

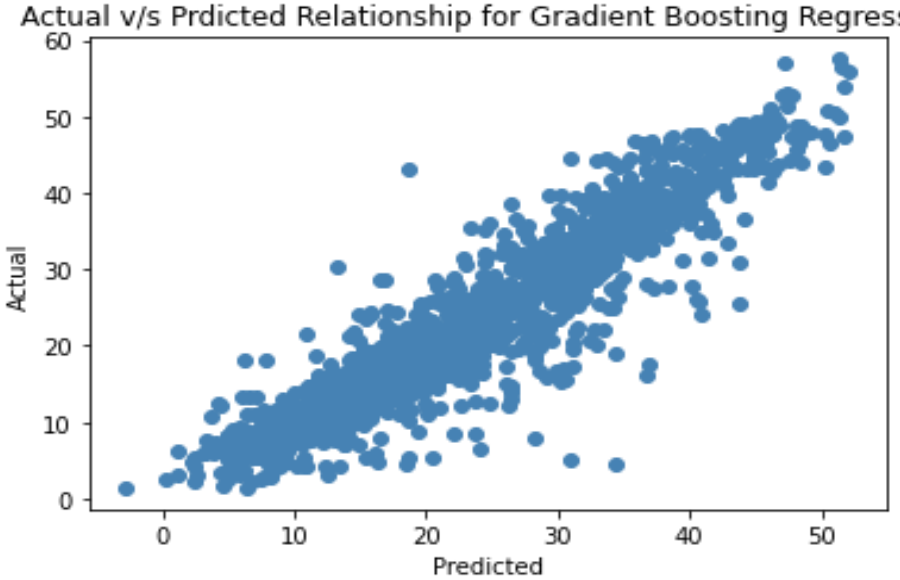
Random Forrest Regressor



`R^2_score: 0.817, RMSE:268.08`

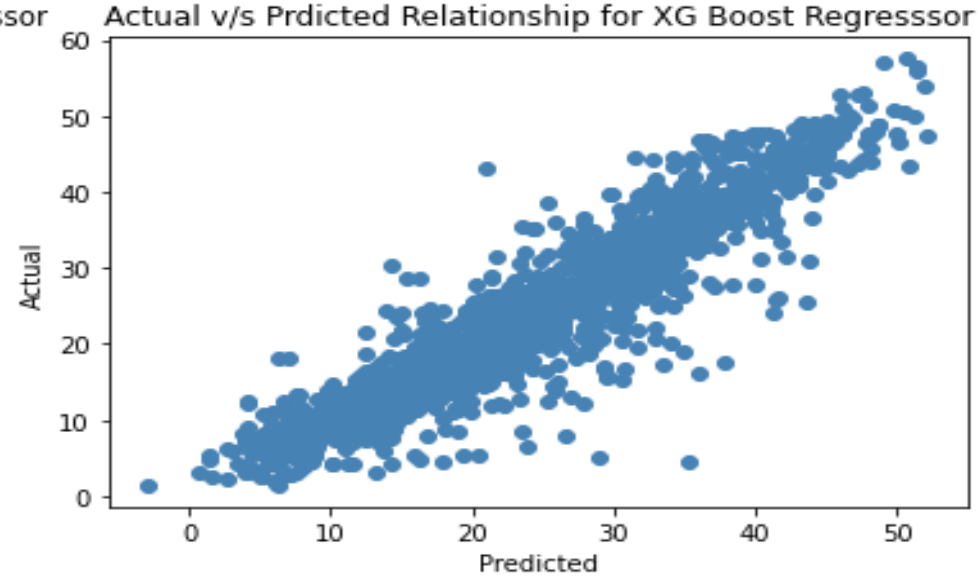
Model Implementation and Results(Cont..)

