Capstone Project

Bike Sharing Demand Prediction

Supervised Machine Learning (Regression)

Index

Seoul Bike Sharing Demand Prediction

❑ 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.

❖ Let’s Understand – “What is Bike Sharing?”

Bike sharing relies on a system of self-service bike stations. Users typically check out a bike using a membership or credit/debit card. They can then ride to their destination and park the bike in a nearby docking station. Bike share bikes are comfortable, have integrated locks and cargo baskets and usually include gearing, fenders and lights that make urban biking safe and enjoyable. Many of them are accessed by a mobile app, so you can usually find a bike nearby from wherever you are at the time. “Bike sharing offers a great chance for people to choose active transportation for short trips. This is a health benefit as well. Riding a bike is good exercise, while also getting to where you need to be.

❖ Data Pipeline

❑ Exploratory Data Analysis:-

Exploratory data analysis is an statistical way of understanding the data which is usually done in a visual way. The graphs plotted in exploratory data analysis are for better understanding of data to the analyst. After loading the dataset we performed this method by comparing our target variable that is Rented Bike Count with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable. For the current data set , Since we have to predict the number of bikes that will be rented, the best way to begin is with the variable to predict ,"count". We can stratify the "count" distribution as boxplots for the categorical variables, and draw the "count" and numeric variables in another plot.

❑ Null values :- Treatment After the data is loaded , The missing data is checked using is.na() or isnul() function . The output depicted that there was no missing values in our dataset. So our dataset does not contain any missing values.

Encoding of Categorical feature:-

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format. Categorical variables- Seasons, Functioning Day, and Holiday- were converted coded into numerical depictions to fit our Model to predict Bike rented count.

➢ **Feature Engineering**

To make the data tenable for understanding and further analysis , the data set was analysed for identifiable statistical trends and patterns. After preliminary analysis, the following steps were undertaken to transform the data into a systematically workable dataset: i) Convert the data-time attribute in proper format and we separate day, month , year and hour into separate columns so that it is easy to perform operations on the data. ii)Divide temperature, humidity and windspeed variables into categories. Doing so we can better accuracy in the model. iii)Create dummy variables for season attribute , here season variable is broken down into 3 binary variables i.e spring, summer and winter.

➢ Standardization of features

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it. The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

➢ Feature Selection

In these steps we used algorithms like Decision Tree, Random Forest, Gradient Boosting to check the results of each feature i.e., which feature is more important compared to our model and which is of less importance. Among the 17 features Temperature and Hour feature are the most important in model prediction. While almost all features are the least important, we found that they help enhance the performance of the models.

❖ Data Description

❑ Dependent variable:

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

❑ Independent variables:

