

Capstone Project

Bike Sharing Demand Prediction

By Nivya T

Bike Sharing System

A bicycle-sharing system, bike share program, public bicycle scheme, or public bike share (PBS) scheme, is a shared transport service in which bicycles are made available for shared use to individuals on a short term basis for a price or free.

Predicting bike sharing demand can help bike sharing companies to allocate bikes better and ensure a more sufficient circulation of bikes for customers.



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.



Objectives

1. Prediction of bike count required at each hour for the stable supply of rental bikes in bike sharing system.
2. Gathering information regarding the factors that affect this prediction the most.



Steps Involved

- 1. Exploring the data:** Analyzing the features and target variable, checking for null values and duplicates, plotting the distribution of target variable etc.
- 2. EDA:** Treating numerical and categorical features separately, VIF Analysis, Encoding, Outlier detection etc.
- 3. Preprocessing of data:** Train test split, Transformation, Scaling etc.
- 4. Creating models:** Create different models and evaluate them using different metrics.

Data Summary

The shape of the dataset is (8760,14)

Data

Features

Target
Variable

Numeric:

- 1.Hour
- 2.Temperature
- 3.Humidity
- 4.Wind speed
- 5.Visibility
- 6.Dew point temperature
- 7.Solar radiation
- 8.Rainfall
- 9.Snowfall

Categorical:

- 1.Season
- 2.Holiday
- 3.Functioning Day

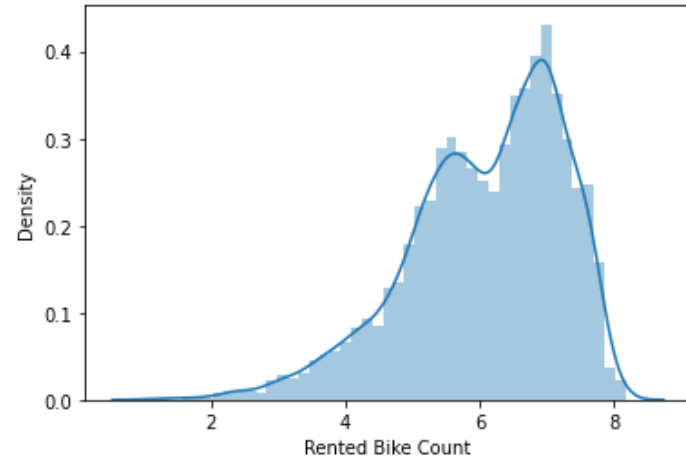
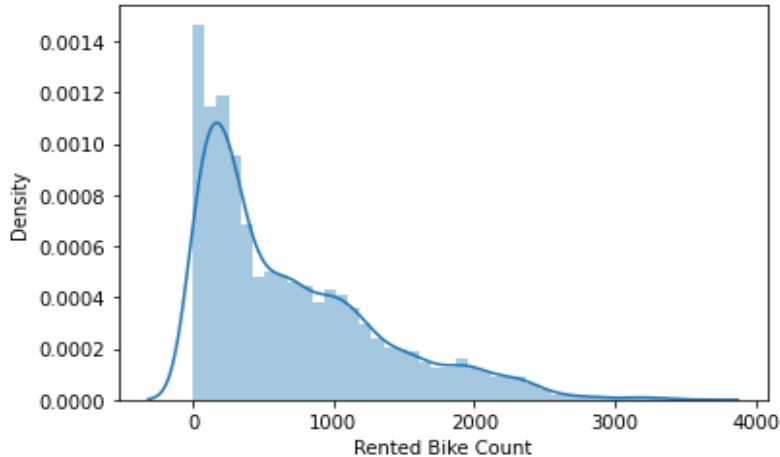
Rented bike
count

Define Dependent Variable

In this project, the dependent variable is 'Rented bike count', the prediction of which gives us the exact number of bikes required per hour in order to reduce the waiting time.

