Capstone Project

Seoul Bike Sharing Demand Prediction

Sonali Kaushal
Somya Hingorani
Nitesh Verma
Harsh Vardhan
Prateek Gupta

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.

Agenda

To discuss the analysis of given bike renting data set from 2017-2018.

Topics covered for the project :

- Data Pipeline
- Data Summary
- Data Description
- Feature engineering
  - Multicollinearity
  - Distribution of target variable
- Exploratory Data Analysis
  - Demand analysis of rented bikes
- Model Overview
- Regression
  - Model’s Evaluation Matrices
  - Features of models
  - Adjusted R2 of Model’s Performed
- Conclusion

Data Pipeline

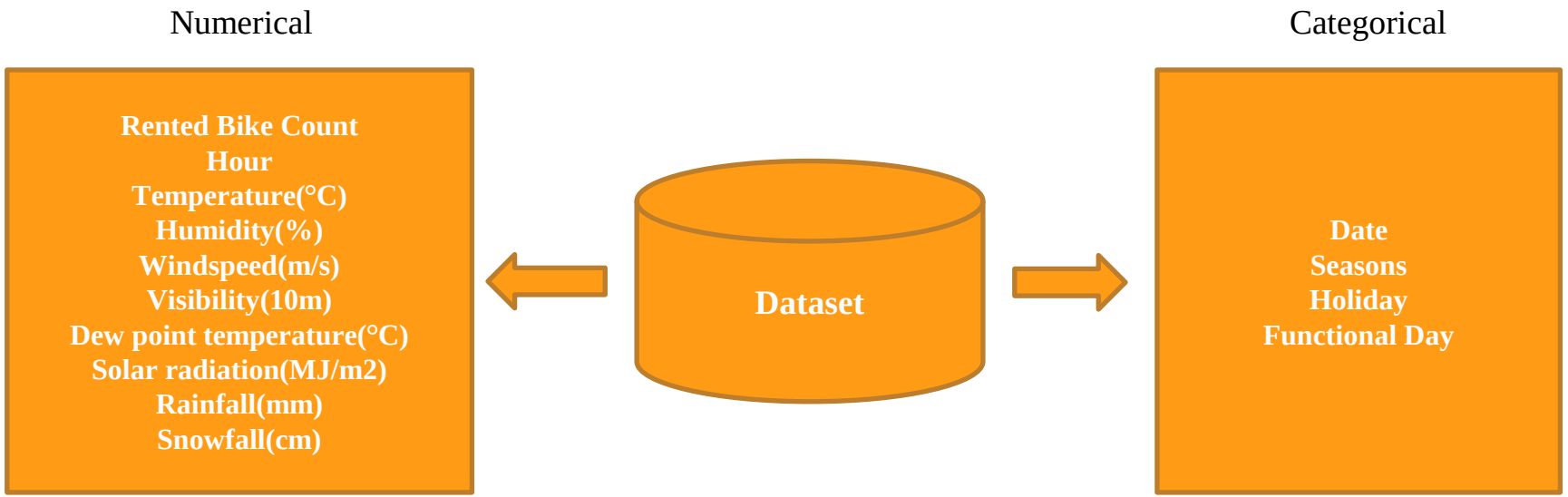
- Data pre-processing: We pre processed the data by dealing with the outliers, null values, and duplicate data.

- Feature engineering: In this part we went through each attributes and encoded the categorical features.

- Exploratory Data Analysis (EDA): In this part we have done some EDA on the features to get insights.

- Model Creation: Finally in this part we created the various models. These various models are being analysed and we tried to study various models so as to get the best performing model for our project.

Data Summary



Data Description

Dependent variable:

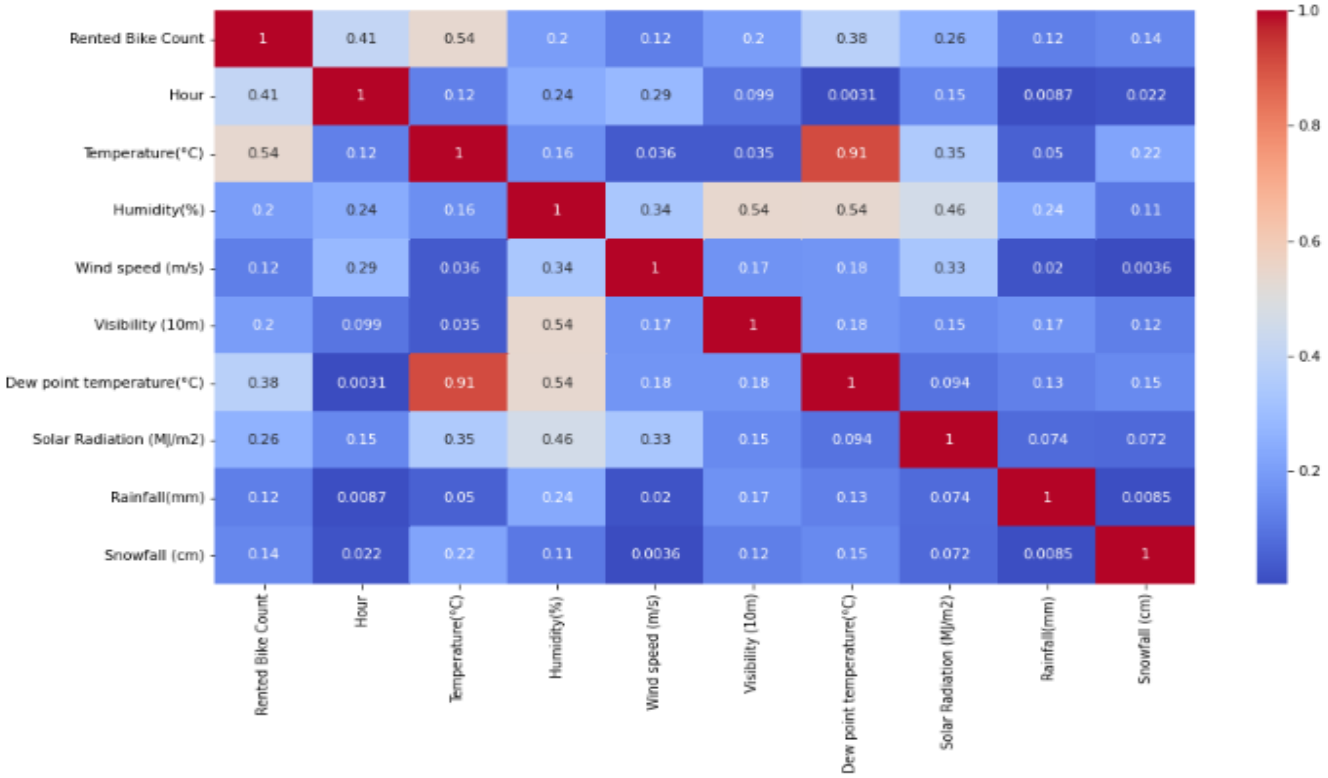
• Rented Bike count - Count of bikes rented at each hour