• Date : year-month-day

• Hour - Hour of he day

• Temperature-Temperature in Celsius

• Humidity - %

• Windspeed - m/s

• Visibility - 10 m

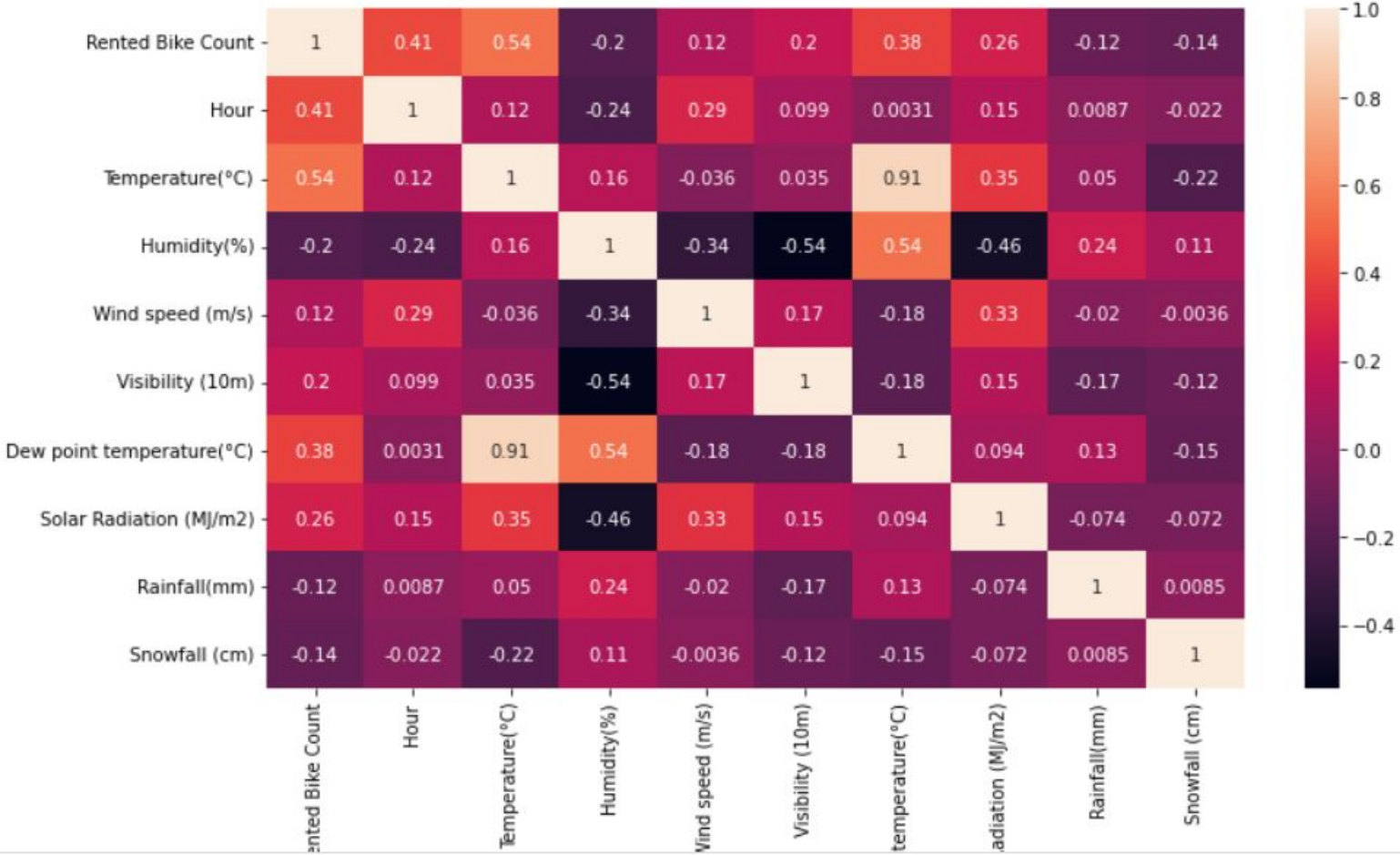
• Dew point temperature - Celsius

• Solar radiation - MJ/m2

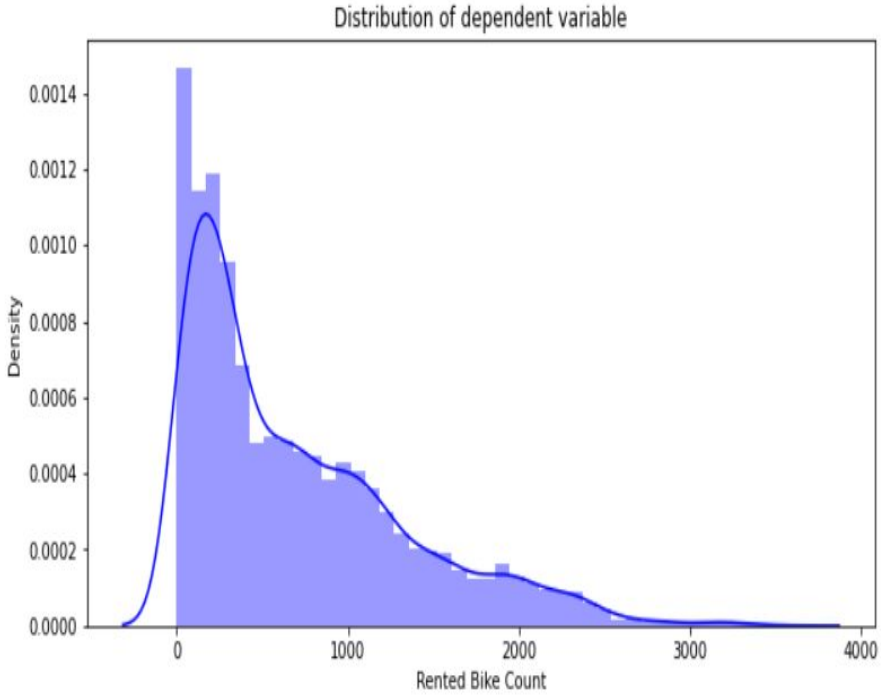
• Rainfall - mm • Snowfall - cm

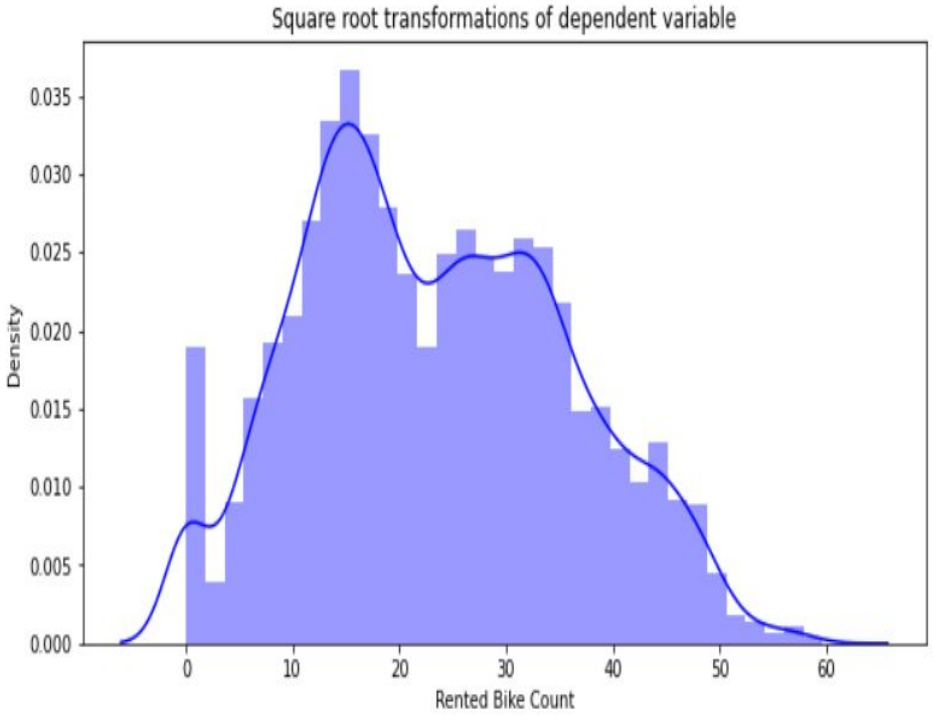
• Seasons - Winter, Spring, Summer, Autumn • Holiday – Holiday / No holiday • Functional Day – No Func (Non Functional Hours), Fun(Functional hours).

