Capstone Project

Supervised ML - Regression

Bike Sharing Demand Prediction

By - Dishant Toraskar

CONTENT

1. Problem Definition
2. Data Description
3. Exploratory Data Analysis
4. Feature Engineering
5. Hyperparameter Tuning
6. Model Building & Evaluation -
   i. Linear Regression
   ii. Random Forest Regressor
   iii. Gradient Boosting Regressor
7. Feature Importance
8. Conclusion

Problem Definition

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.

Data Description

The dataset contains 8760 rows and 14 features including target variable.

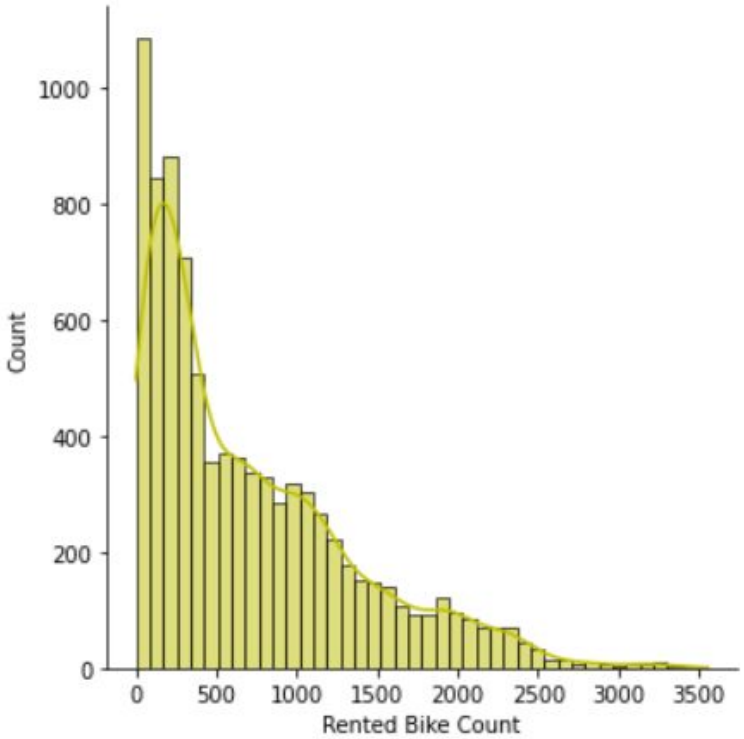
Target feature: Rented Bike Count (Per Hour).

Numerical Features: Hour, Temperature, Humidity, Windspeed, Visibility, Dew point temperature, Solar radiation, Rainfall, Snowfall.

Categorical Features: Holiday, Functional Day.