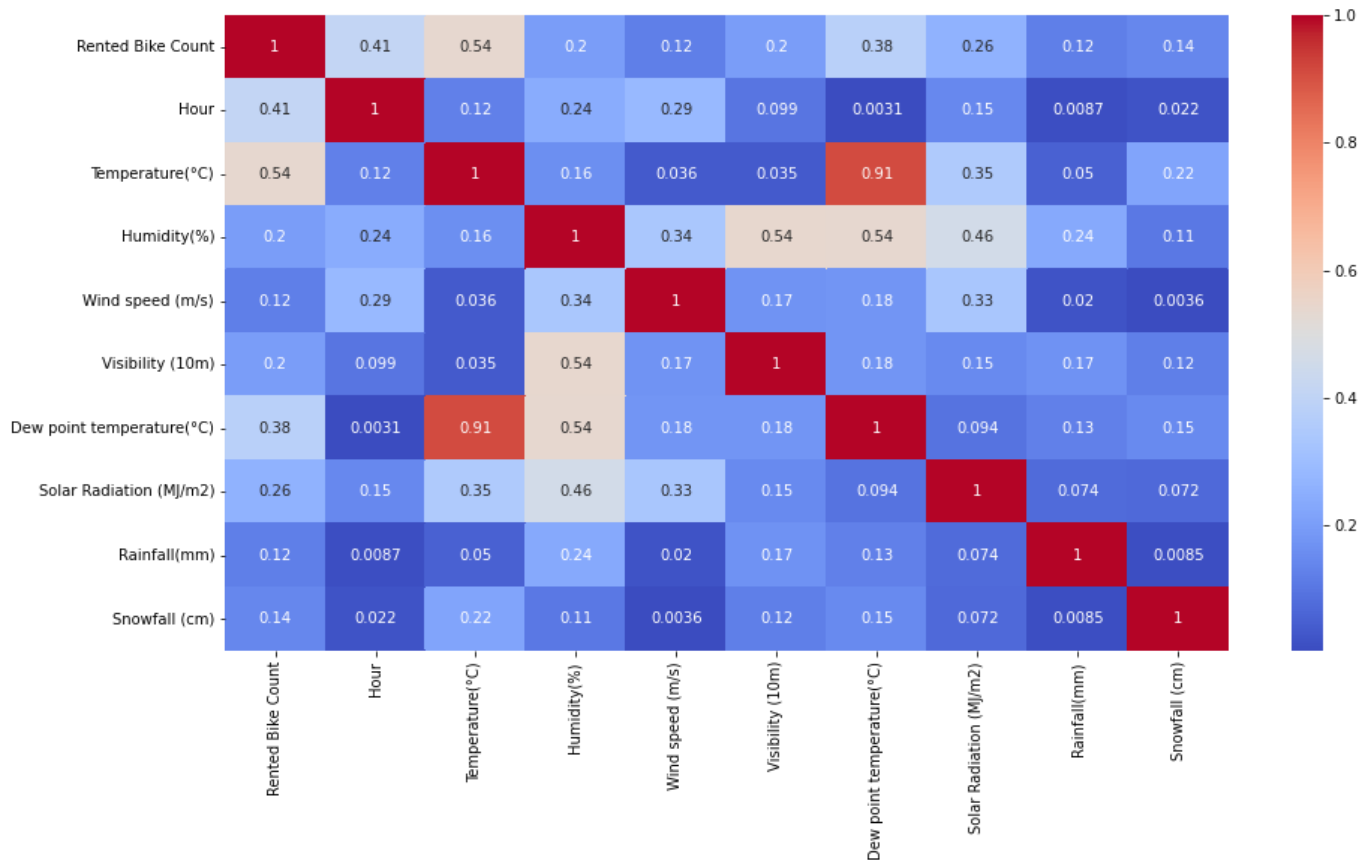


Plotting the distribution of dependent variable



The distribution of the dependent variable is skewed. Therefore we use $\log(1+x)$ transformation.

Collinearity Between Variables



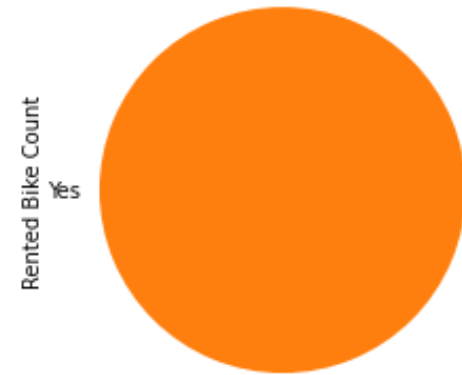
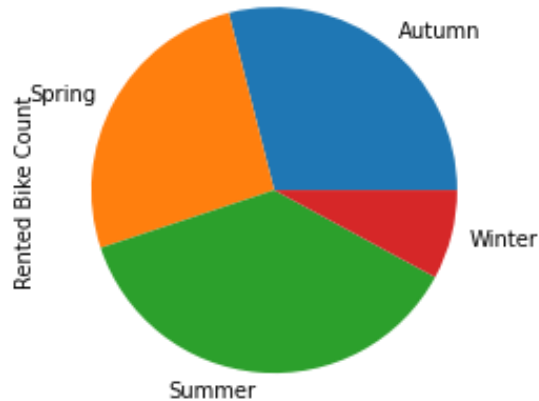
VIF Analysis

