Capstone Project-2

Bike Sharing Demand Prediction

(Supervised Machine Learning regression)

BY

NEHAL S JAMBHULKAR

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 final aim of this project is the prediction of bike count required at each hour for the stable supply of rental bikes.

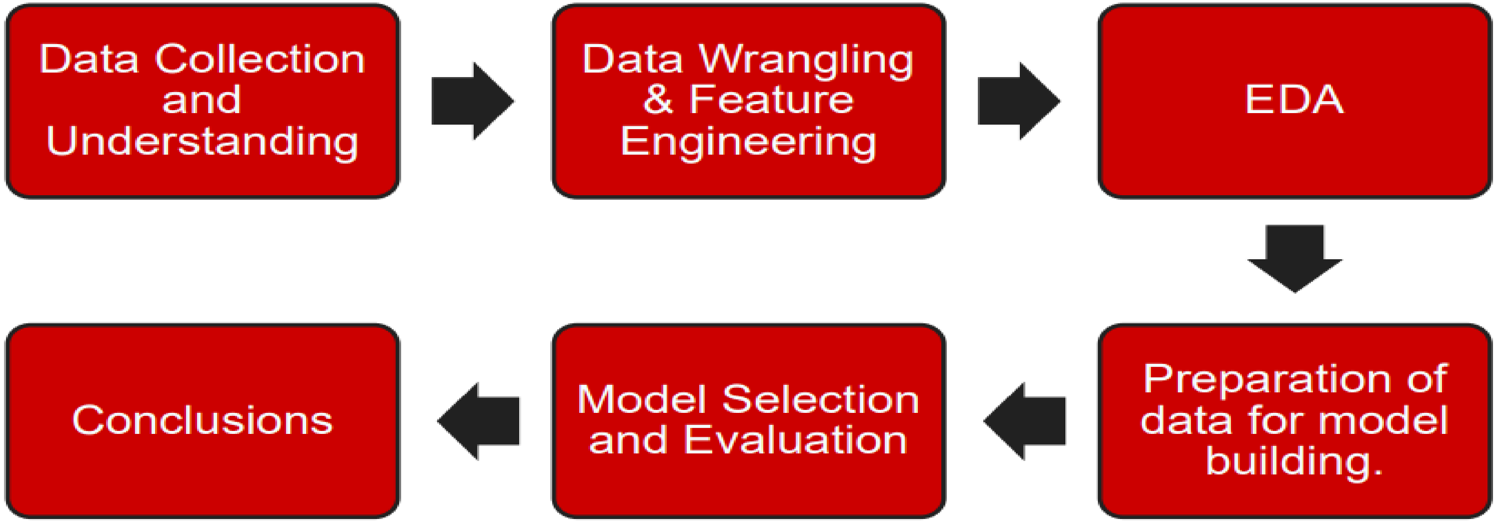
BUSINESS UNDERSTANDING

- Bike rentals have became 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 which that’s why they prefer rental bikes.
- Therefore, the business to strive and profit more, it has to be always ready and supply no. of bikes at different locations, to fulfil the demand.
- Our project goal is a pre planned set of bike count values that can be a handy solution to meet all demands.

Work Flow :

- So we will divide our work flow into following steps

.



Data Collection and Understanding

Data Wrangling & Feature Engineering

EDA

Preparation of data for model building.

Model Selection and Evaluation

Conclusions

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.
- We have 8740 rows and 14 columns, out of which 11 are numerical and 3 are categorical. Rented bike Count is our dependent variable.

Description of columns:

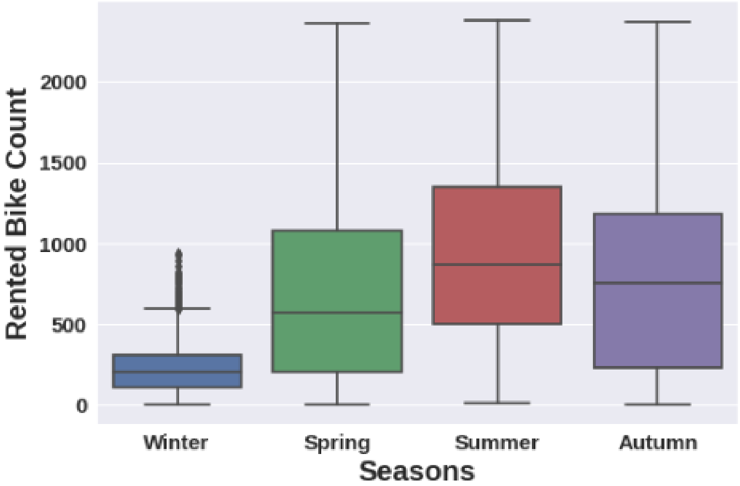
- Rented Bike count - Count of bikes rented at each hour - Int64
- Date : year-month-day - Object
- Hour - Hour of he day - int64
- Temperature-Temperature in Celsius - float64
- Humidity - % - float64
- Wind-speed - m/s - float64
- Visibility - 10m -int64
- Dew point temperature - Celsius -float64
- Solar radiation - MJ/m2 - float64
- Rainfall - mm - float64
- Snowfall - cm - float64
- Seasons - Winter, Spring, Summer, Autumn - object
- Holiday - Holiday/No holiday - object
- Functional Day - NoFunc(Non Functional Day), Fun(Functional Day) - object

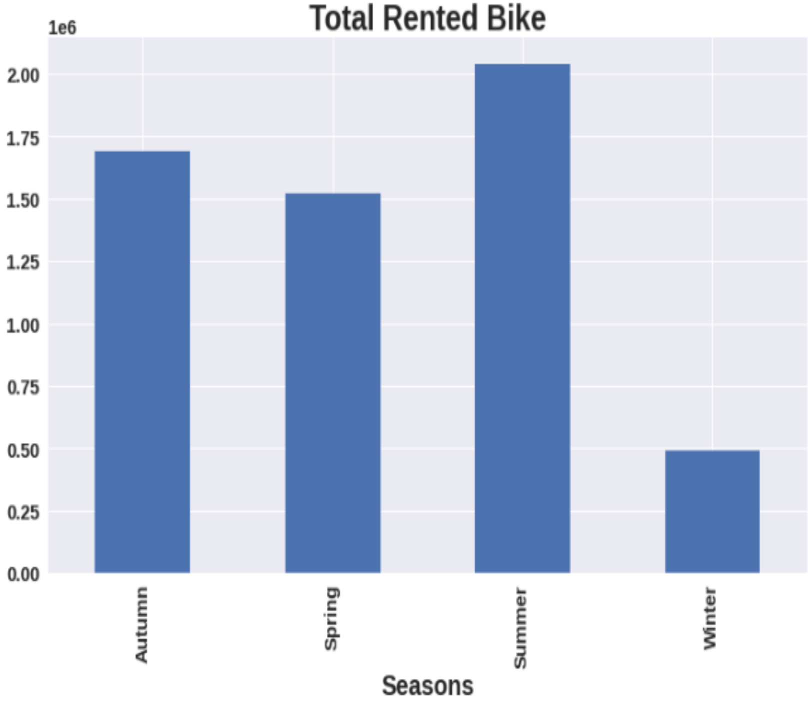
Data wrangling

- Converting date column to date-time and extracting day, month and year.
- Outlier detection
- There are No Duplicate values present
- There are No Missing Values present
- And finally we have dependent feature 'rented bike count' variable which we need to predict for new observations .
- The dataset shows hourly rental data for one year (1 December 2017 to 31 November(2018)(365 days).we consider this as a single year data