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)

Feature engineering

Multicollinearity

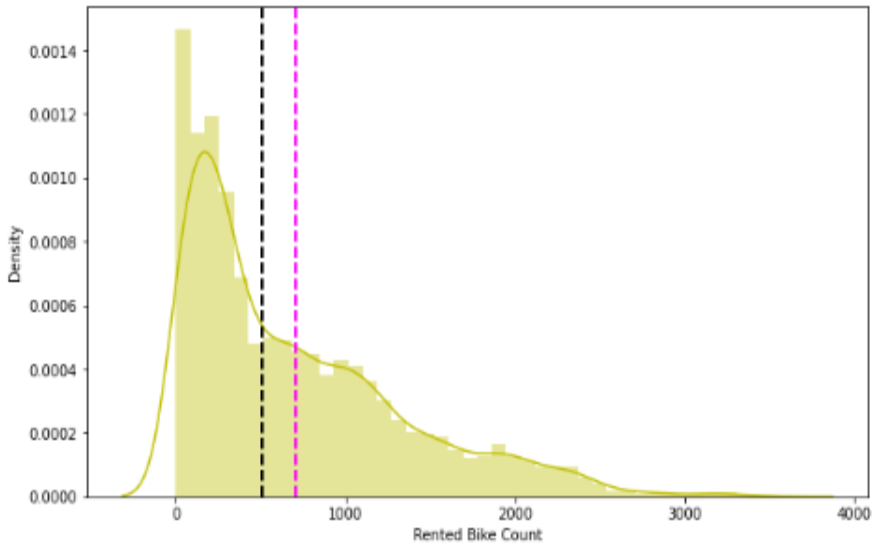


	variables	VIF
0	Hour	3.921832
1	Temperature(°C)	3.228318
2	Humidity(%)	4.868221
3	Wind speed (m/s)	4.608625
4	Visibility (10m)	4.710170
5	Solar Radiation (MJ/m2)	2.246791
6	Rainfall(mm)	1.079158
7	Snowfall (cm)	1.120579

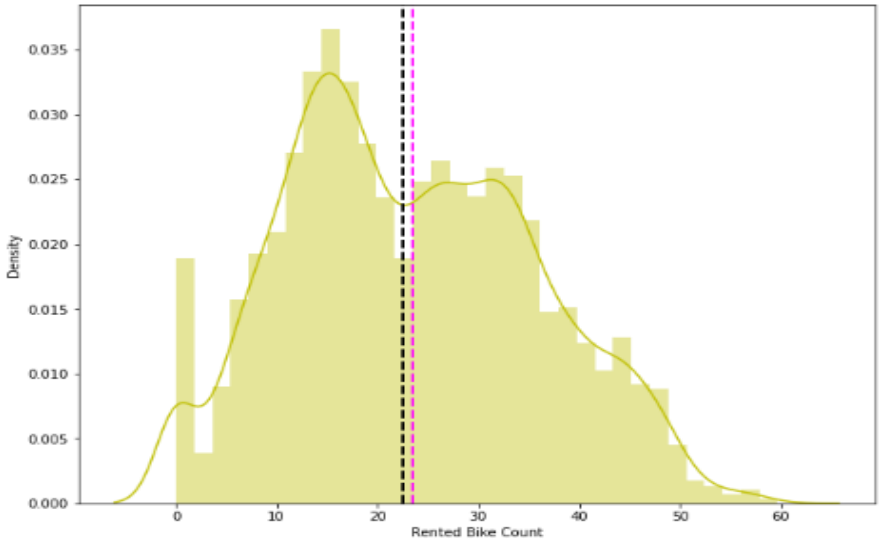
- Dew point temperature and Temperature are highly correlated
- Values of VIF's are less than 5 which is acceptable, there are less chances of multicollinearity

Distribution of target variable

Graph 1



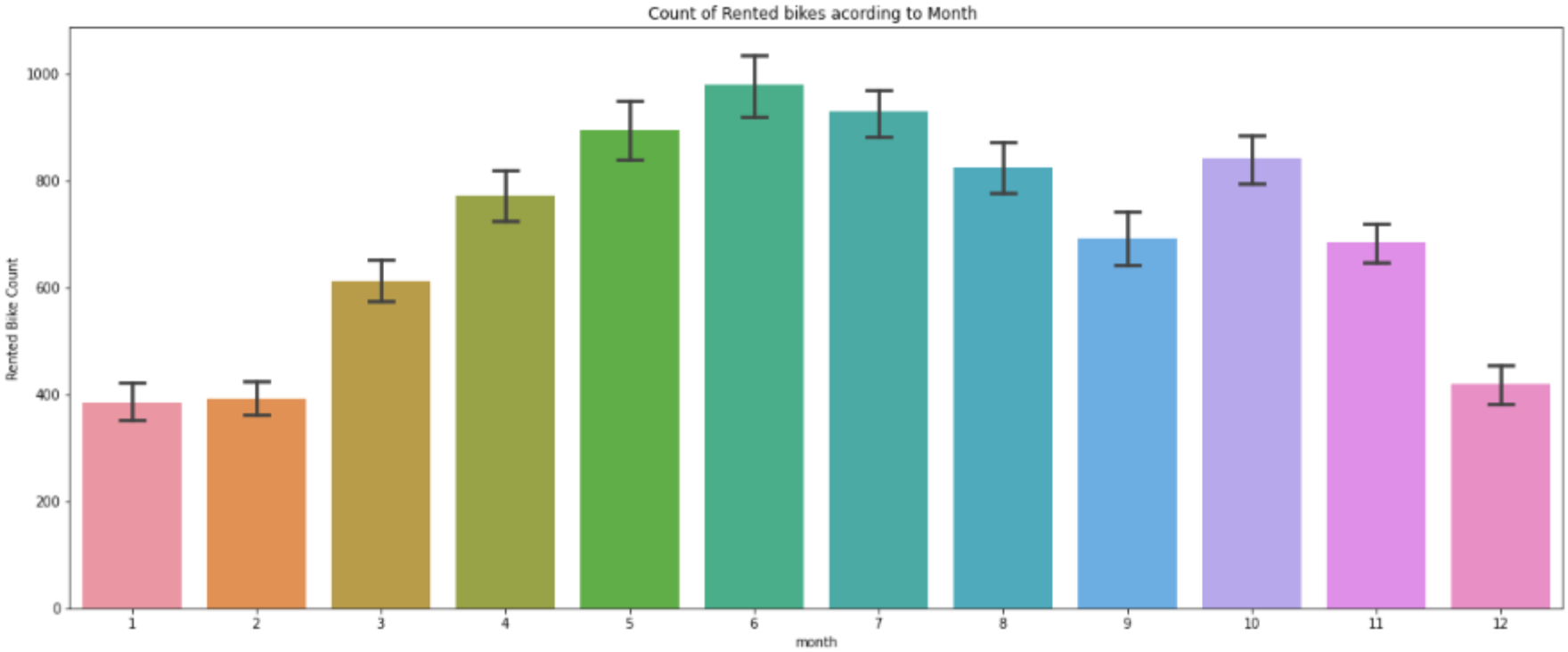
Graph 2



- Rented bike count column follow right tail distribution. (Graph 1)
- After doing square root of Rented Bike Count, it follows normal distribution. (Graph 2)