❖ EDA - Feature Correlation Graph



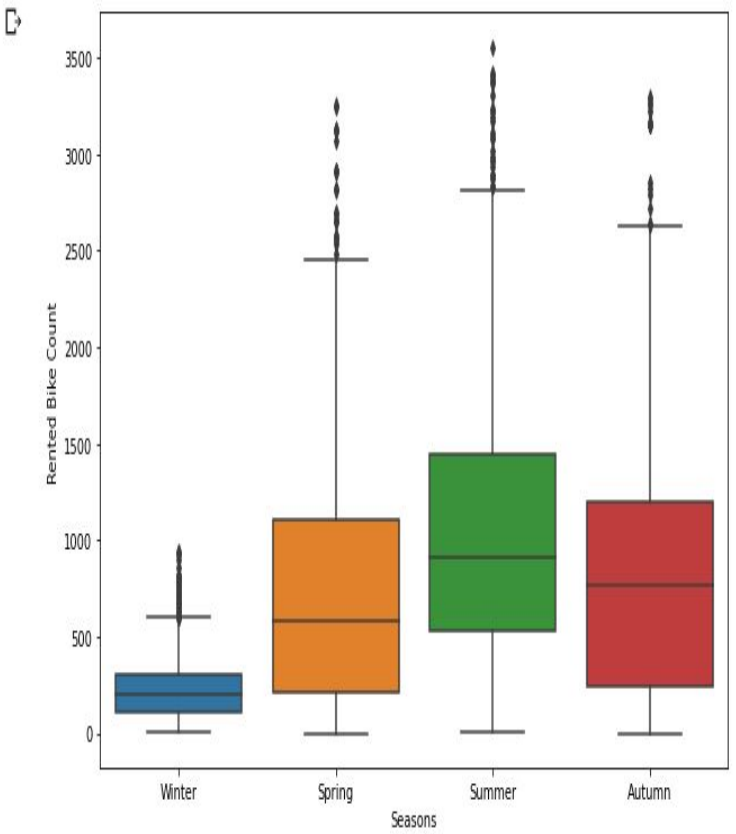
❖ Distribution Graph Of Dependent Variable





❖ EDA - Rented Bike Count By Season

In this plot, it is observed that there is high demand of bikes in Summer season (approx. 23 lacs), followed by Autumn season has second highest demand in the year (approx 18 lacs) thereafter Spring season is with approx. 16 lacs demand of bikes but Winter is the only season in the year having the lowest demand of bikes (even less than 5 lacs).

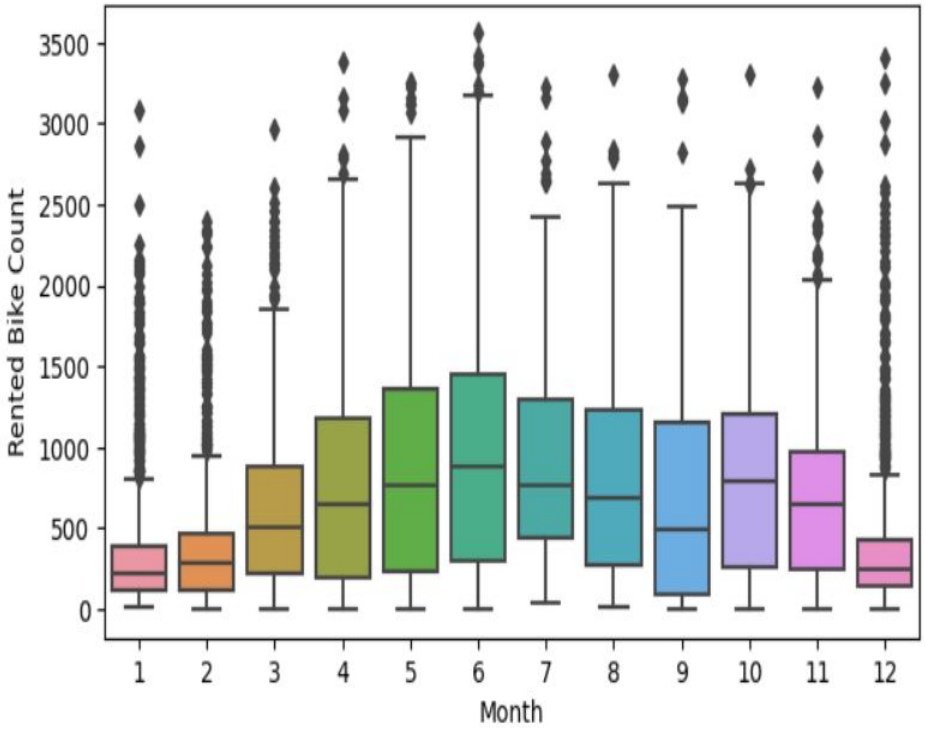


❖ EDA – Rented Bike Count By Month

❑ We can see that there is less demand of Rented bike in the month of December, January, February i.e. during winter season.