EDA (Exploratory Data Analysis):

Analysing Count of Rented Bikes for different seasons

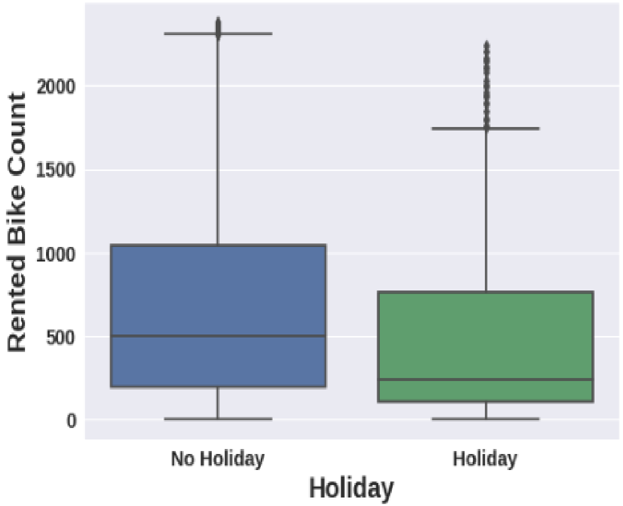


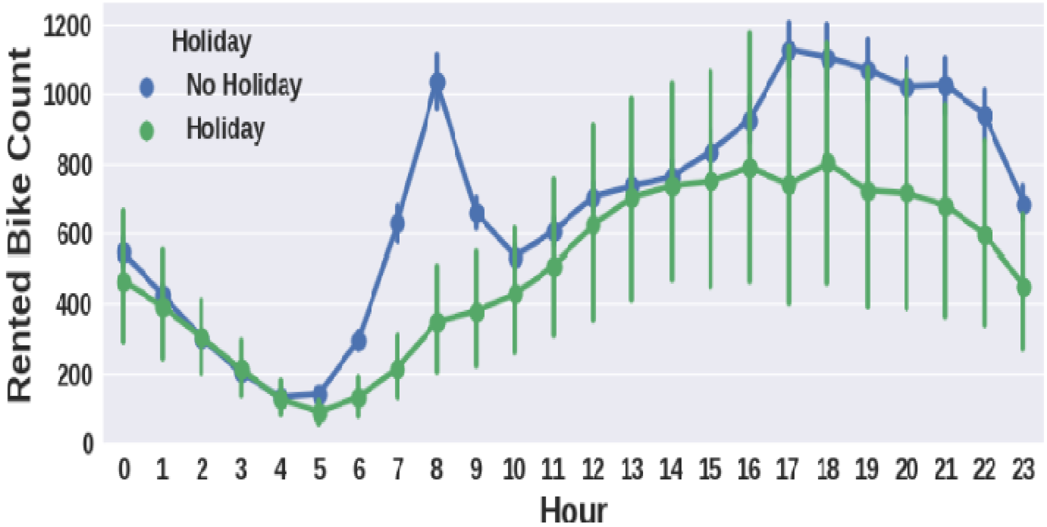


.

- Summer season was the peak (2208) of all the activity with the most number of Rented bike count.
- Whereas; Winter was the least (2160) popular season with the bike counts recorded.

Analysing at which hour the Rented Bike count is maximum with respect to Functional day

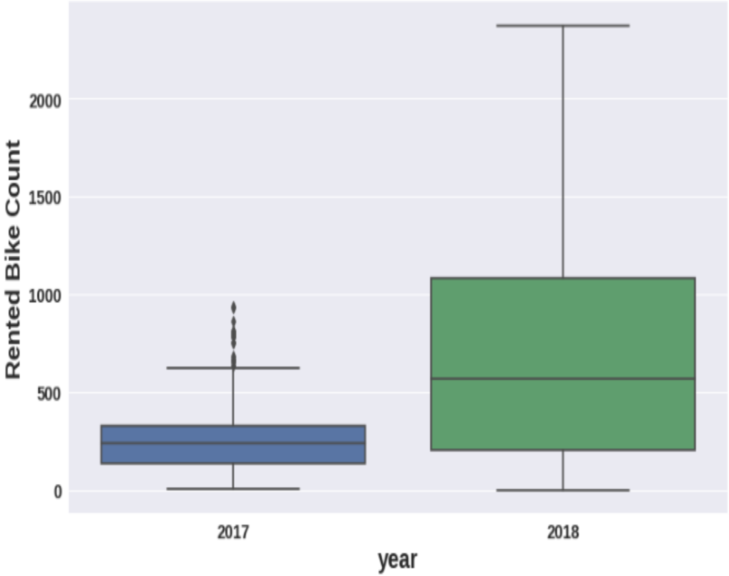


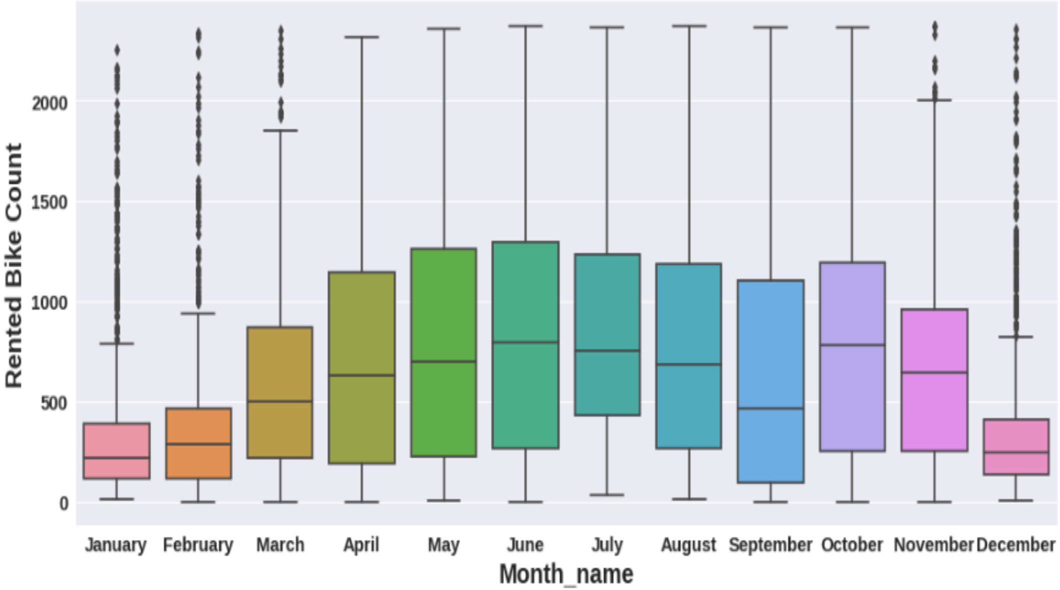


● The above trends indicate that during the holidays the demands of bikes plummet down. Maybe due to lower travel activity and people prefer to stay at homes more.