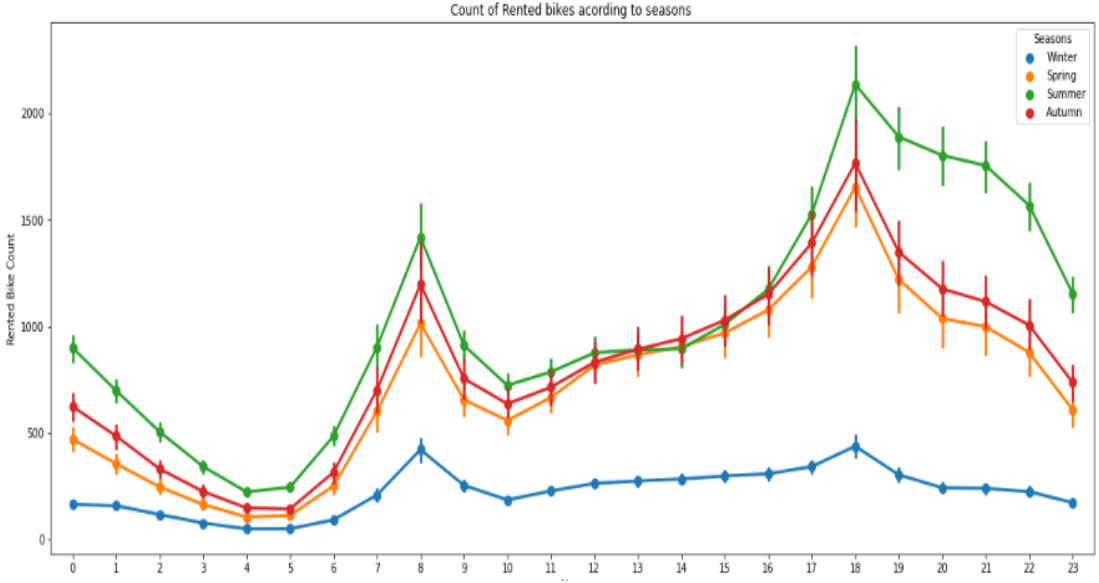
Exploratory Data Analysis

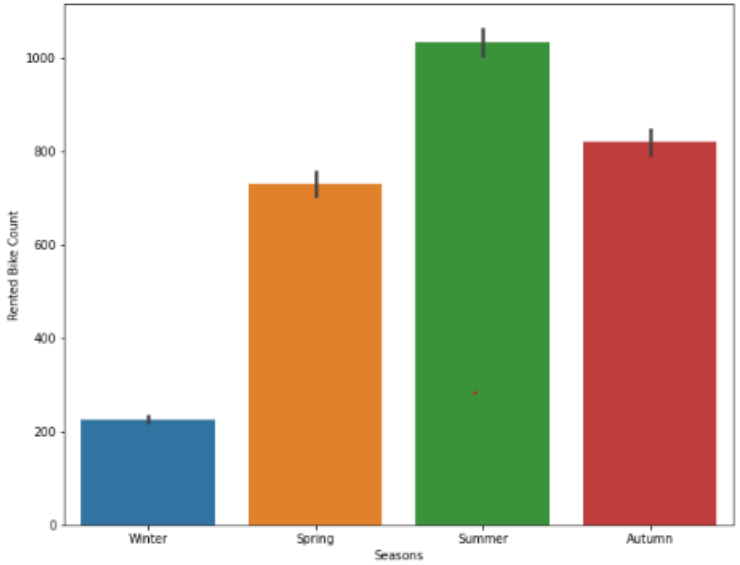
Monthly demand



- From the above bar plot we can clearly say that in month of June, there is high demand for rental bikes.
- The least demand of rented bikes is seen in the months of January and February.