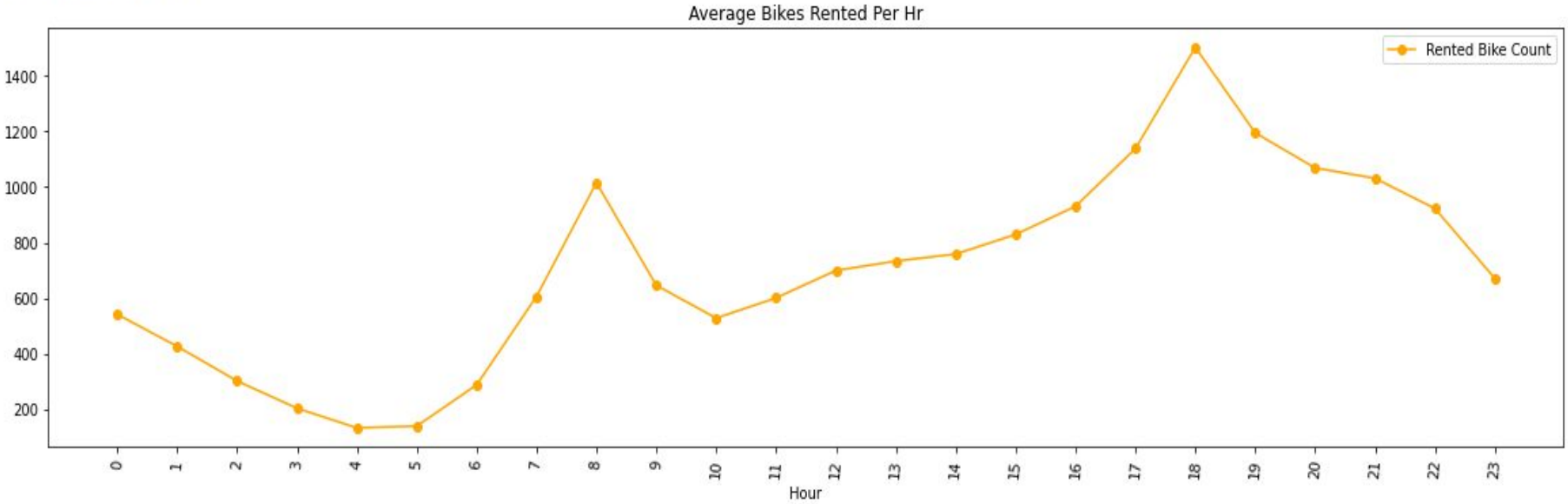
❑ Also demand of bike is increasing from March and the June is at peak level throughout the year.

❑ This trend of increase in demand get started to decrease from July where January is at lowest position in terms of bike demand.