| | 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 |



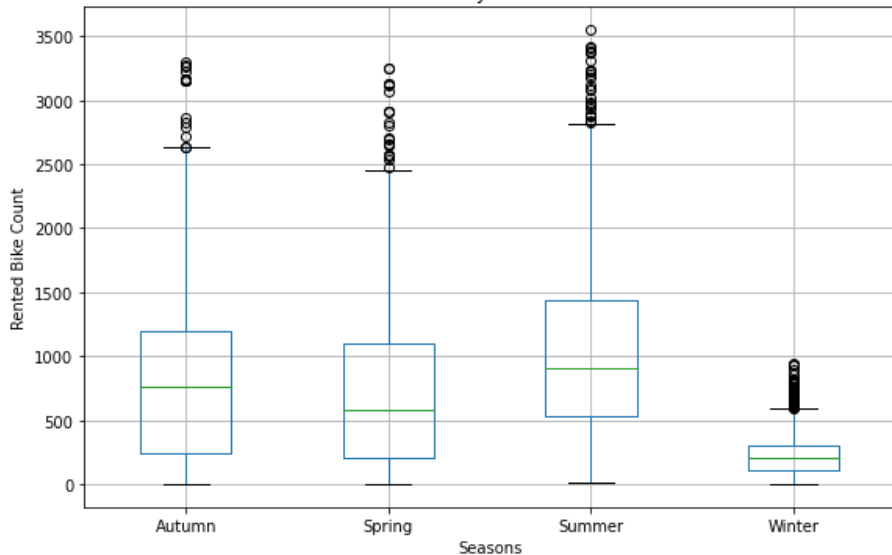
VIF is under 5 for all the variables. Therefore we can neglect the chances of multicollinearity.

Bike Demand Based on Seasons, Holiday and Functioning Day

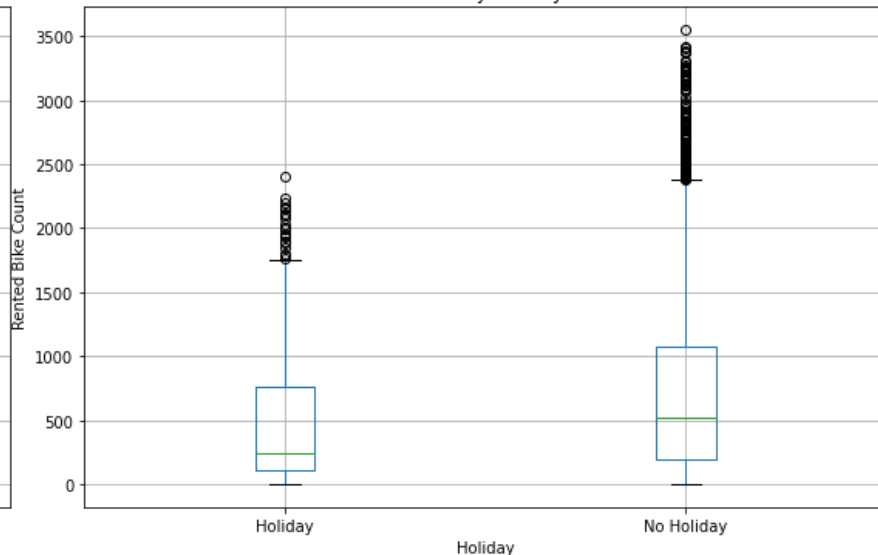


Outlier Detection: Box Plot

Boxplot grouped by Seasons
Label by Seasons



Boxplot grouped by Holiday
Label by Holiday



Preparing Dataset for Modelling

| hour | Temperature(°C) | Humidity(%) | Wind speed (m/s) | Visibility (10m) | Dew point temperature(°C) | Solar Radiation (MJ/m2) | Rain |
|------|-----------------|-------------|------------------|------------------|---------------------------|-------------------------|------|
| 0 | -5.2 | 37 | 2.2 | 2000 | -17.6 | 0.0 | |
| 1 | -5.5 | 38 | 0.8 | 2000 | -17.6 | 0.0 | |
| 2 | -6.0 | 39 | 1.0 | 2000 | -17.7 | 0.0 | |
| 3 | -6.2 | 40 | 0.9 | 2000 | -17.6 | 0.0 | |
| 4 | -6.0 | 36 | 2.3 | 2000 | -18.6 | 0.0 | |