● Whereas on “No holidays” - the demand is very high around 6-9 and 18-22 hour of the day, as it maybe a convince to get home after work.

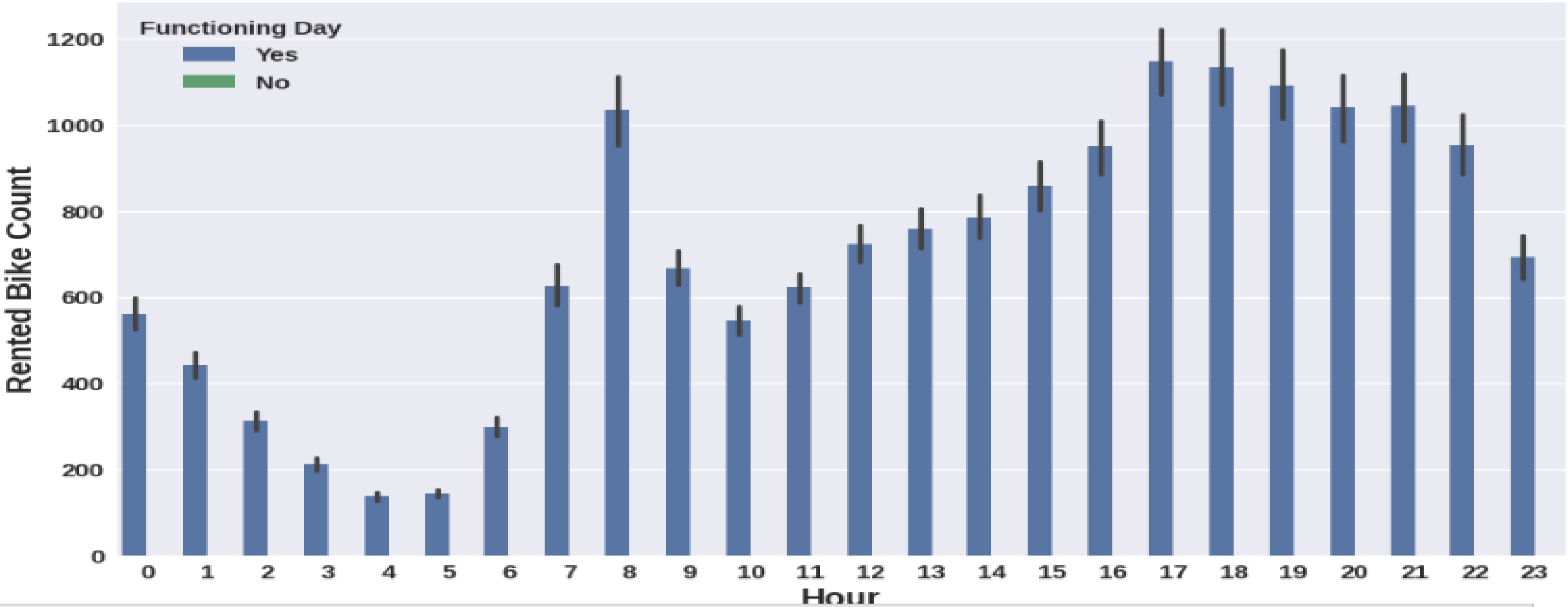
Analysing in which Year and Month the Rented Bike Count was maximum:





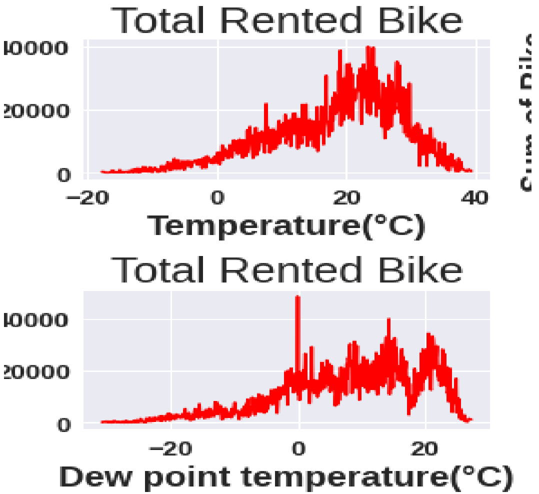
- ➢ The demand increased drastically from 2017-2018.
- ➢ **We see that number of ride count drastically increases between April to june which are comparitively summer season**
- ➢ **From the above boxplot we can say that at starting and ending of year demand of bike decrease (because of winter season) apart of these month demand of bike in other month is equal.**

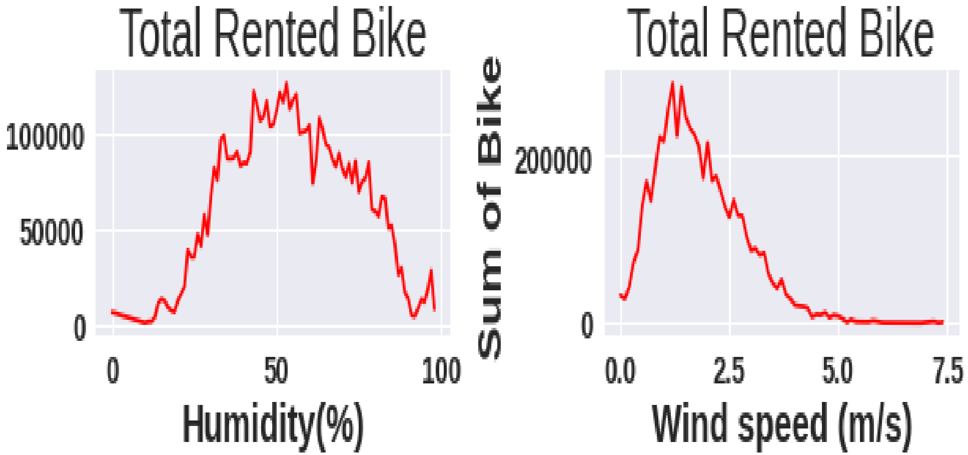
How demand of bike change with Functioning days