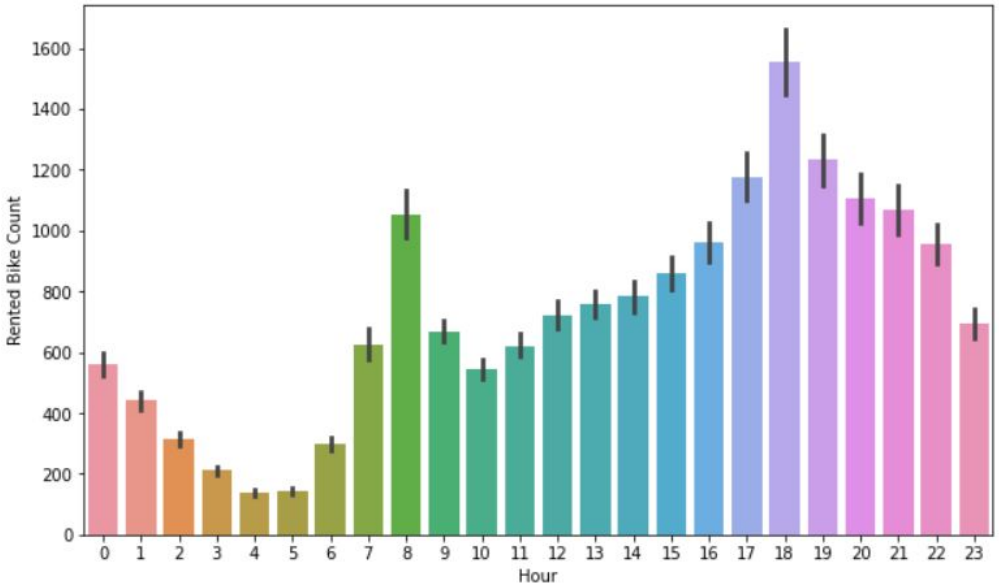
Exploratory Data Analysis

Target Feature - Rented Bike Count



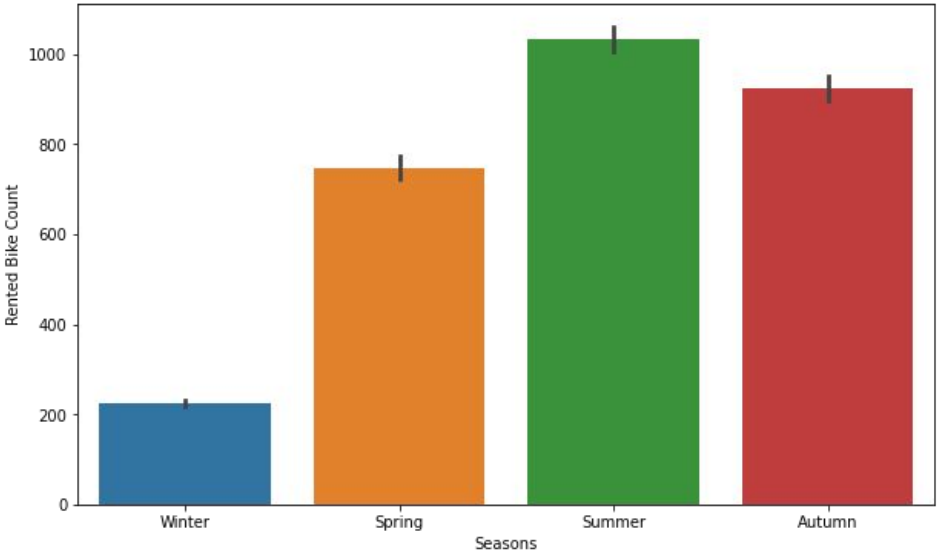
We can observe that our target feature i.e., **Rented Bike Count** is not distributed normally which can be the problem for linear regression which assumes that the output variable or dependent feature must be normally distributed.

Rented Bike Count w.r.t. Hour



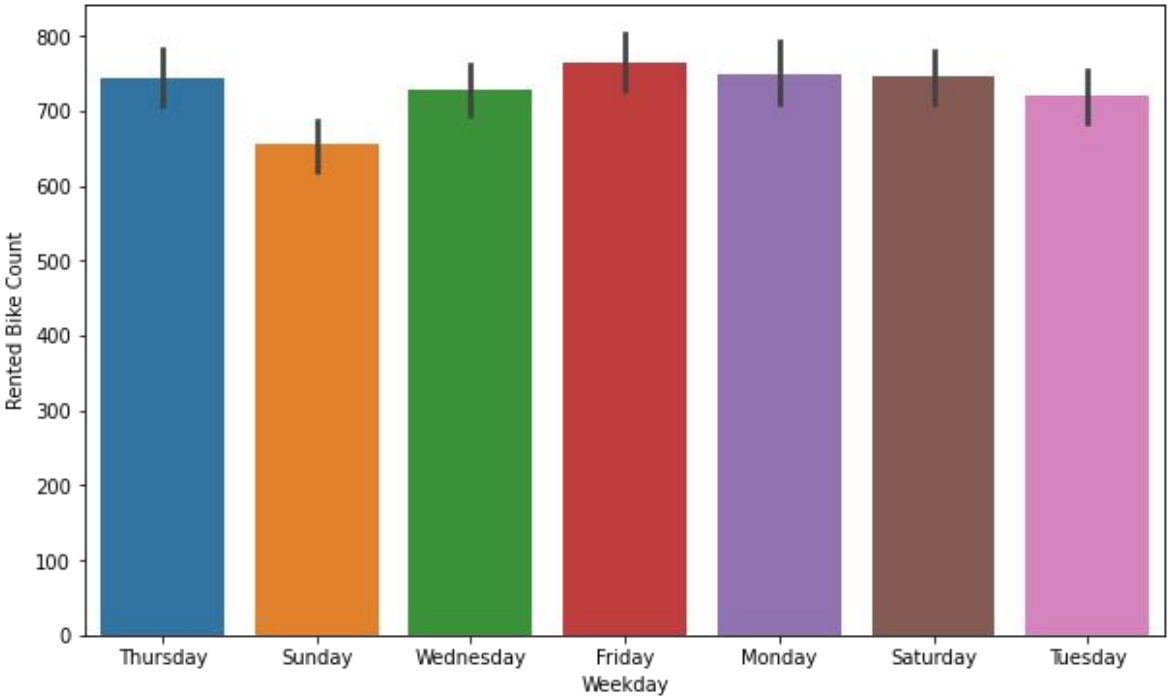
18:00 hrs was the peak time for bike rentals. Whereas, **4-5** in the **morning** were the timing with lowest number of bikes rented. Most numbers of bike are rented in between **17:00** hrs to **21:00** hrs.