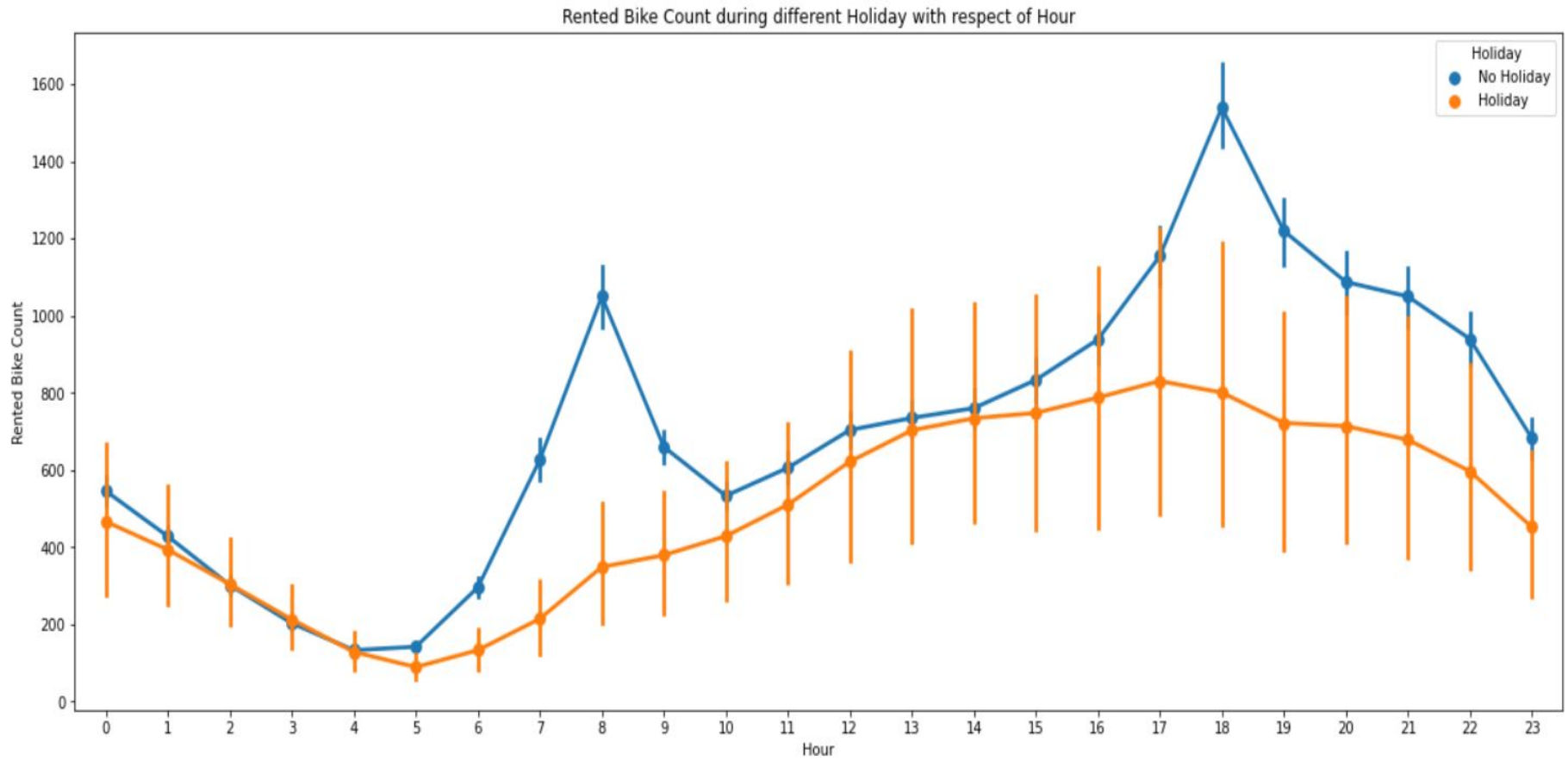


❖ EDA – Rented Bike Count By Hour

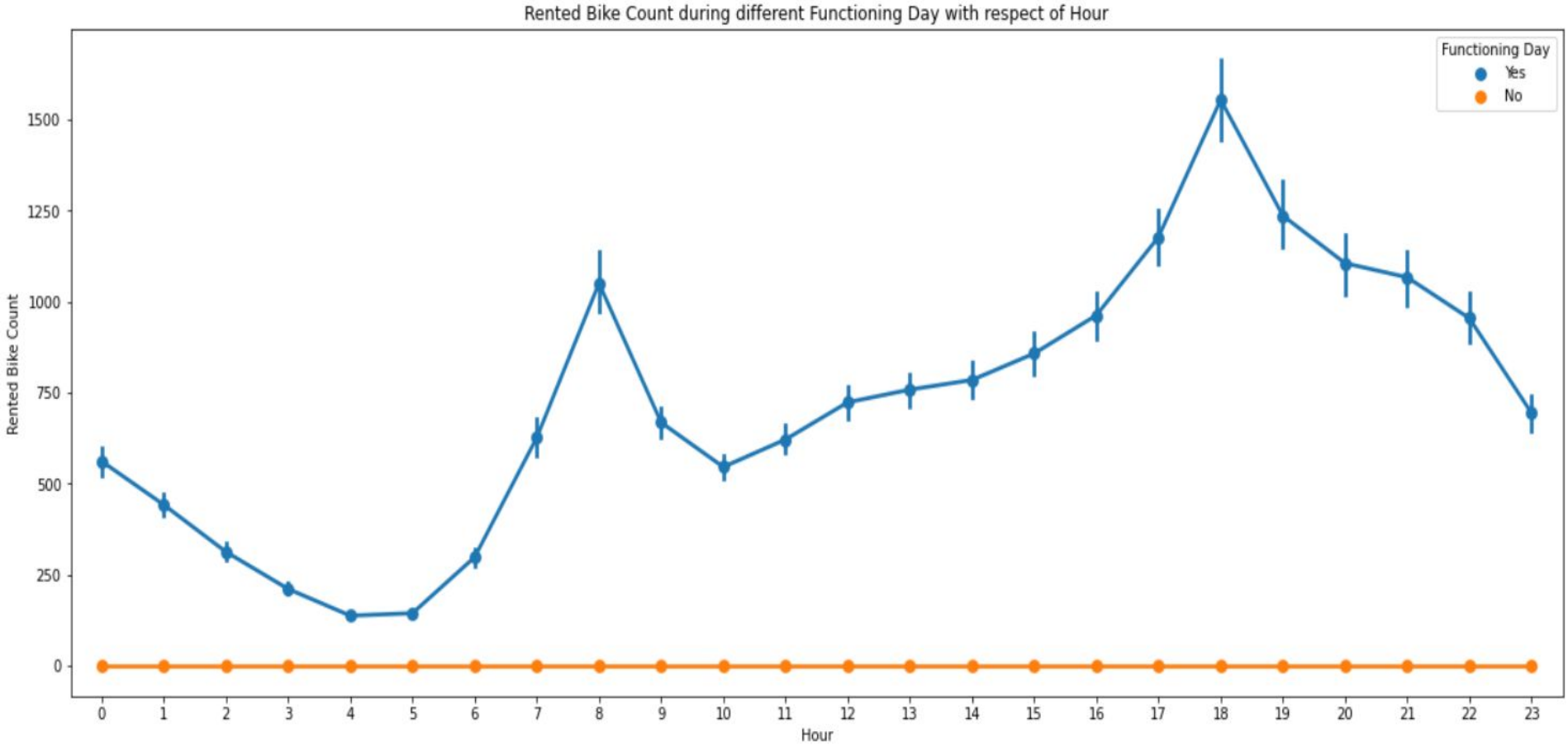
There is a surge of high demand in the morning 8 AM and in evening 6 PM as the people might be going to their work at morning 8 AM and returning from their work at the evening 6 PM.