1. From above bar graph we get to know that there was no bike rent on non functioning day

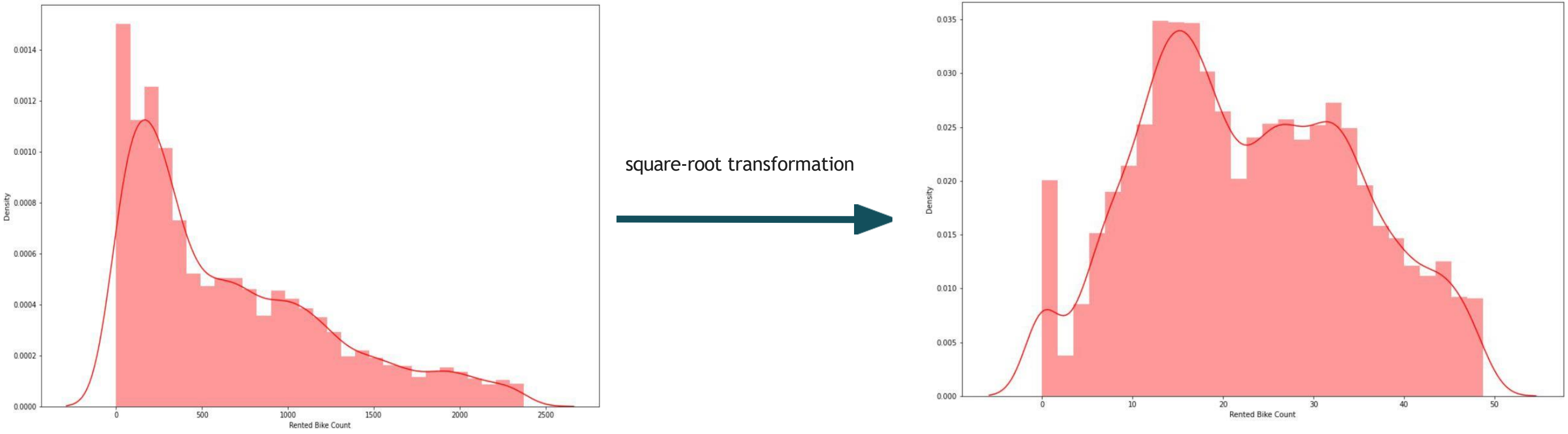
Number of bike(bike trend) at different-different wheather condition:





- As we can seen from above garph that at lower and higher tempreture required number of bike is less,same trend with humidity
- when wind speed is higher ,number of required bike is less. we think these are very basic things because of health concern. during high/low tempreture and high wind speed people not preffer bike for travalling from one place to other place.
- Also during rainfall,snowfall,less visibility required bike is very less ,it's obvious, right.as we always concerned about our health and safety so during rainfall ,snowfall, solar radiation, no body prefer bike for travalling

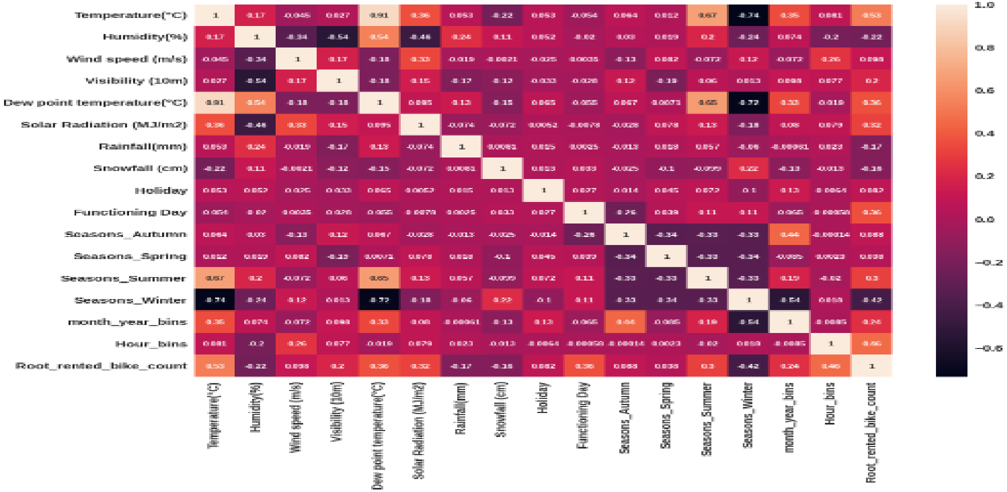
Distribution of Dependent Variable



As it was right skewed, so we have taken square root of dependent variable to visualize it in a better way...

To normalize the distribution we applied square root method. After normalization no outliers were found.

Preparation of data for model building:



Observation

As we can see from above correlation heatmap that tempreture is highly correlated with Dew point tempreture ,so there is collinearity between these two features.

To remove multicollinearity from our data we need to remove either tempreture column or Dew point tempreture column

As we can see from heatmap that tempreture column has high correlation with our target variable (rented bike count) compare to Dew point tempreture ,so let's remove Dew point tempreture column from our dataframe.

Apart from these two features ,there is no high correlation between independent features of our data.

Season_summer and Season_winter is also highly correlated with tempreture,it's obvious becuase tempreture increase in summer and decreases in winter. so let's remove these two columns from dataframe.

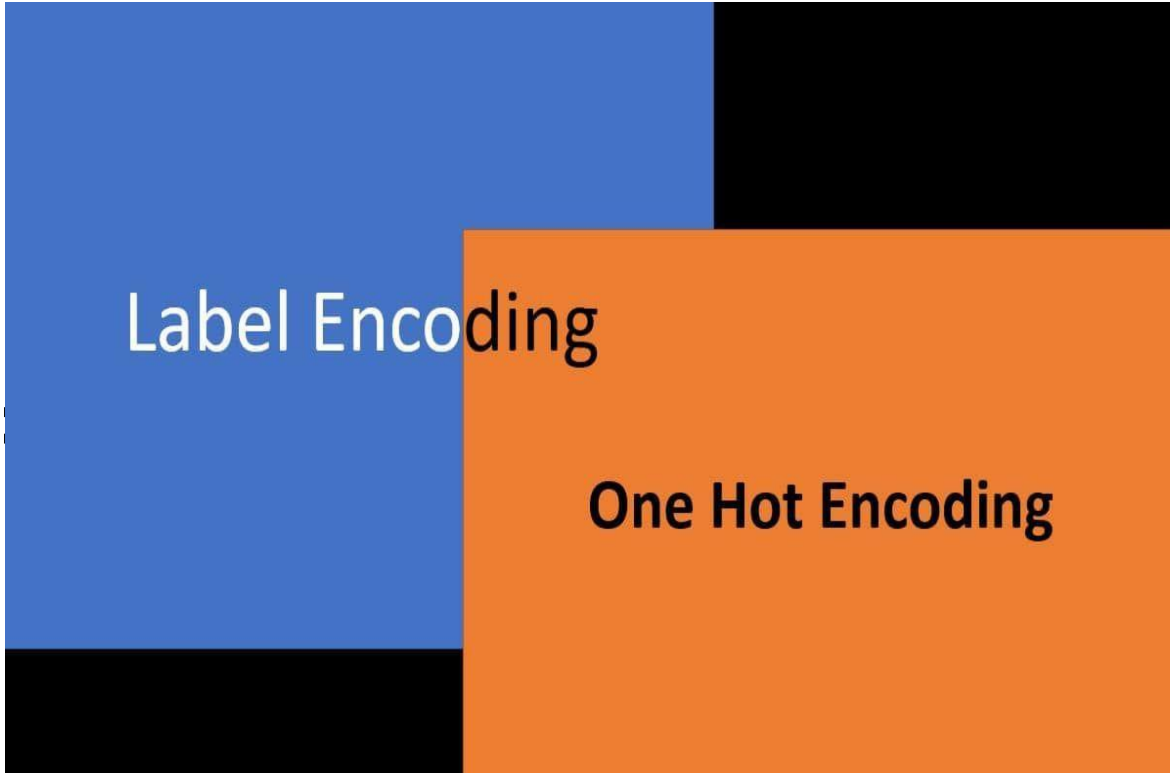
Varienece inflation factor help us in removing multicollinearity.