Rented Bike Count w.r.t Season



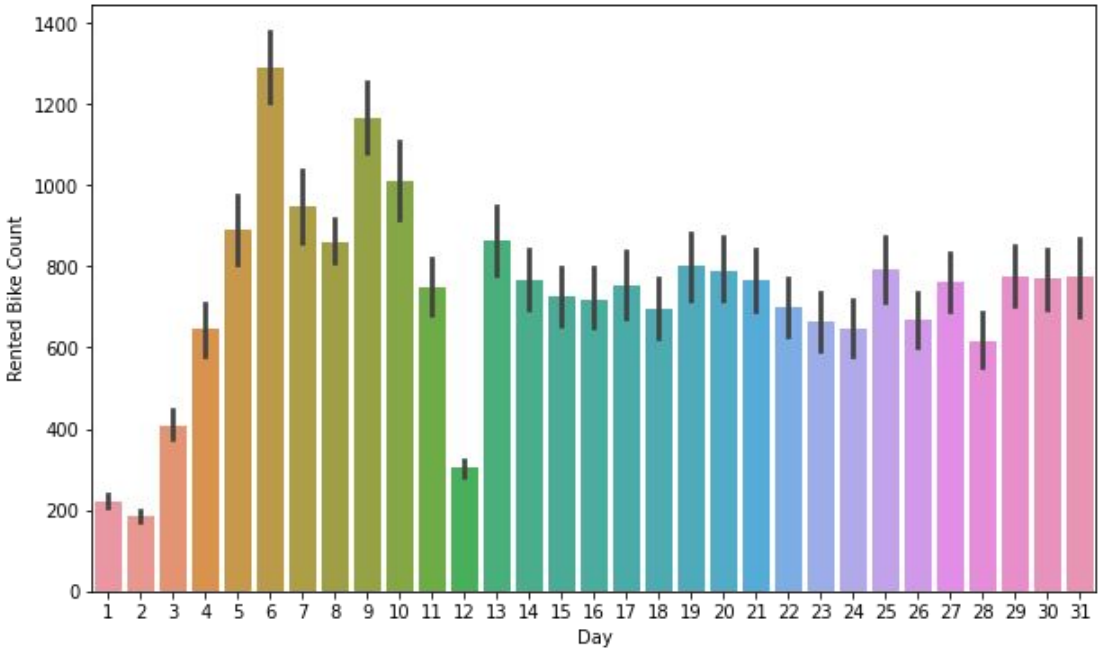
We can observe that, **most number** of bikes are rented during the **summer** season whereas, **winter** is the season with **lowest number** of bikes rented.

Rented Bikes Count w.r.t. Weekday



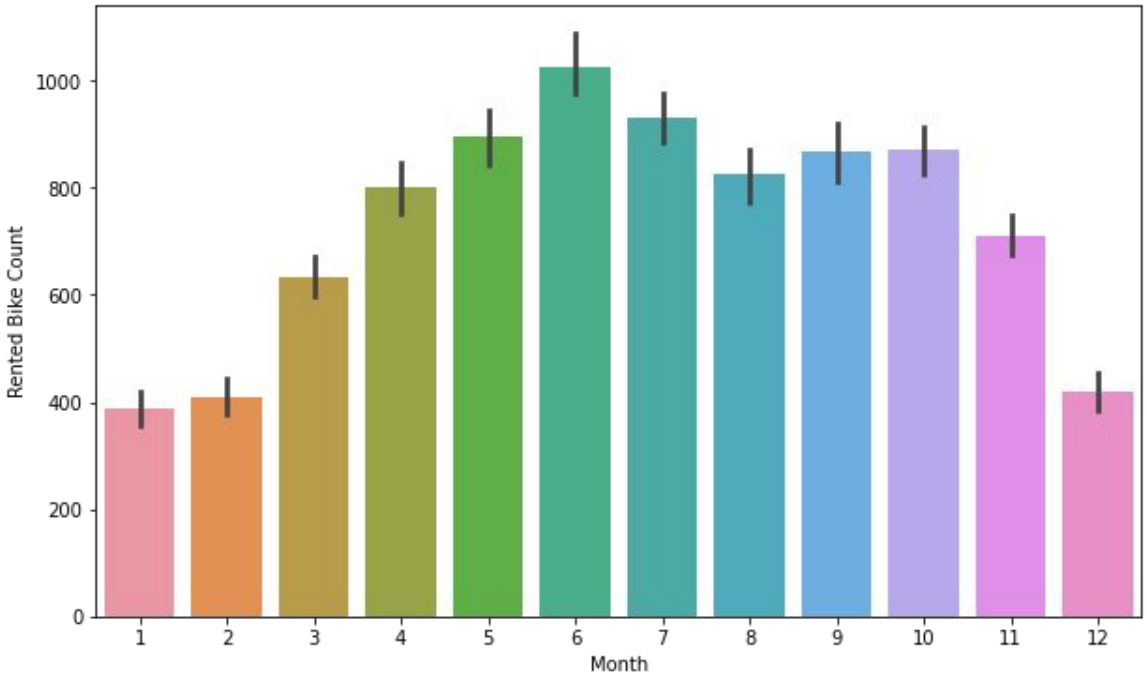
Sunday has the **lowest** number bikes rented. Whereas, every other weekday share approximately same amount of rented bikes count.