Gradient Boosting Regressor



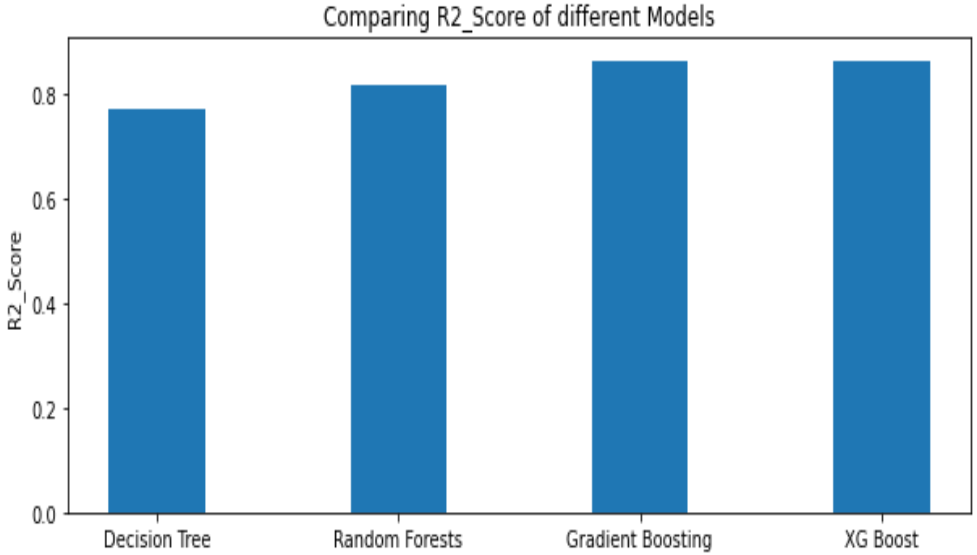
R^2_score: 0.863, RMSE:232.03

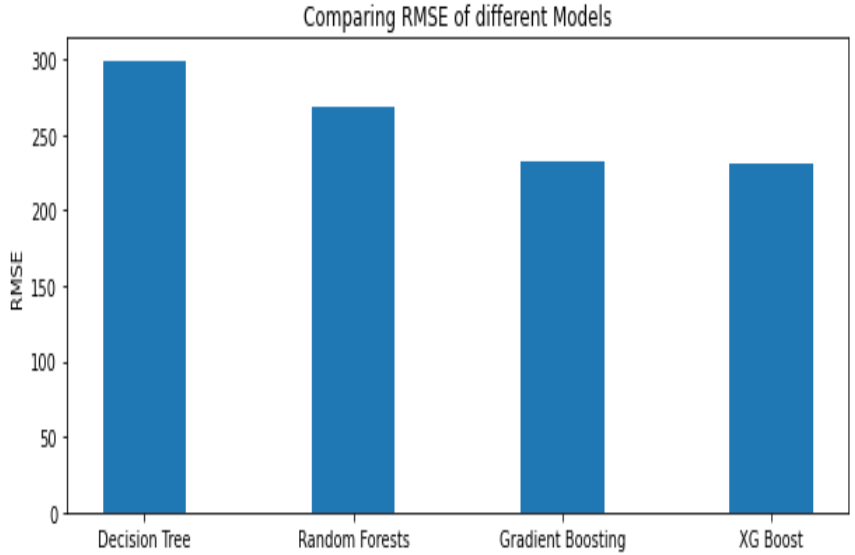
XG Boost Regressor



R^2_score: 0.863, RMSE:231.57

Comparing the results:





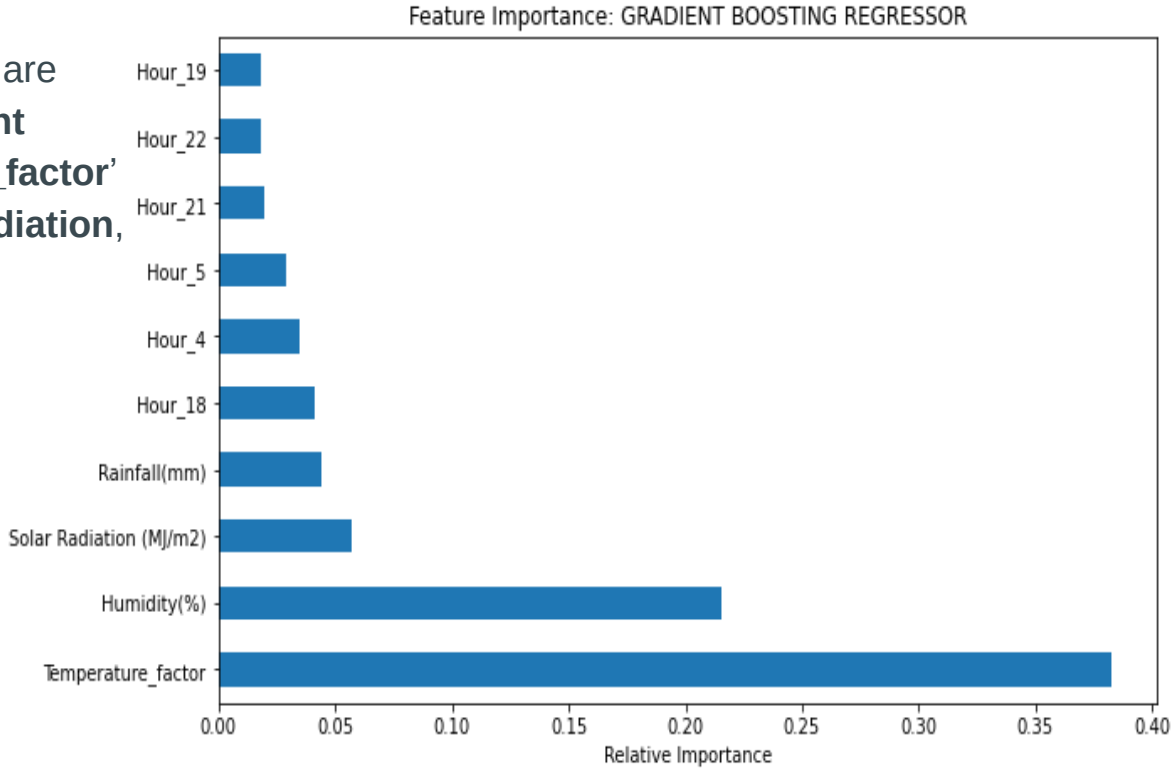
- Gradient Boosting and XG Boost model has approximately same result having highest r2_score and lowest RMSE value amongst all models.