As we calculated varience inflation factor for our dataframe features and we can say from above dataframe that for all numeric features varience inflation factor is less than 10, so we can keep these feature to fit in our linear regression machine learning model.

Pre-processing of Data :-

Feature Engineering:

- Label Encoding:
- One hot Encoding



Feature Engineering:

- Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.
- Feature engineering, in simple terms, is **the act of converting raw observations into desired features using statistical or machine learning approaches**.
- It is the process of designing artificial features into an algorithm. These artificial features are then used by that algorithm in order to improve its performance, or in other words reap better results.

Label Encoding:

- Label Encoding refers **to converting the labels into a numeric form so as to convert them into the machine-readable form**.
- Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.
- Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.
- Here we encode the variables like Functioning Day and Holiday in the form of 0 and 1. also convert the seasons column into dummy variables like Spring, Summer, Rainy and Winter.

One hot Encoding:-

- A one hot encoding is **a representation of categorical variables as binary vectors**.
- This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary
- In this Feature Engineering, we apply lambda function to convert respective columns in the form of 0 and 1. Ex. We convert Visibility column in the form of 1 when it is greater than 2000, also for rainfall if the value is greater than
- 0.148 then it is converted into 1 otherwise 0. Same procedure follows for snowfall and solar radiation

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 with regularizations**
- **Polynomial Regression**
- **Decision tree**
- **Random forest**
- **Gradient Boosting**
- **eXtreme Gradient Boost**

Assumptions of regression line:

1.The relation between the dependent and independent variables should be almost linear.

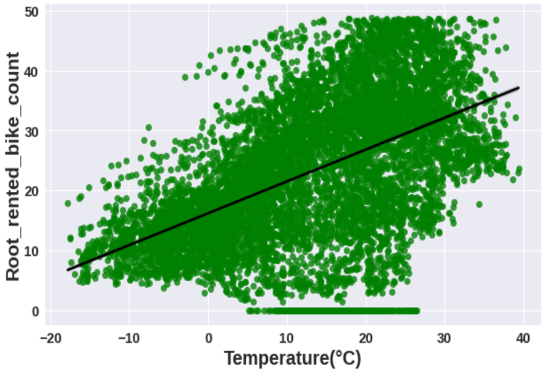
2.Mean of residuals should be zero or close to 0 as much as possible. It is done to check whether our line is actually the line of "best fit".

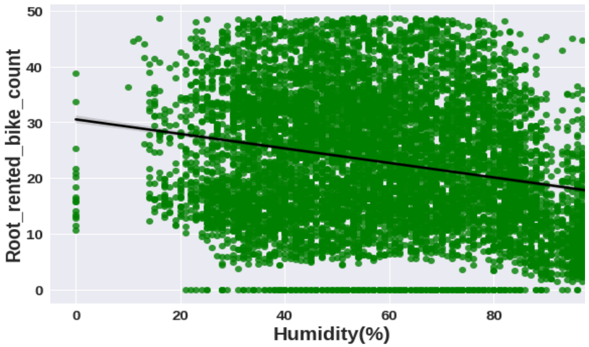
3.There should be homoscedasticity or equal variance in a regression model. This assumption means that the variance around the regression line is the same for all values of the predictor variable (X).

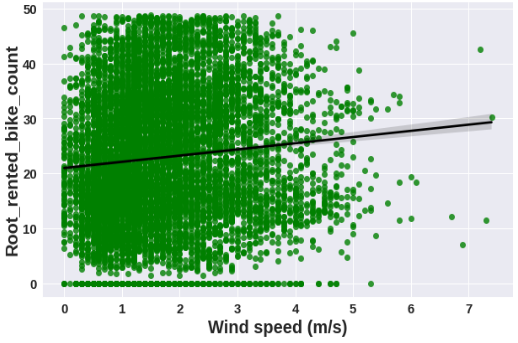
4.There should not be multicollinearity in regression model. Multicollinearity generally occurs when there are high correlations between two or more independent variables.

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.

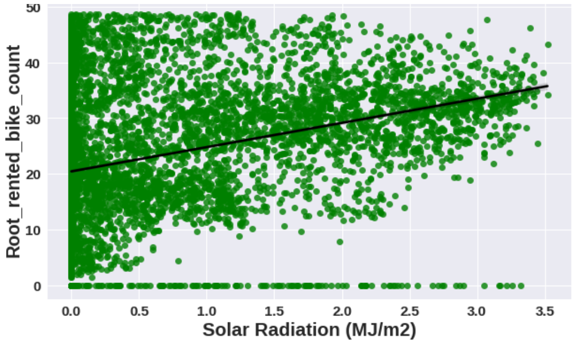
Model Selection and Evaluation :