Rented Bikes Count w.r.t. Date



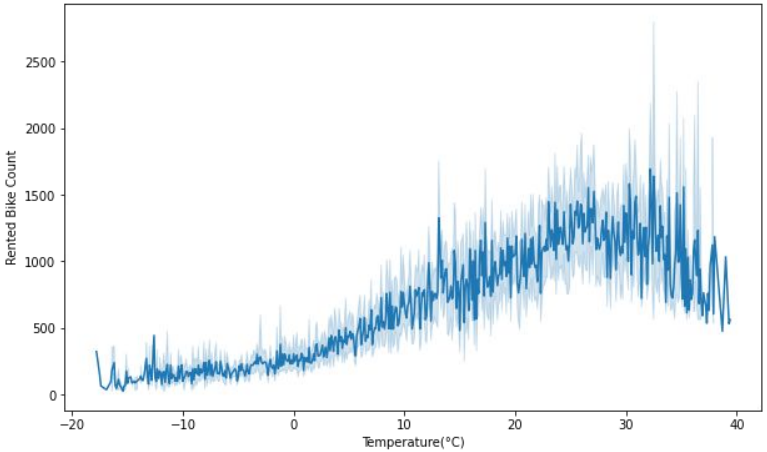
1st, 2nd & 12th have the lowest number for bike count whereas **6th** has the highest figures. Most number of bikes rented are between **5th to 10th**. From **13th onwards** till the **last date** of the month number of bike rented is approximately same.