❖ EDA - Rented Bike Count By Holiday And NO Holiday



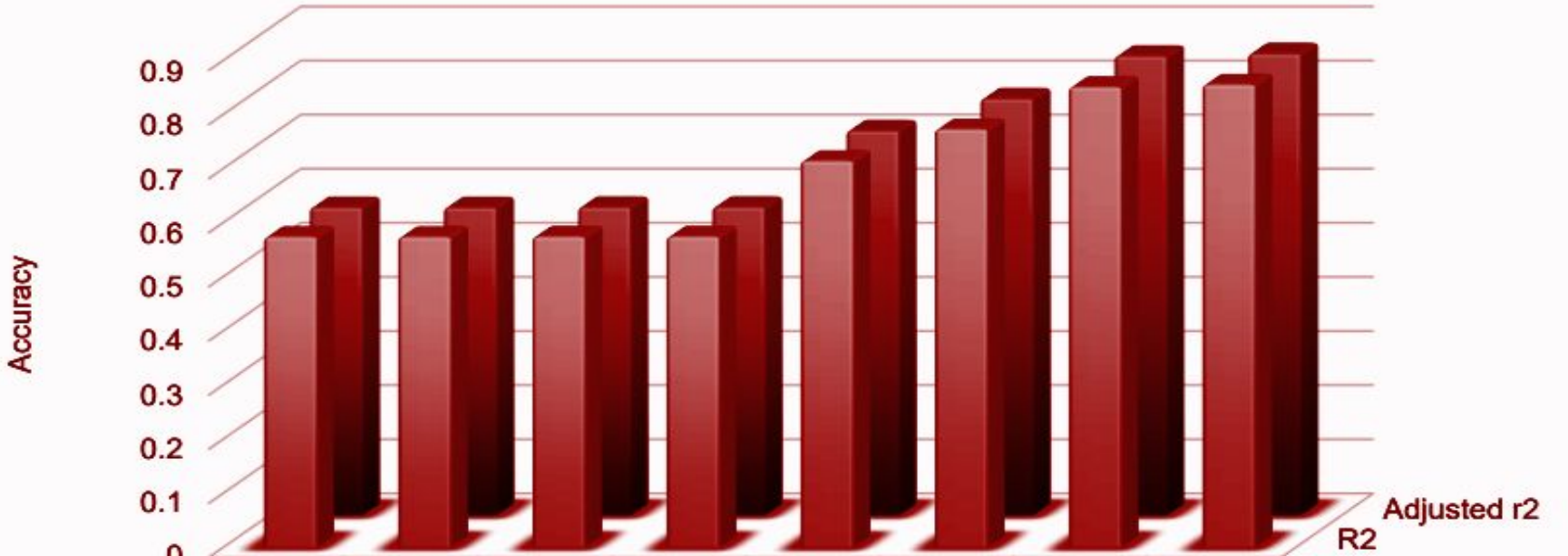
❖ EDA - Rented Bike Count By Functioning Day



❖ Model's Performed

- ✓ Linear Regression
- ✓ Lasso Regression
- ✓ Ridge Regression
- ✓ Polynomial Regression
- ✓ Decision Tree
- ✓ Random Forest
- ✓ Gradient Boosting

❖ Model Performance With R2 And Adjusted R2



	Linear	Lasso	Ridge	Elastic net	Ploynomial	Decision Tree	Random Forest	Gradient Boosting
R2	0.5741	0.5739	0.5745	0.5746	0.7145	0.7732	0.8512	0.8557
Adjusted r2	0.5699	0.5697	0.5703	0.5705	0.7117	0.7709	0.8497	0.8542