Task :Regression
Train Set :(7008, 11)
Test Set: : (1752,11)



Linear Regression

The scores obtained while performing Linear Regression is:

• ty

Lasso Regression

| | Metric | Train Score | Test Score |
|---|--------|-------------|------------|
| 0 | MAE | 287.46 | 279.09 |
| 1 | MSE | 202234.96 | 195153.97 |
| 2 | RMSE | 449.71 | 441.76 |
| 3 | r2 | 0.51 | 0.53 |
| 4 | adj_r2 | 0.51 | 0.53 |

| | Metric | Train Score | Test Score |
|---|--------|-------------|------------|
| 0 | MAE | 287.44 | 279.07 |
| 1 | MSE | 202203.58 | 195114.04 |
| 2 | RMSE | 449.67 | 441.72 |
| 3 | r2 | 0.51 | 0.53 |
| 4 | adj_r2 | 0.51 | 0.53 |

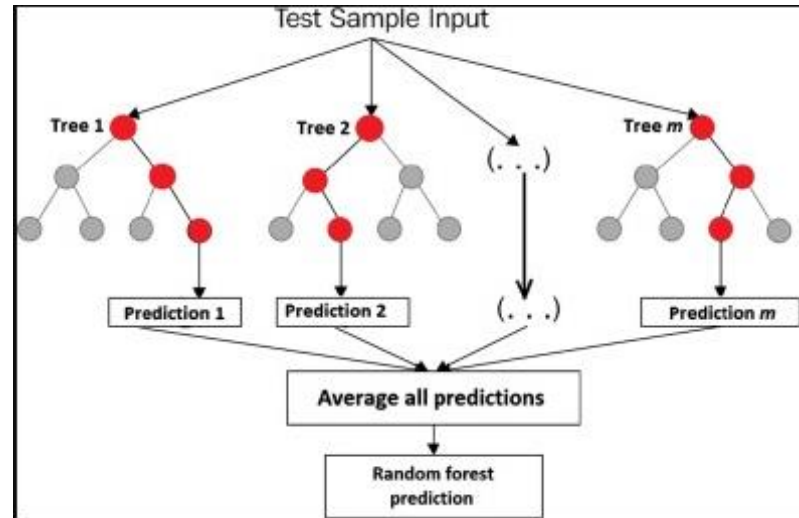
Ridge Regression

| | Metric | Train Score | Test Score |
|---|--------|-------------|------------|
| 0 | MAE | 287.45 | 279.08 |
| 1 | MSE | 202213.50 | 195127.98 |
| 2 | RMSE | 449.68 | 441.73 |
| 3 | r2 | 0.51 | 0.53 |
| 4 | adj_r2 | 0.51 | 0.53 |

Random Forest

The scores obtained while performing Random Forest Regression is:

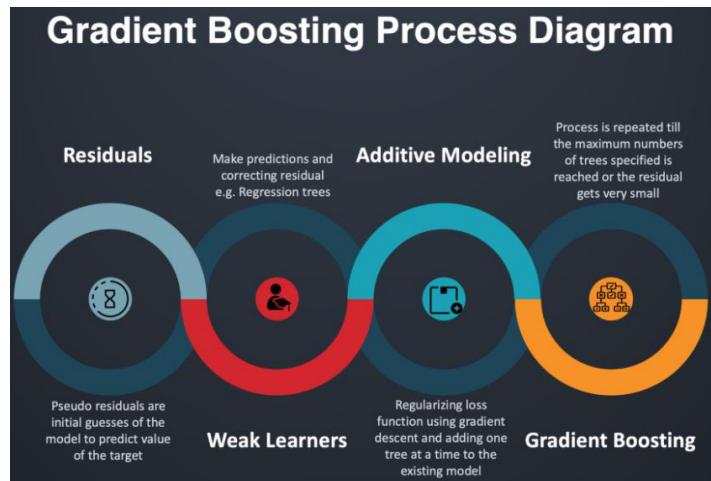
| | Metric | Train Score | Test Score |
|---|--------|-------------|------------|
| 0 | MAE | 53.83 | 140.68 |
| 1 | MSE | 9083.60 | 55662.80 |
| 2 | RMSE | 95.31 | 235.93 |
| 3 | r2 | 0.98 | 0.87 |
| 4 | adj_r2 | 0.98 | 0.87 |