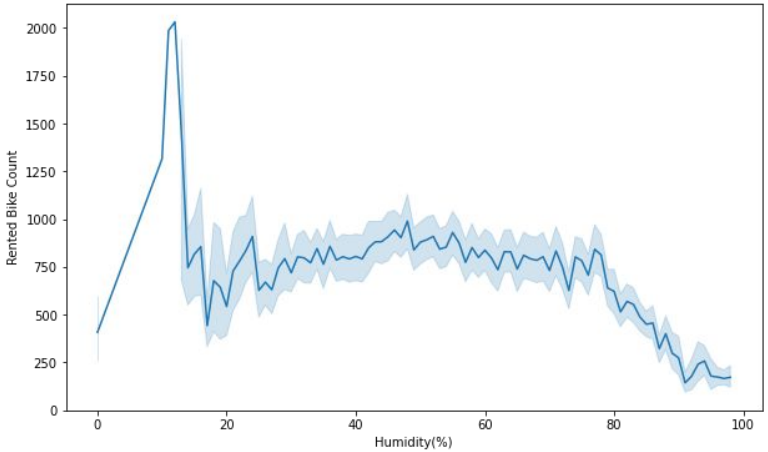
Rented Bikes Count w.r.t. Month

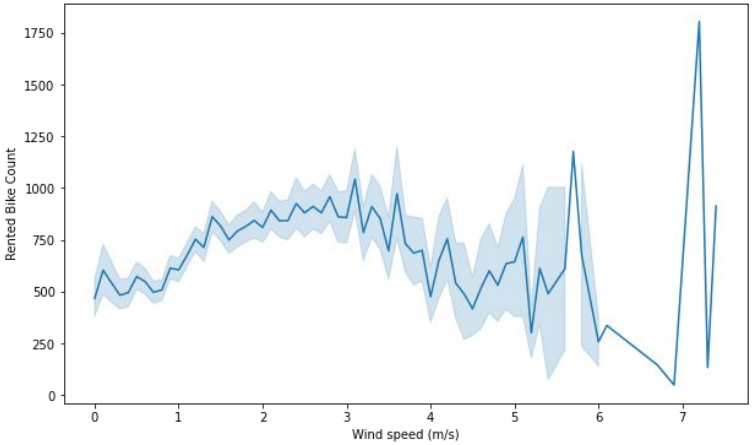


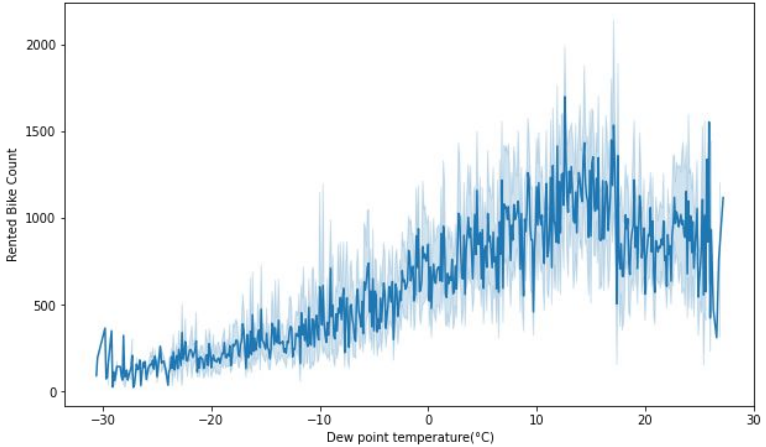
June, July & May have the **highest number** of count of rented bikes accordingly. Whereas **Jan, Feb and December** falls on the **lowest** side.