- Keeping model complexity in mind, we will go with Gradient Boosting Regressor Model.

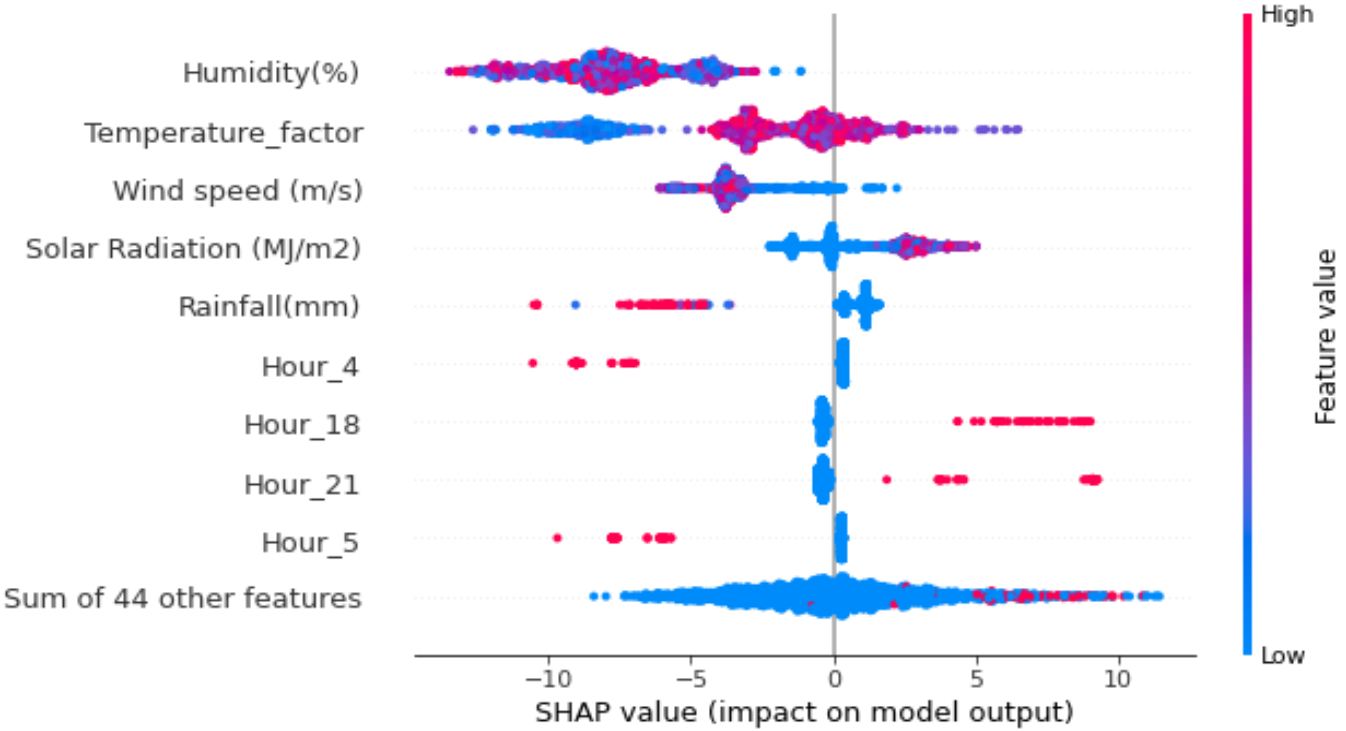
Model Explainability:

- It is important to explain the **behavior of the model** to the **stakeholders** so that they can understand which **factors** are controlling the **performance of the model**.

- The **most important features** which are affecting the performance of **Gradient Boosting Model** are '**Temperature_factor**' and '**Humidity**' followed by **Solar radiation**, **Rainfall** and **Hour**.



SHAP Value (Beeswarm plot)



Summary and Conclusions

1. After importing and analyzing the data, following insights were observed-

- The number of bikes rented is higher during the rush hours - which might be due to the office/schools/college timings.
- People prefer to rent bike most in summer and least in winter which can be also observed in month v/s number of bike rented plot as it is rented most during the month of May, June, July which is summer.

- The number of bikes rented is higher on working days than on Holidays which is due to employees and students.
- On a non-functioning day, there were no any bikes rented at all.
- The number of bikes rented based on weekday is hard to understand but it is for sure that it is least on Sunday.
- As compared to 2017, there is a huge growth in the bike renting demand for the year 2018.

Summary and Conclusions (Cont..)

2. Since the data contains many outliers, linear regression is not a good in this condition, following models are applied to this dataset:

- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- XG Boost Regressor

3. Gradient Boosting model is found to best suited for this dataset having high test r2_score and less test RMSE.

4. However the best model depends on the requirement of stakeholders, if the accuracy is of high concern rather than the explainability of model then Gradient Boosting is best. While if model explainability is of high concern then Decision Tree will be the best choice as it is easy to explain the model compared to other models.

THANK YOU