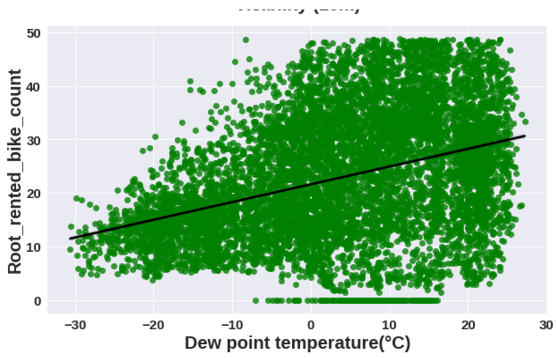


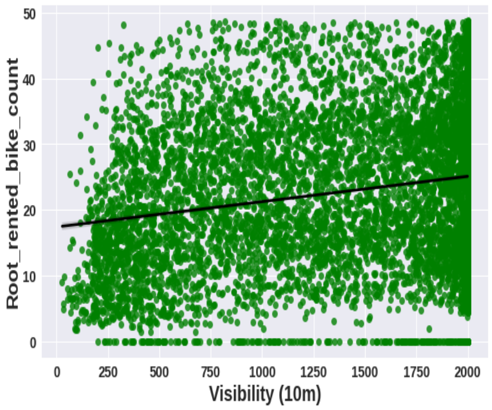




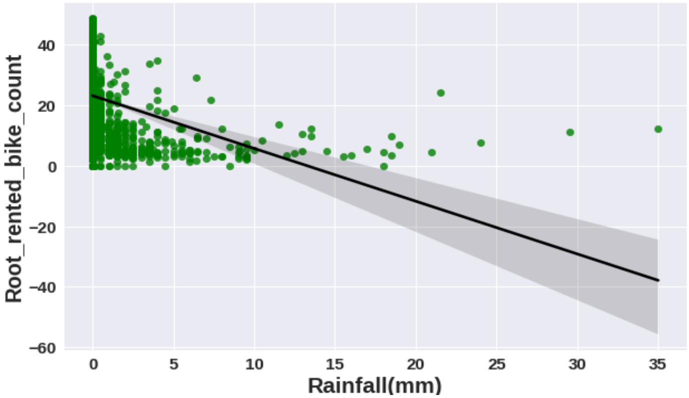
- From the above regression plot of all numerical features we see that the columns 'Temperature', 'Wind_speed', 'Visibility', 'Dew_point_temperature', 'Solar_Radiation' are positively relation to the target variable,which means the rented bike count increases with increase of t

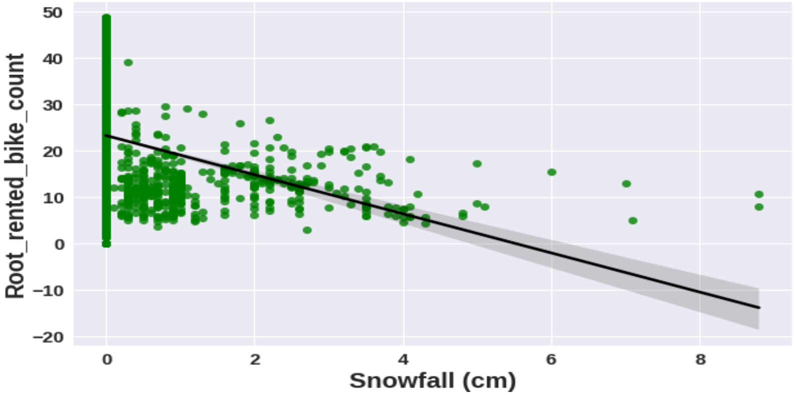






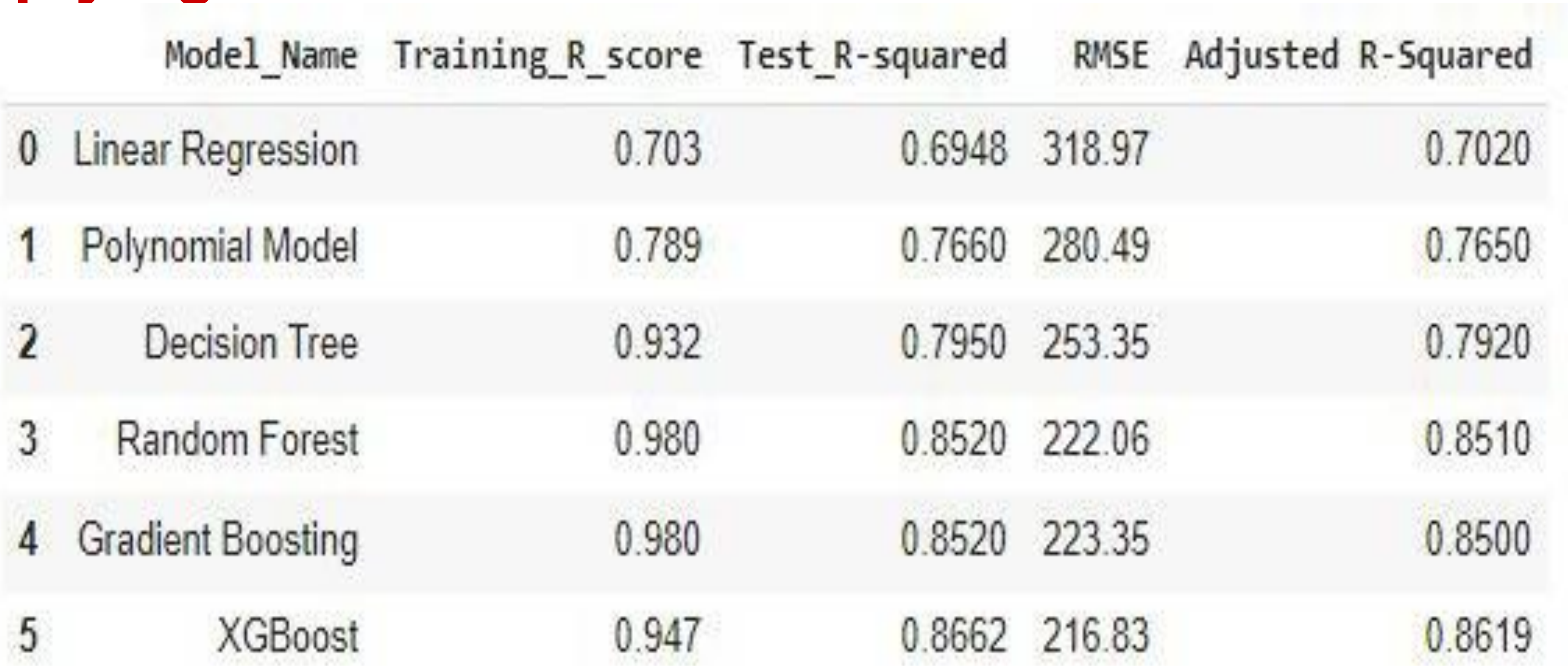
Model Selection and Evaluation :





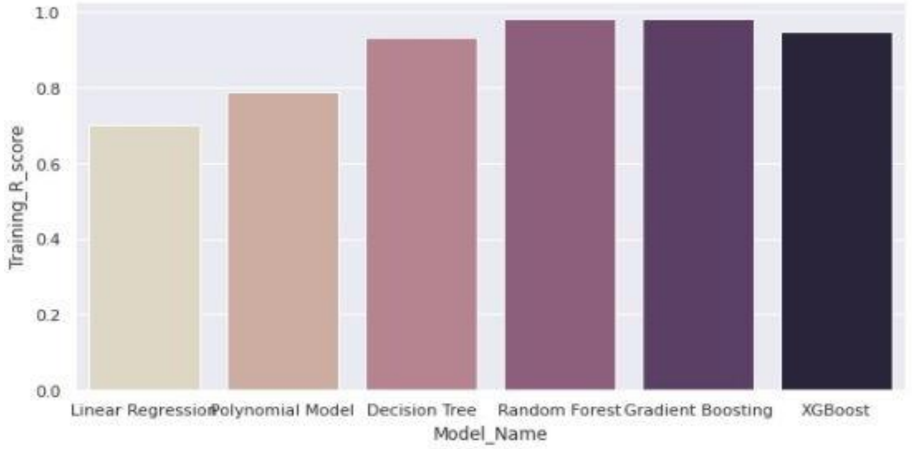
- 'Rainfall',‘ Snowfall’, 'Humidity' these features are negatively related with the target variable which means the rented bike count decreases when these features increase.