Rented Bikes Count w.r.t. Other Numerical Features

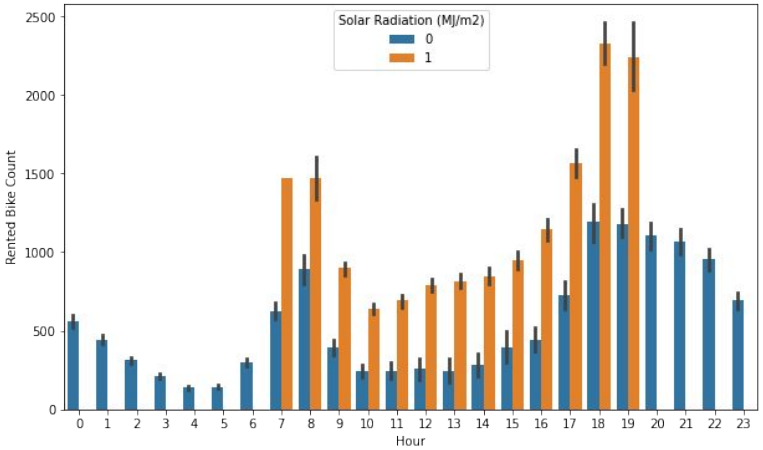


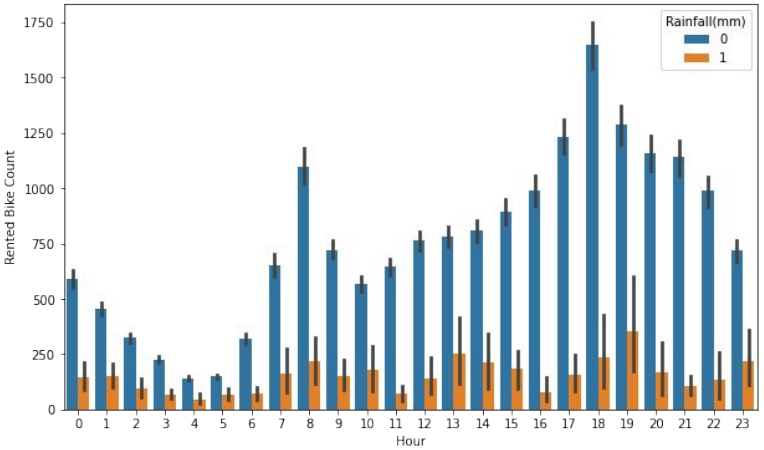


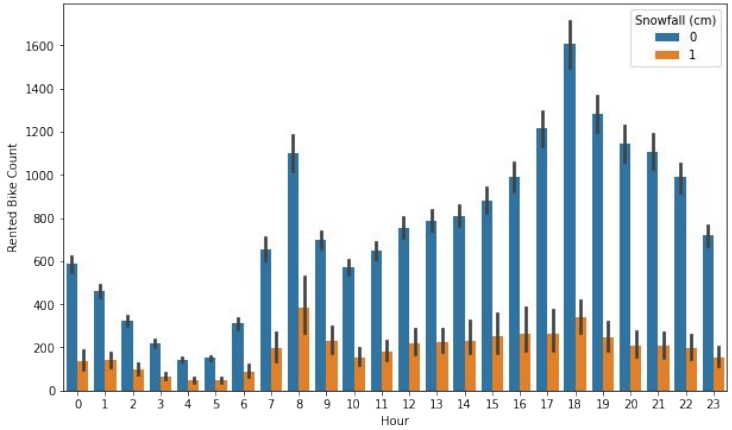




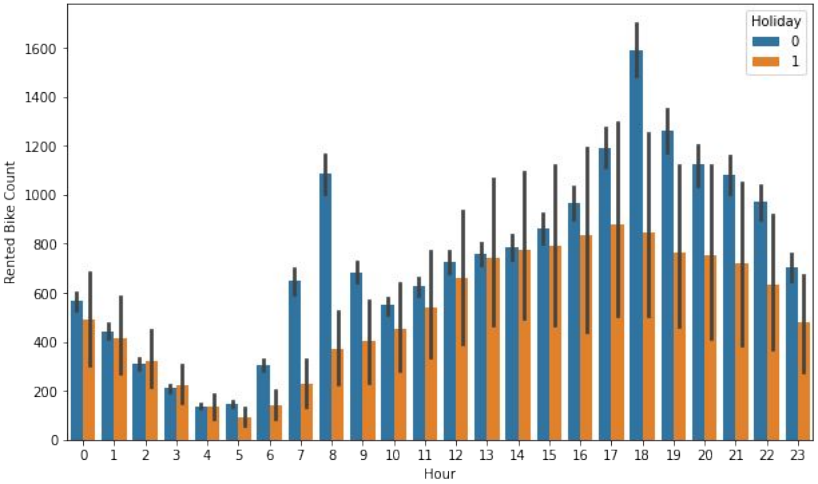
Rented Bikes Count w.r.t. Hour & Solar Radiation, Rainfall & Snowfall