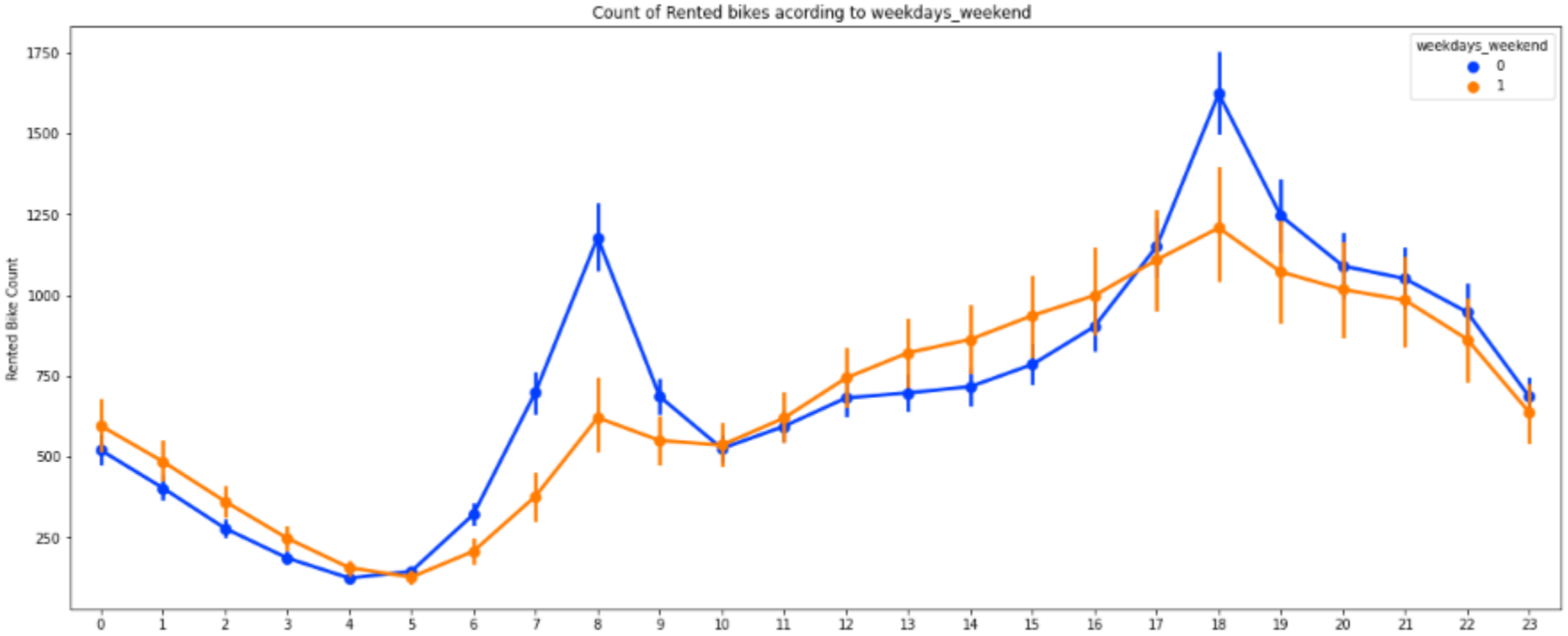
Seasonal demand for rented bikes





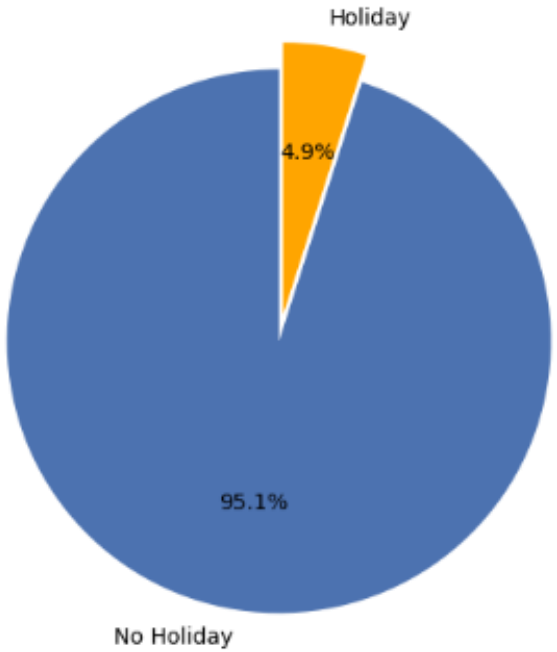
- As per the graphs Summer is the busiest season in regards to rented bike demand.
- Peak demand for rented bike is during 6pm (18:00).
- Rented bike demand is less during winter season, in comparison to other seasons.

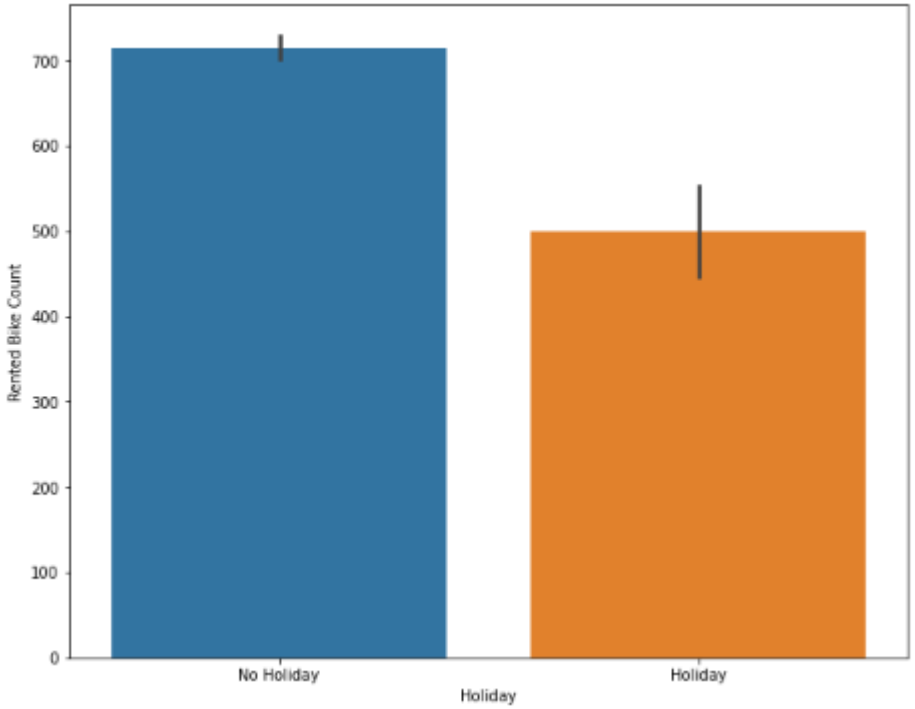
Demand of rented bikes (Weekdays Vs Weekends)



- According to point plot, weekdays (blue line) shows higher demand of rented bikes, in comparison to weekends (orange line).
- Peak demand for rented bike in weekend and weekdays is during 6pm (18:00).