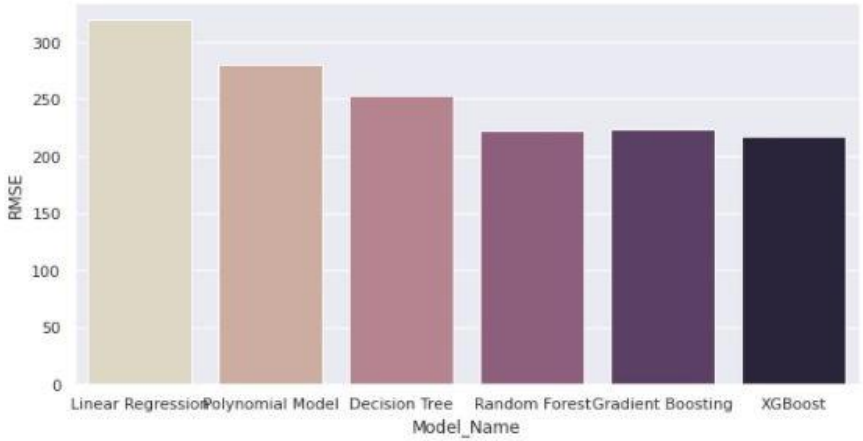
Applying Models - Validation & Selection

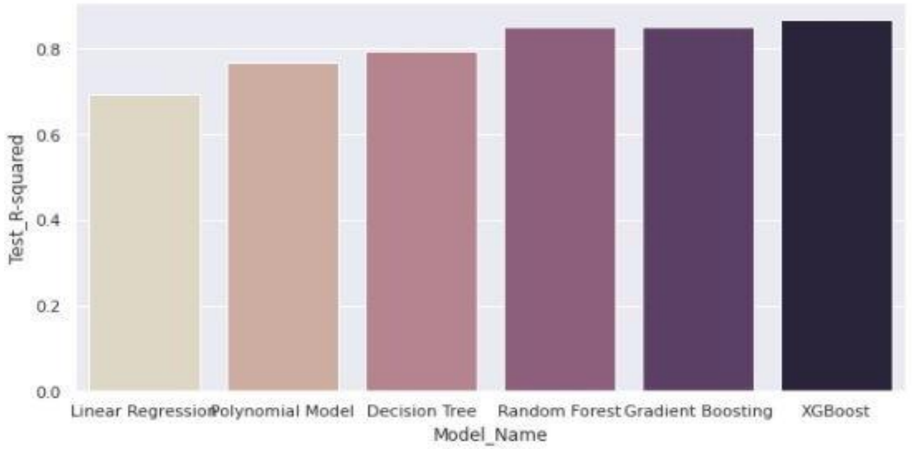


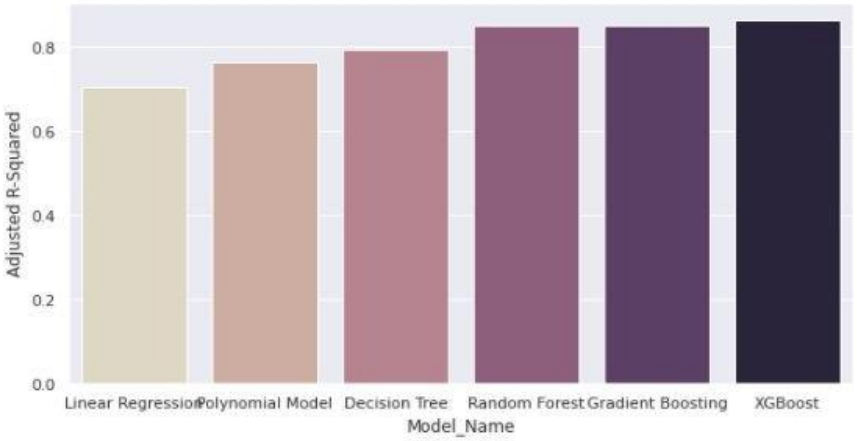
	Model_Name	Training_R_score	Test_R-squared	RMSE	Adjusted R-Squared
0	Linear Regression	0.703	0.6948	318.97	0.7020
1	Polynomial Model	0.789	0.7660	280.49	0.7650
2	Decision Tree	0.932	0.7950	253.35	0.7920
3	Random Forest	0.980	0.8520	222.06	0.8510
4	Gradient Boosting	0.980	0.8520	223.35	0.8500
5	XGBoost	0.947	0.8662	216.83	0.8619

Model - Validation & Selection









Model - Validation & Selection

Observation 1: As observed from the table above Linear Regression did not generated great results, some improvement in the results were achieved by Polynomial linear regression and Decision tree, but had lower Test_R_squared values.

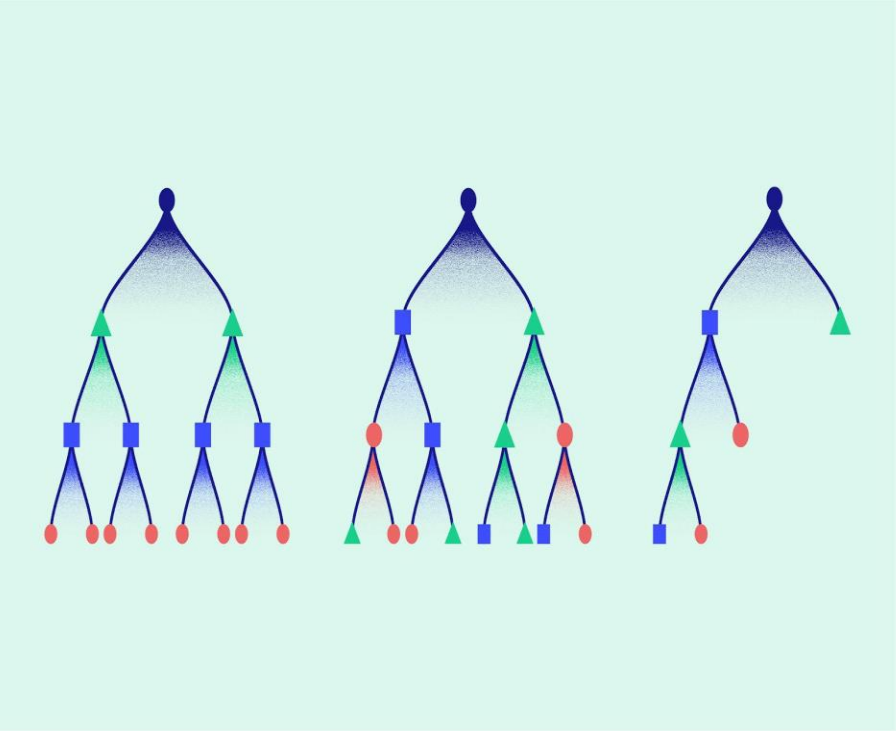
Observation 2: Random forest & Gradient Boosting have performed equally good, but XGBoosting take the best place of all.