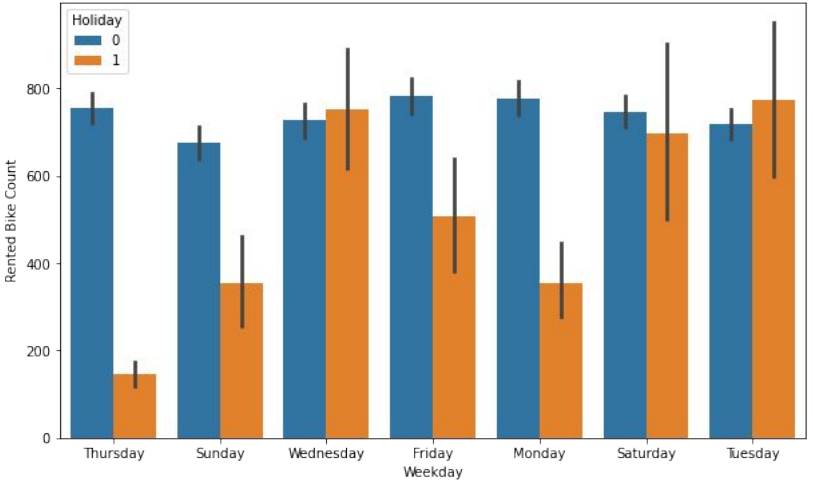




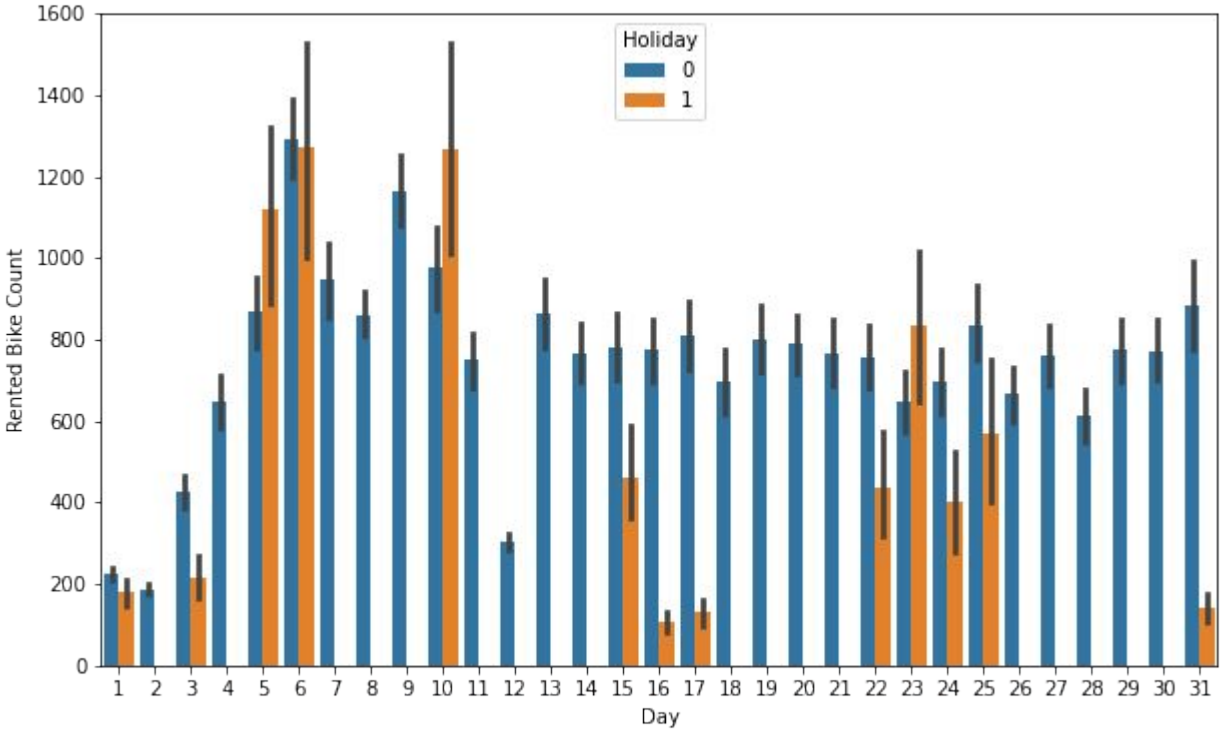


a)Rented Bikes Count w.r.t. Hour & Holiday
b)Rented Bikes Count w.r.t. Weekday & Holiday

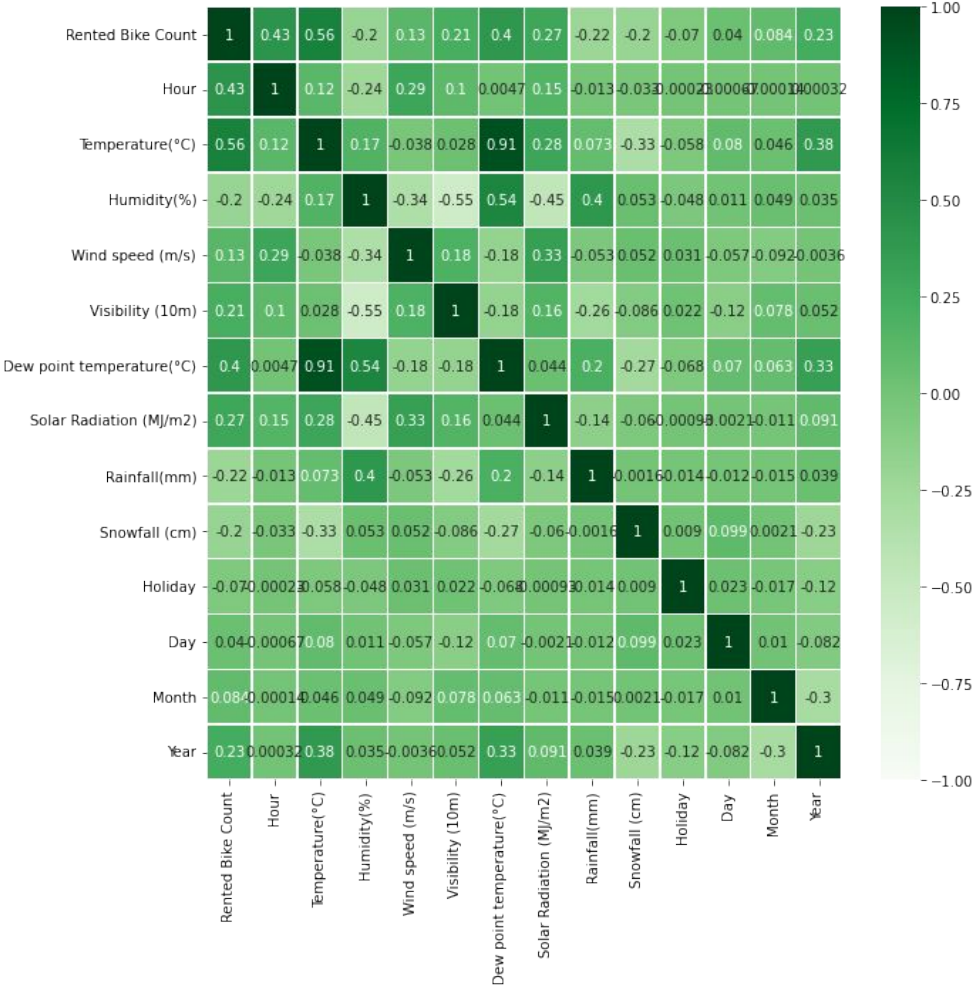




Rented Bikes Count w.r.t. Date & Holiday



Data Correlation Heatmap



Feature Engineering

Feature engineering is the process of transforming existing feature & producing new feature which are beneficial for model performance.