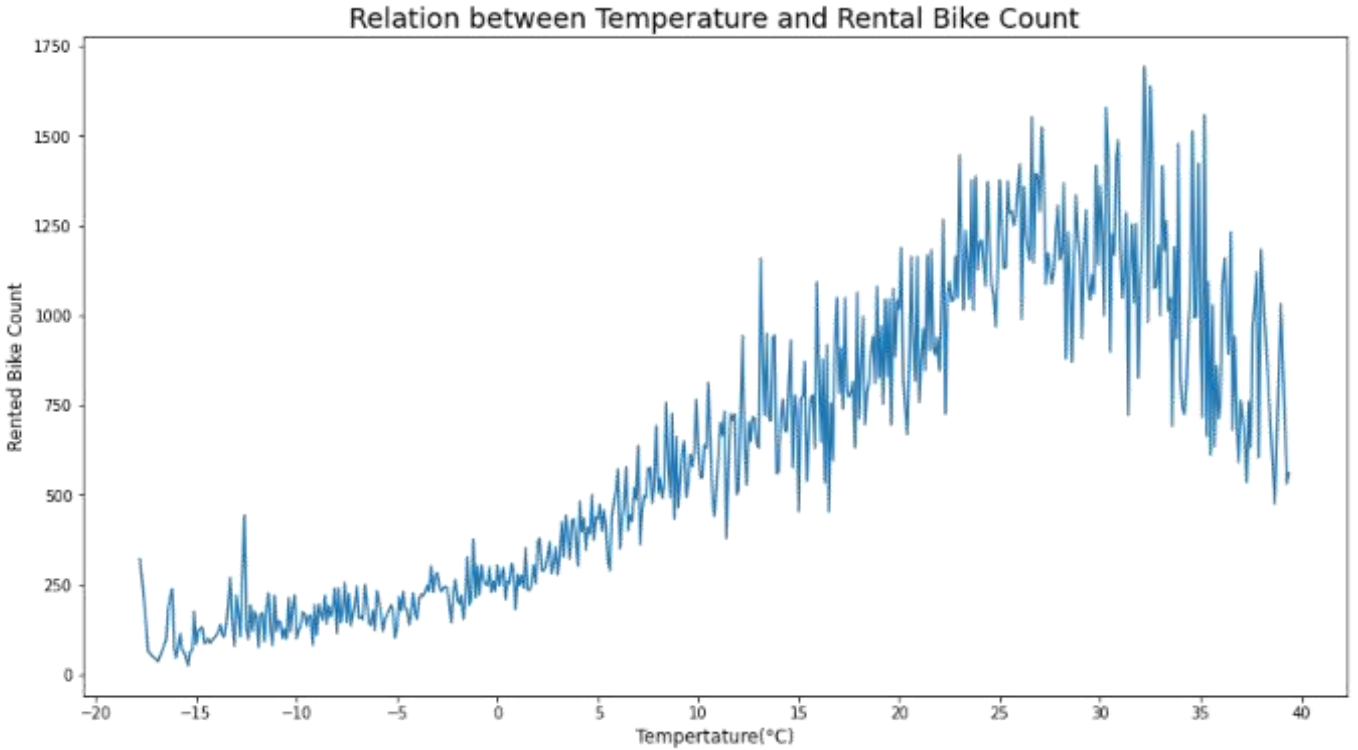
Demand on holidays and no holidays





- According to the pie chart, 95.1% of the data represent working day and only 4.9% of the data represent holiday.
- On working day, the rental bikes demand is high in comparison to holiday.