Observation 3: From the above observation we have come to a conclusion that we would choose our regression model from XGBoost.

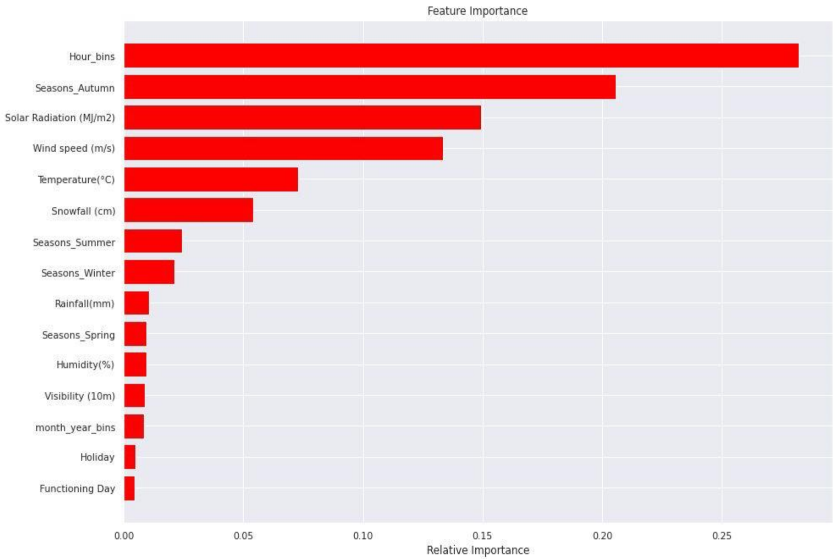
About the model

Since we have chosen XGBoost as our regression model. Below are the best hyperparameters:

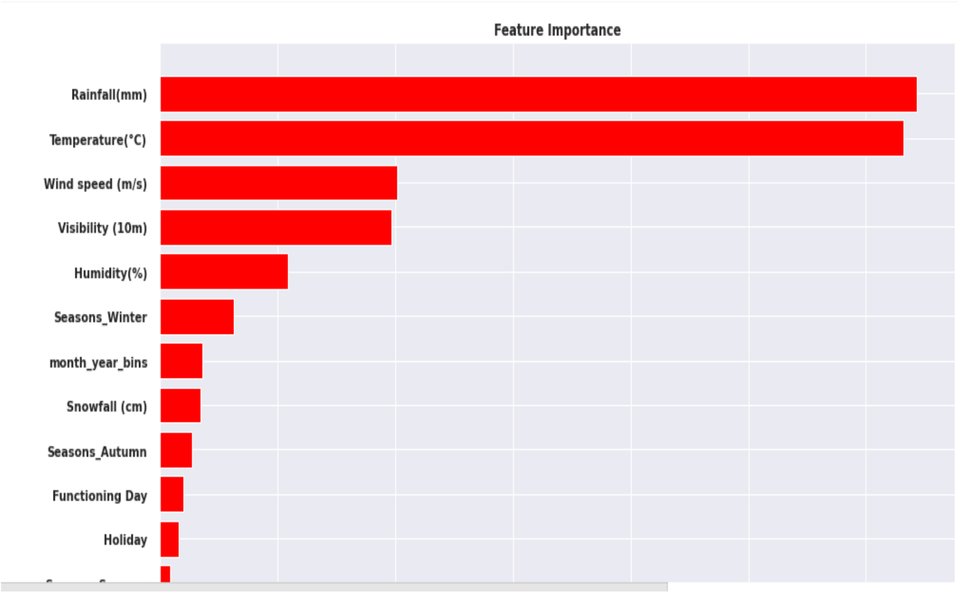
```
'max_depth': 6
'min_child_weight': 12
'n_estimators': 200
```



Feature Importance



XGBOOST



Decision tree

Challenges

- As dataset was quite big enough which led more computation time.
- Data Cleaning
- A huge amount of data needed to be dealt while doing the project which is quite an important task and also even small inferences need to be kept in mind.

Conclusion

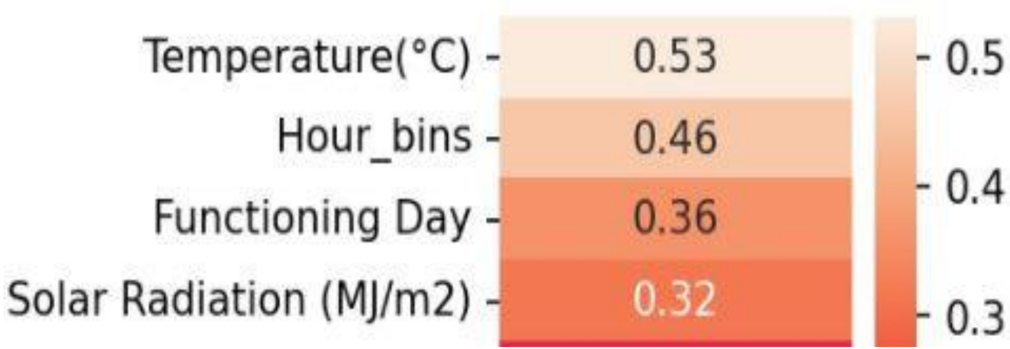
- From above model we can conclude the below points:
- Rented Bike Count is very much dependent :-

1) At what Temperature the Bike is rented.

2) On what Hour the Bike is rented.

3) How much Humidity present in the atmosphere.

4) Is it a functioning day or not.



Conclusion

- In project, after trying combinations of features with linear regression the model underfit. It seemed obvious because data is spread too much. It didn't seem practical to fit a line.
- Rainfall is the most influencing feature and winter is at the second place for Linear Regressor.
- Temperature is the most important feature and Hour is at second place for Decision Tree, Random Forest and Gradient Boosting Regressor.
- Winter is the most important feature and Functioning day[yes] is the second most for XGBoost Regressor.
- The feature temperature is on the top list for all the regressors except XGBoost and Linear Regessor
- The experimental results prove that the XGBoost model predicts best the trip duration with the highest R2 and with less error rate compared to Linear Regression, Decision Tree, Random Forest, Gradient Boosting.

Thank You