Gradient Boosting Machine

The scores obtained while performing Gradient Boosting Machine is:

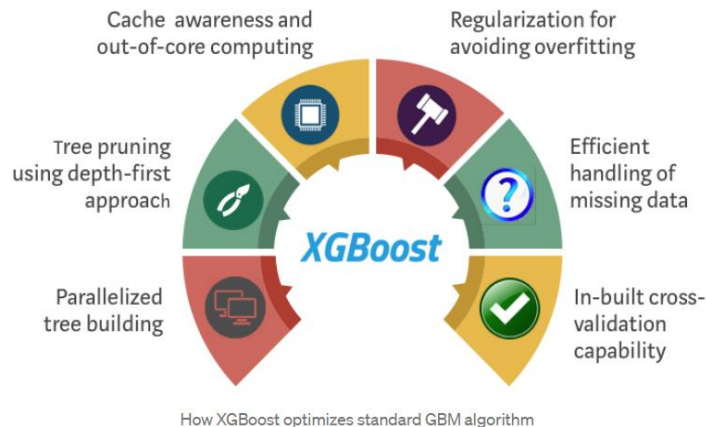
| | Metric | Train Score | Test Score |
|---|--------|-------------|------------|
| 0 | MAE | 196.04 | 196.02 |
| 1 | MSE | 104082.42 | 108884.17 |
| 2 | RMSE | 322.62 | 329.98 |
| 3 | r2 | 0.75 | 0.74 |
| 4 | adj_r2 | 0.75 | 0.74 |



XGBoost

The scores obtained while performing XGBoost is:

| | Metric | Train Score | Test Score |
|---|--------|-------------|------------|
| 0 | MAE | 120.88 | 146.87 |
| 1 | MSE | 39415.53 | 59618.05 |
| 2 | RMSE | 198.53 | 244.17 |
| 3 | r2 | 0.91 | 0.86 |
| 4 | adj_r2 | 0.90 | 0.86 |



Scores After Cross Validation and Hyperparameter Tuning

| | Model | Train MAE | Test MAE | Train MSE | Test MSE | Train RMSE | Test RMSE | Train r2 | Test r2 | Train adj r2 | Test adj r2 |
|---|------------------|-----------|----------|-----------|-----------|------------|-----------|----------|---------|--------------|-------------|
| 0 | Linear | 287.44 | 279.07 | 202203.58 | 195114.04 | 449.67 | 441.72 | 0.51 | 0.53 | 0.51 | 0.53 |
| 1 | Lasso | 287.44 | 279.07 | 202203.58 | 195114.04 | 449.67 | 441.72 | 0.51 | 0.53 | 0.51 | 0.53 |
| 2 | Ridge | 287.45 | 279.08 | 202213.50 | 195127.98 | 449.68 | 441.73 | 0.51 | 0.53 | 0.51 | 0.53 |
| 3 | Random Forest CV | 176.08 | 179.53 | 84522.73 | 89452.62 | 290.73 | 299.09 | 0.80 | 0.79 | 0.80 | 0.78 |
| 4 | GBMCV | 125.96 | 142.83 | 42643.99 | 56999.00 | 206.50 | 238.74 | 0.90 | 0.86 | 0.90 | 0.86 |
| 5 | XGboost CV | 75.20 | 139.62 | 16533.00 | 54594.92 | 128.58 | 233.66 | 0.96 | 0.87 | 0.96 | 0.87 |

Observations

- Comparing the R^2 score of all the models, one can see that XGBoost performs better.
- R^2 score in Linear Regression is 0.51 for the train data and 0.53 for the test data. Clearly, Linear Regression model fails in this case.
- Gradient Boosting Machine has a test accuracy of 86% making it the second-best model.
- Random Forest is also found to perform well on the data.