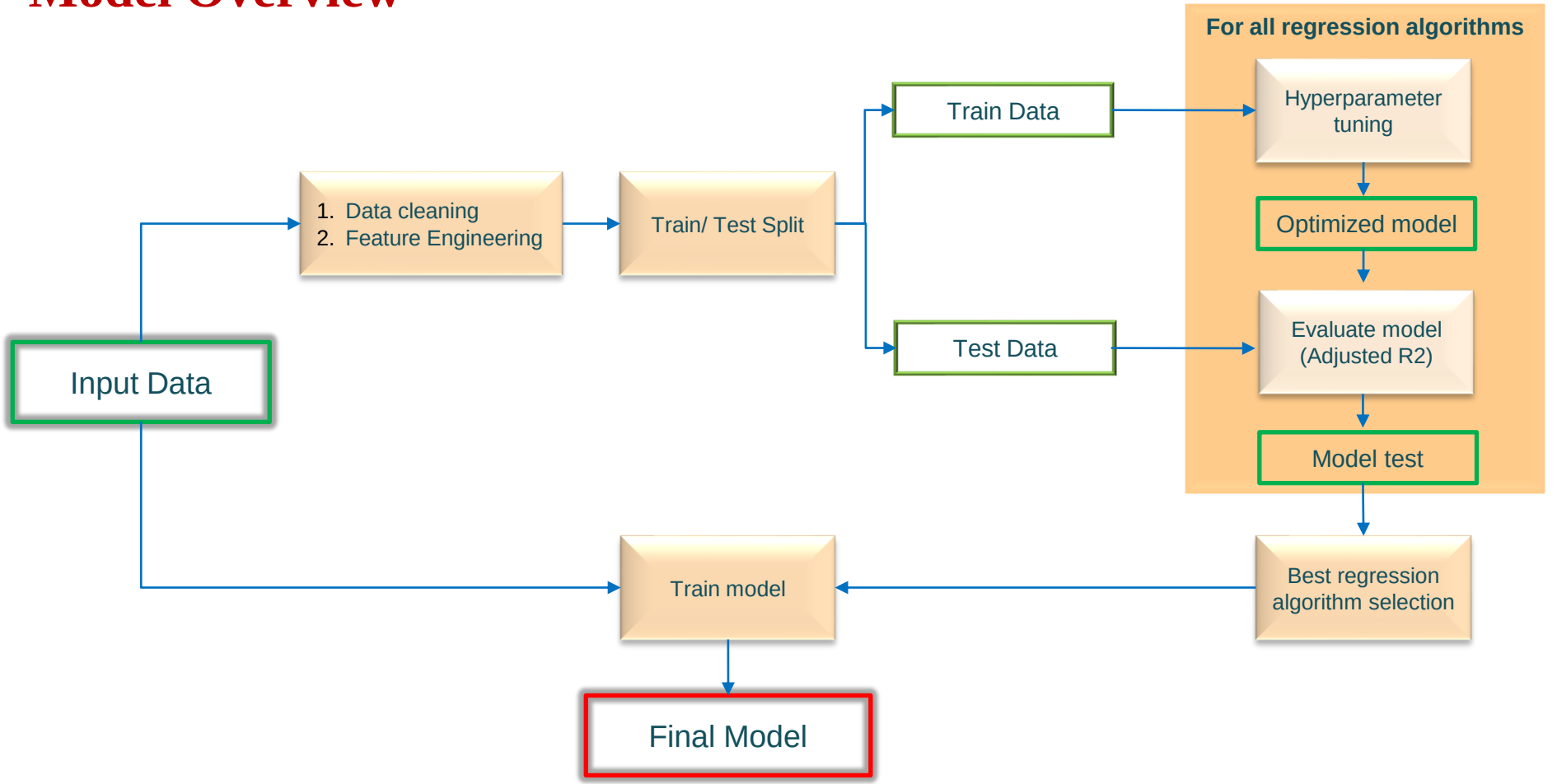
Demand on the basis of temperature



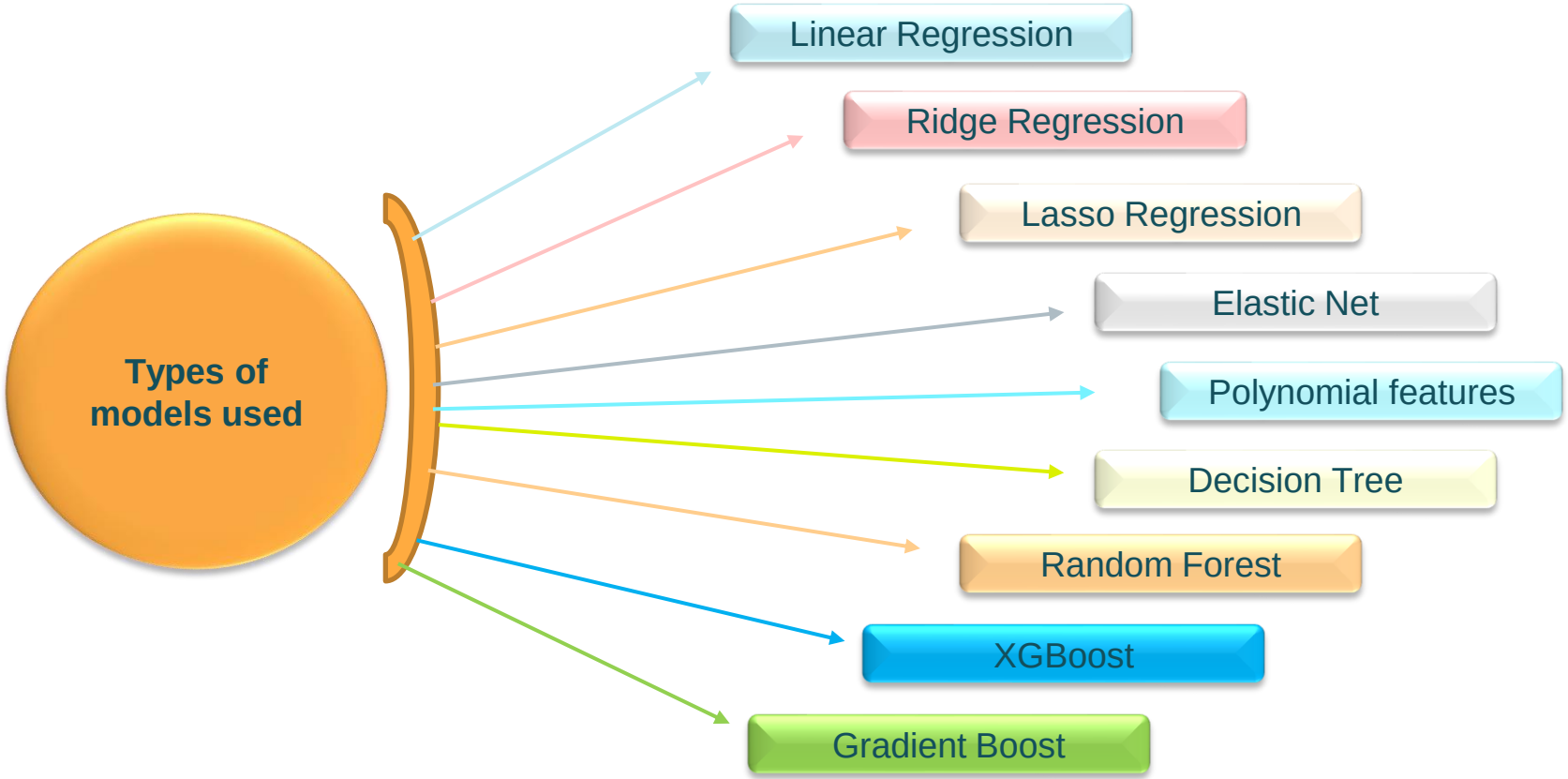
- According to the line chart, highest demand for the rented bike is seen when there is moderate temperature that is around 25°C to 35°C.

Model Overview

Model Overview



Types of models used

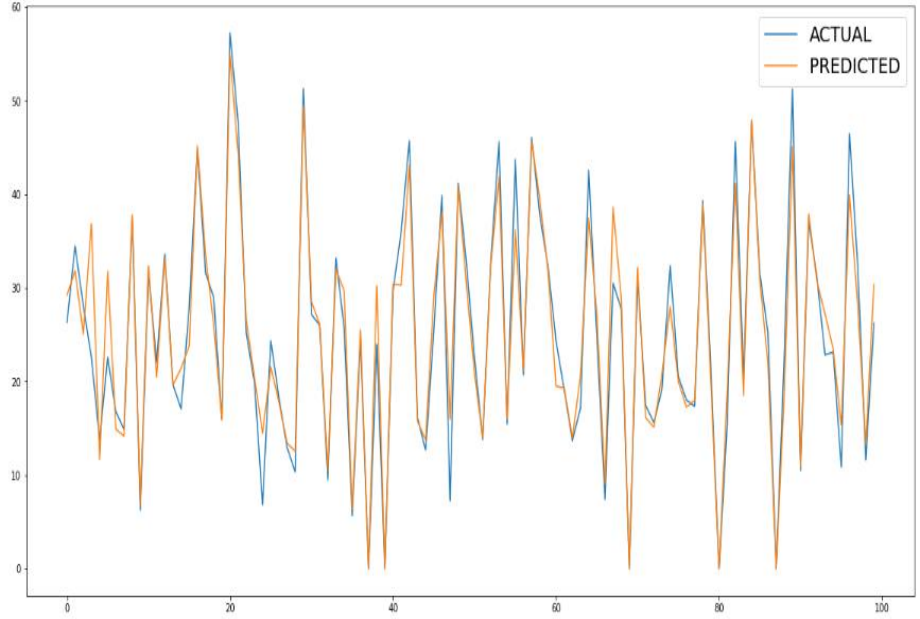


Model's Evaluation Matrices

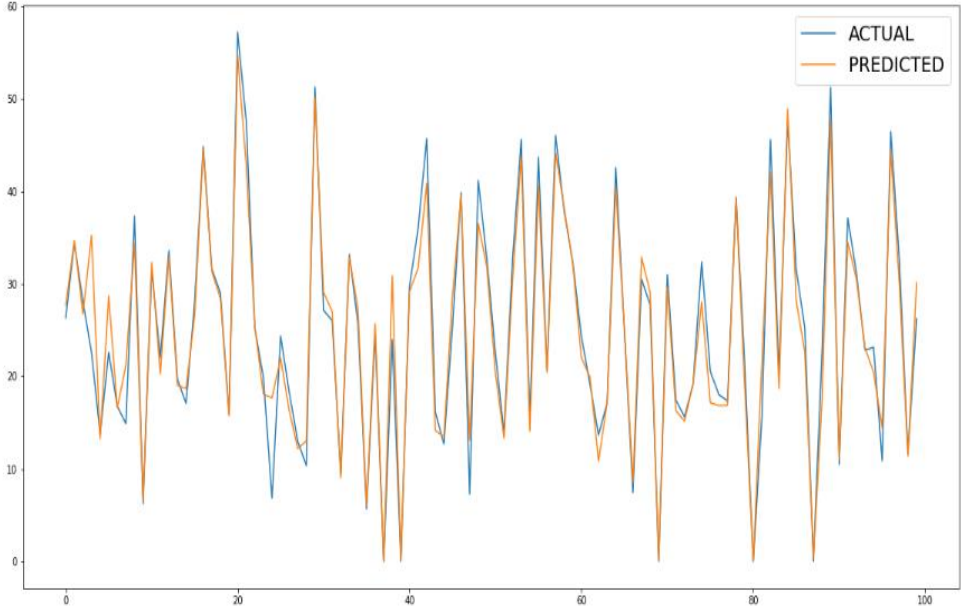
	Models	Mean_square_error	Root_Mean_square_error	R2	Adjusted_R2
0	Linear	173680.505898	416.749932	0.577557	0.574642
1	Lasso	174034.671726	417.174630	0.576695	0.573775
2	Ridge	173959.546128	417.084579	0.576878	0.573959
3	Elasticnet	174324.678423	417.522069	0.575990	0.573064
4	Polynomial	115078.893750	339.232802	0.720093	0.718162
5	Decision_Tree	23.227076	4.819448	0.848371	0.847324
6	Random_Forest	18.325110	4.280784	0.880371	0.879546
7	Gradient_Boosting	14.539109	3.813018	0.905087	0.904432
8	Xtreme_GB	11.875198	3.446041	0.922477	0.921942