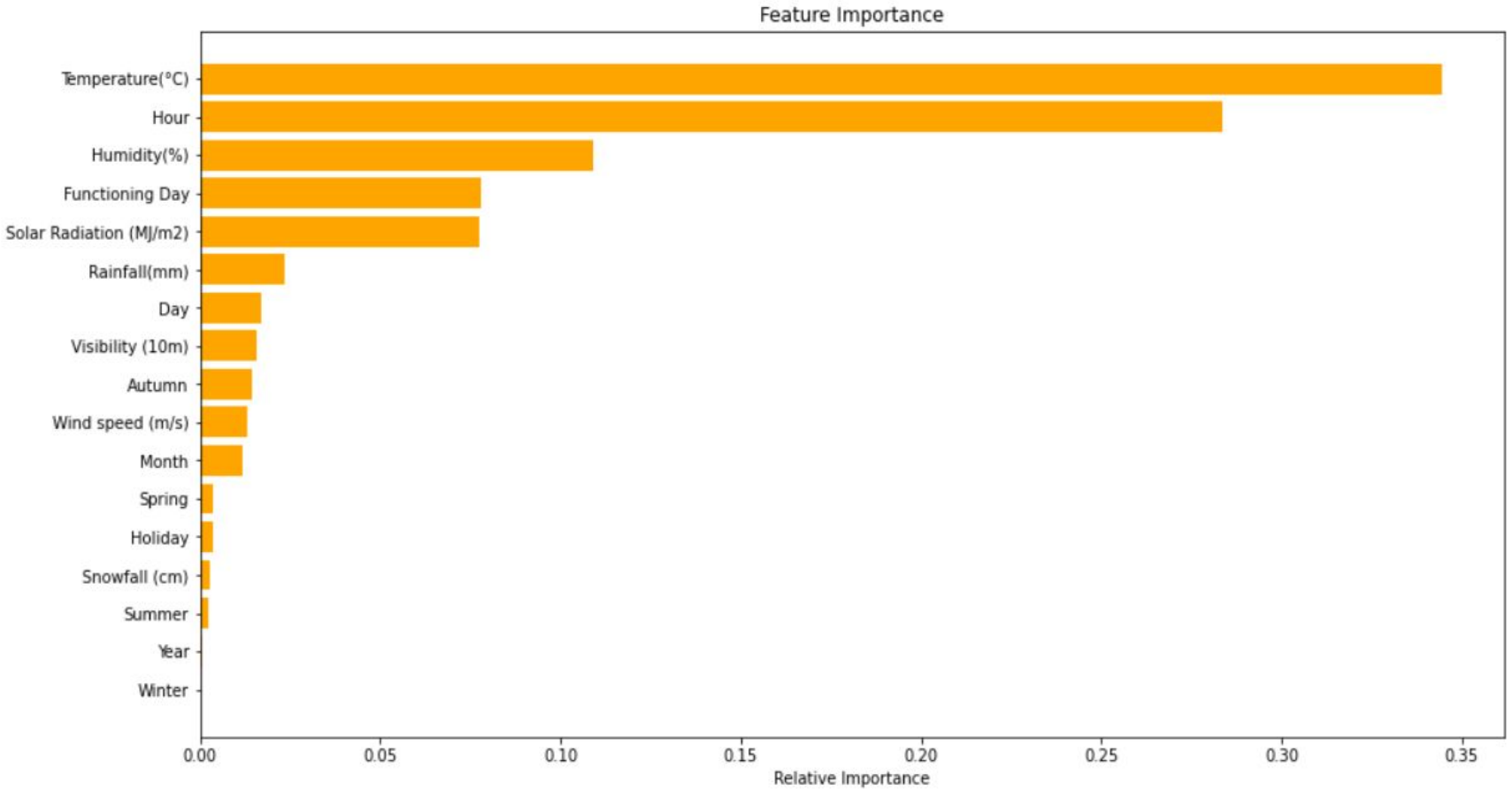
❖ Model Validation And Selection

❑ Observation 1: As seen in the previous slide, Linear Regression, Lasso, Ridge, Elastic net is not giving great results.

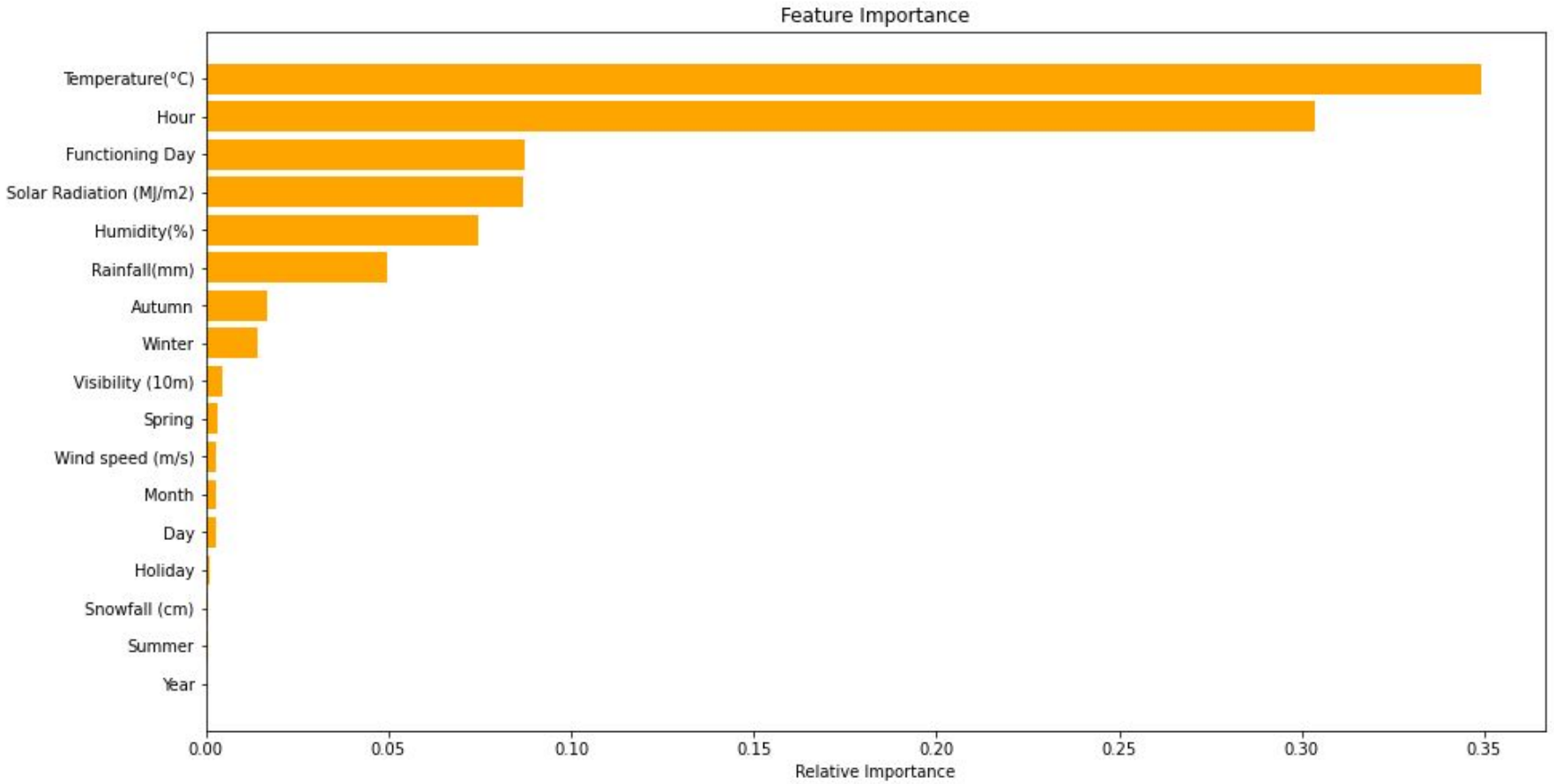
❑ Observation 2: R2 and Adjusted R2 are improving from Polynomial Regression and Decision Tree.