Best Parameters:

- **Random Forest:**

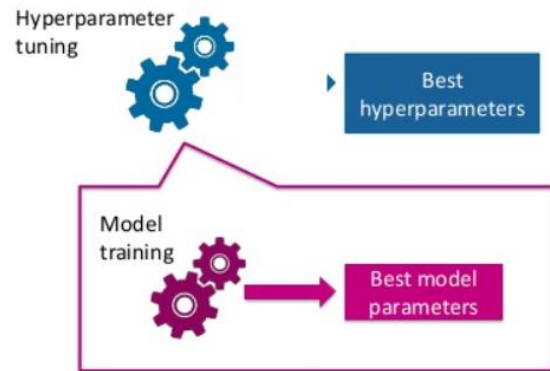
- max_depth': 8
- 'min_samples_leaf': 40
- 'min_samples_split': 50
- 'n_estimators': 100

- **GBM:**

- max_depth=8
- min_samples_leaf=40
- min_samples_split=50
- n_estimators=80

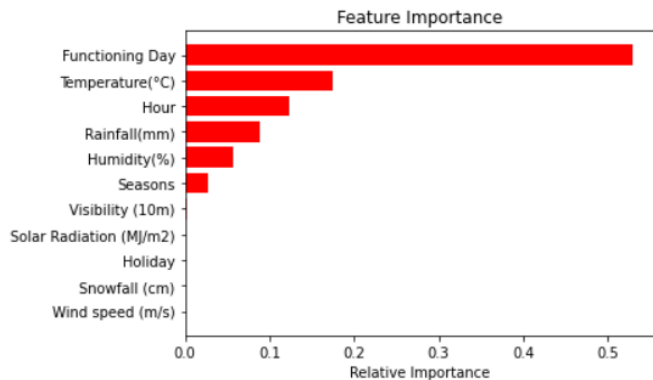
- **XGBoost:**

- 'eval_metric': 'rmse',
- 'max_depth': 6,
- 'n_estimators': 500,
- 'objective': 'reg:squarederror'

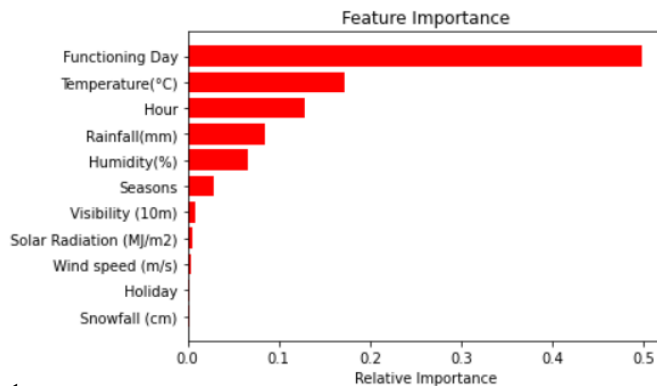


Feature Importance

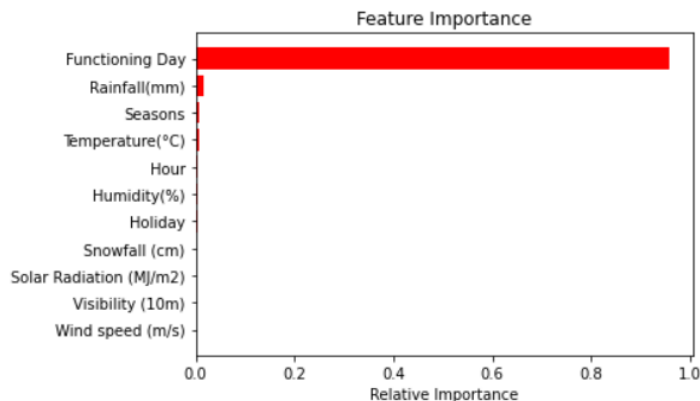
Random Forest



GBM



XGBoost



Observations

- The feature 'Functioning Day' has the highest impact on the dependent variable 'Rented Bike Count'.
- In Random Forest and GBM, 'Temperature' is making an impact while 'Rainfall' is the second most important factor in XGBoost.
- Random Forest and GBM give importance to 6-7 features while XGBoost considers only the top 3-4 features and almost neglects the rest.



Conclusion

This project focus on predicting the bike-sharing demand using Seoul Dataset.

The results show that XGBoost, Random Forest and GBM algorithms perform well on the dataset whereas Linear Regression fails in this case. Among these three models, XGBoost is found to have better performance.

Therefore XGBoost algorithm can be used as an effective tool for Bike Sharing Demand Prediction.

We did a variable analysis to identify the hidden relationship between the variables. For all the models, functioning day, temperature, and rainfall were ranked as the most influential variables to predict the rental bike demand at each hour.

This project identifies the curious relationship among the variables which can directly impact the dependent variable, 'Rented Bike Count'.

This project will be helpful for the company to predict the hourly bike demand and enrich the user experience.

Thank You