Model - Actual Vs Predict

Gradient Boosting

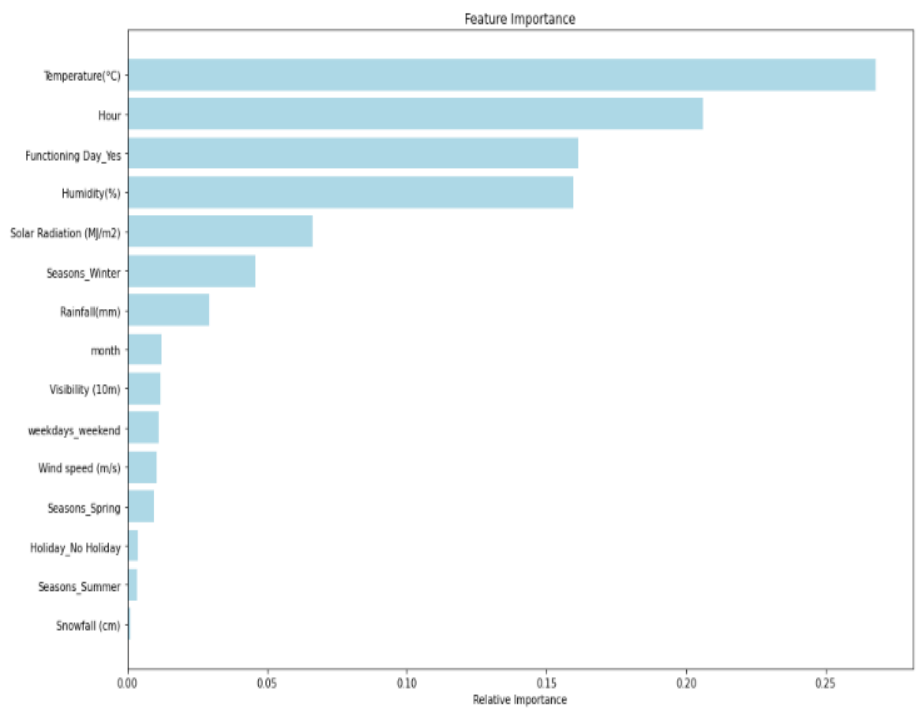


eXtreme Gradient Boosting

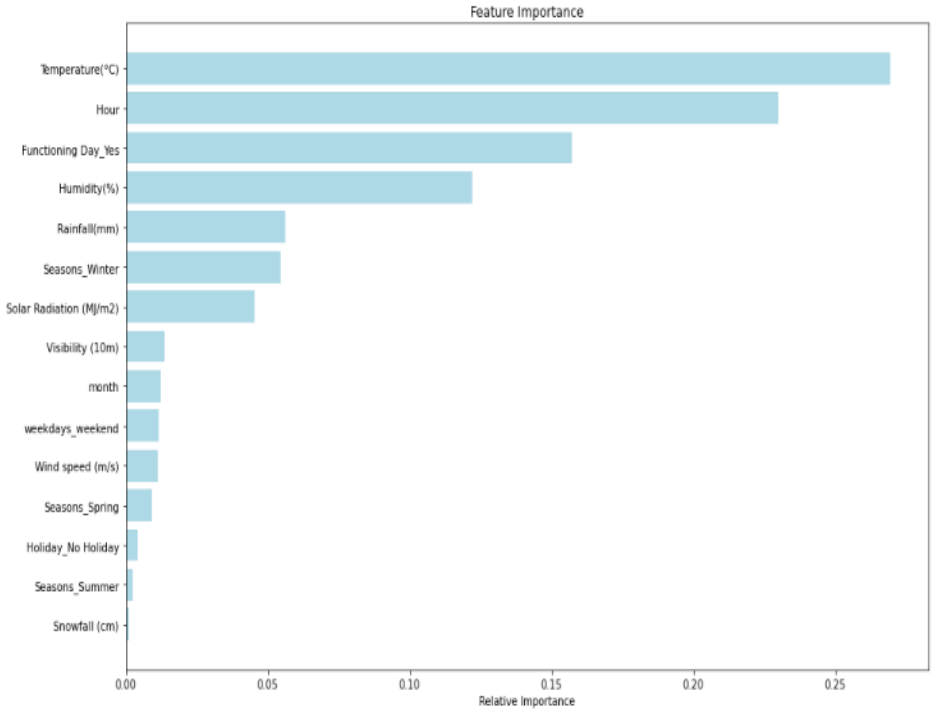


Model Features

Decision Tree



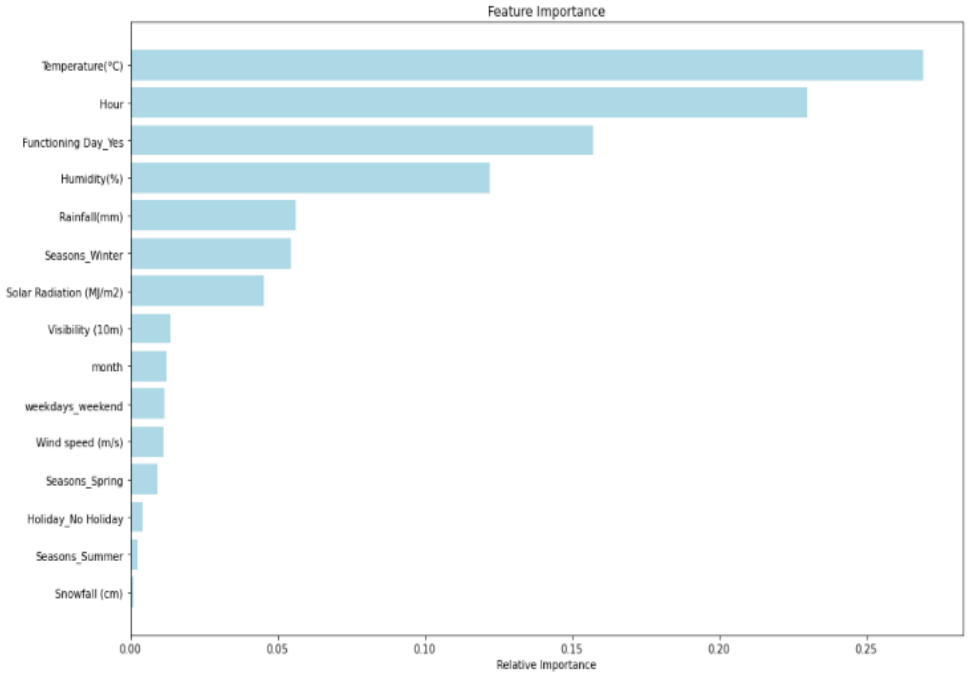
Random Forest



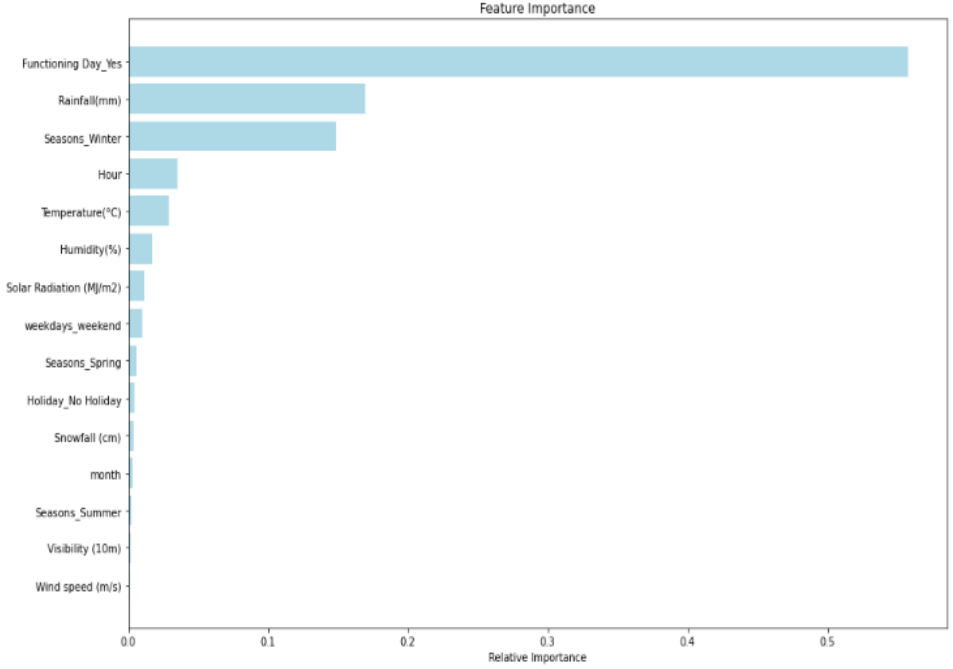
- According to Decision tree and random forest tests, temperature is the most important feature and has the highest impact on rented bike demand.

Model Features

Gradient Boosting

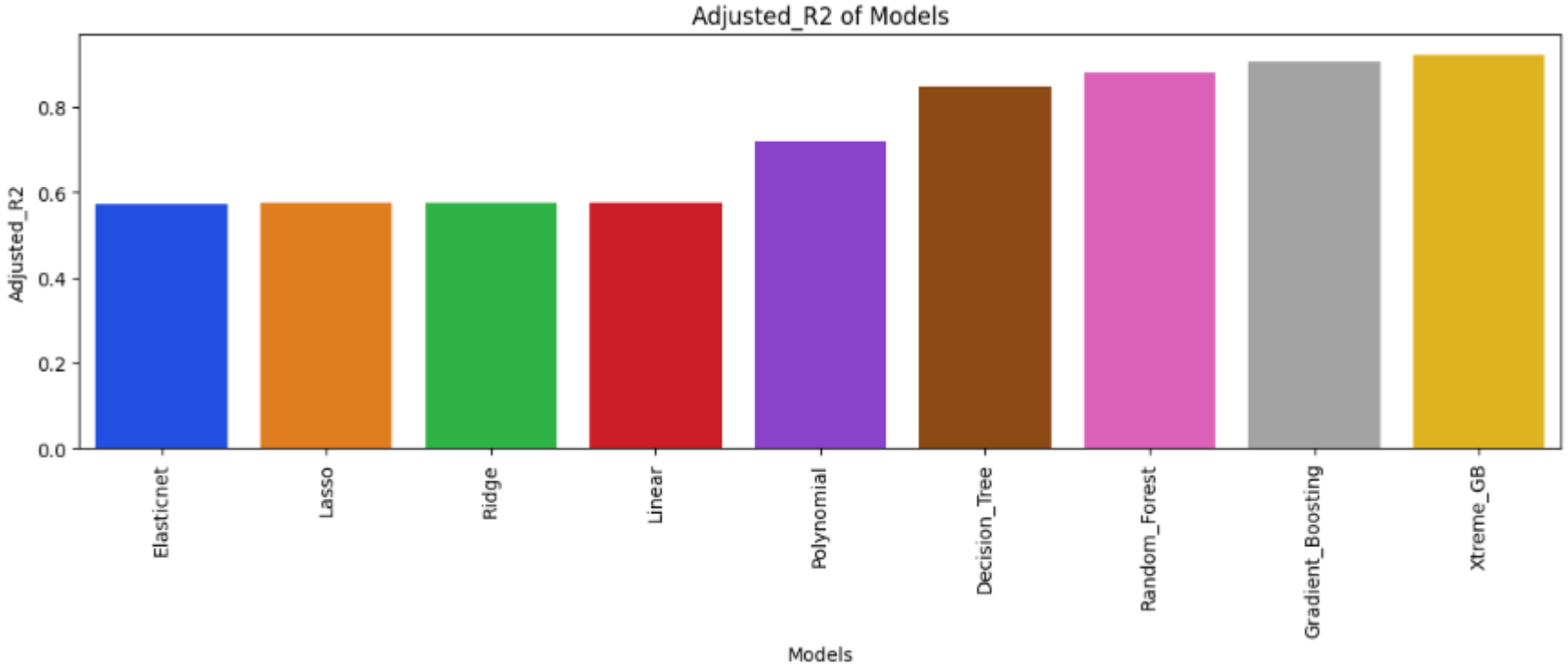


eXtreme Gradient Boosting



- According to Gradient Boosting, temperature is the most important feature and has the highest impact on rented bike demand
- As per eXtreme Gradient Boosting model, Functional day has the highest importance.

Adjusted R2 of Model's Performed



- On the basis of R2 and Adjusted-R2, Gradient Boosting model and eXtreme Gradient Boosting model have the best accuracy

Conclusion - EDA

- The highest demand for rental bikes was noted in June
- Summer is the busiest season in regards to rented bike demand
- Peak demand for rented bike is around 6pm (18:00)
- Rented bike demand is less during winter season, in comparison to other seasons
- On working day, the rental bike demand is high in comparison to holiday

Conclusion – Regression Model

- On the basis of R2 and Adjusted-R2, Gradient Boosting model and eXtreme Gradient Boosting model are best as the accuracy of these models are above 90% and also their Adjusted-R2 values are less than R2 values.
- Gradient Boosting model's R2 and Adjusted-R2 values are 0.905 and 0.904 respectively.
- eXtreme Gradient Boosting model's R2 and Adjusted-R2 values are 0.922 and 0.921 respectively.
- Linear model, Lasso model, Ridge model, and Elasticnet model have Adjusted-R2 value below 60%.

Thank You