❑ Observation 3: Random forest & GBR have best performed equally in terms of R2 and Adjusted R2. We can use either Random Forest or Gradient Boosting model for the bike rental stations.

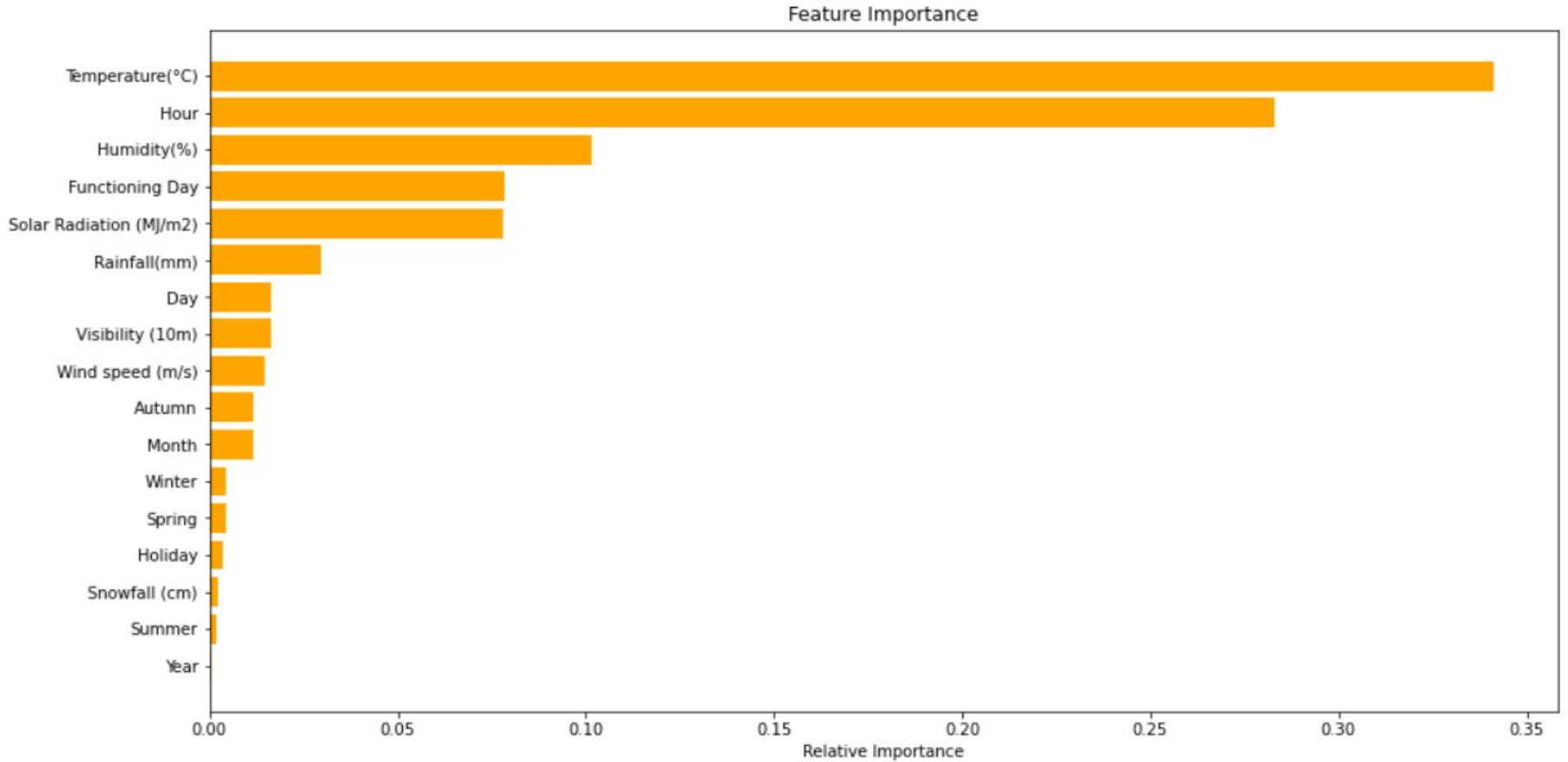
❖ Feature Importance For Decision Tree



❖Feature Importance For Random Forest



❖ Feature Importance For Gradient Boosting



❖ Conclusion

1.In holiday or non-working days there is demands in rented bikes.
2.There is a surge of high demand in the morning 8AM and in evening 6PM as the people might be going to their work at morning 8AM and returning from their work at the evening 6PM.
3.People preferred more rented bikes in the morning than the evening.
4.When the rainfall was less, people have booked more bikes except some few cases.
5.The Temperature, Hour & Humidity are the most important features that positively drive the total rented bikes count.
6.After performing the various models the Random Forest and Gradient Boosting 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 Random Forest and Gradient Boosting model.
7.R2 value for Random Forest and Gradient Boosting are 0.851 and 0.855 respectively.
8.We can use either Random Forest or Gradient Boosting model for the bike rental stations.

❖Challenges

❑ A huge amount of data needed to be deal while doing the project which is quite an important task and also even small inferences need to be kept in mind.

❑ As dataset was quite big enough which led more computation time.

❑ Handling the numerical and categorical data to build high accuracy model.

THANK YOU