1. Converted **Date** column for **object** data type to **DateTime** data type
2. Extracted **day**, **month**, **year** & **weekday** from **Date** column and then **drop** the **Date** column.
3. Converted **snowfall**, **rainfall** & **soldar radiation** feature from **numeric** to **categorical** as more than 50% of the value were 0.
4. Performed one hot encoding for **seasons & weekday** categorical features.
5. Dropped **dew point temperature** column as it was very highly correlated with **temperature** column.
6. Dropped **functional day** column as every day was a functioning day.

Hyperparameter Tuning

Hyperparameter tuning is the process of finding an optimal parameter for the model for training which can have a great impact on the performance.
We can use the following methods -

1. Grid Search CV - Grid Search is one of the most basic hyper parameter technique used and so their implementation is quite simple. All possible permutations of the hyper parameters for a particular model are used to build models. The performance of each model is evaluated and the best performing one is selected. Since GridSearchCV uses each and every combination to build and evaluate the model performance, this method is highly computational expensive.

2. Randomized Search CV - In randomizedsearchcv, instead of providing a discrete set of values to explore on each hyperparameter, we provide a statistical distribution or list of hyper parameters. Values for the different hyper parameters are picked up at random from this distribution.

Thus, we randomized search cv is much more time efficient and it is suitable for large dataset. Whereas, grid search cv is more time consuming and is suitable for small dataset. I have used randomized search cv.

Model Building & Evaluation

1. Linear Regression

Linear regression is the most basic & simplistic machine learning algorithm. In linear regression we finds a best fit line for y | X. i.e., for output(y) w.r.t. the given features(X).

Linear Regression Model Performance & Evaluation -

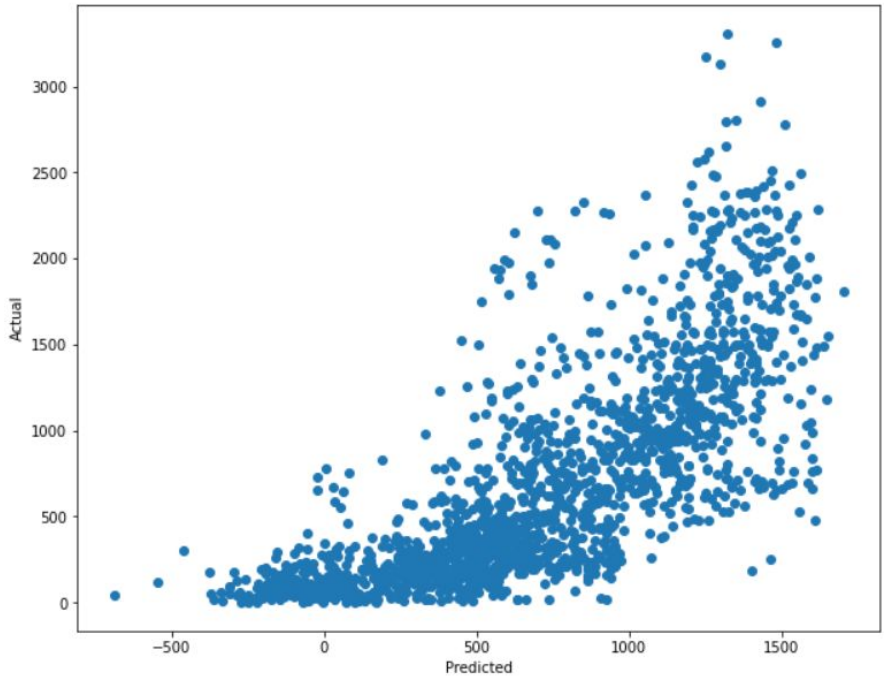
```
R^2 Score : 0.5511739358478436
Adj R^2 is 0.544716006867237
RMSE is: 419.99745801962837
```

r2 - score, **adjusted r2 - score** and **rmse** are some of the performance metrics for regression based machine learning problems.

r2 - score is the amount of variation of output dependent variable which is predictable from the input variable(s).

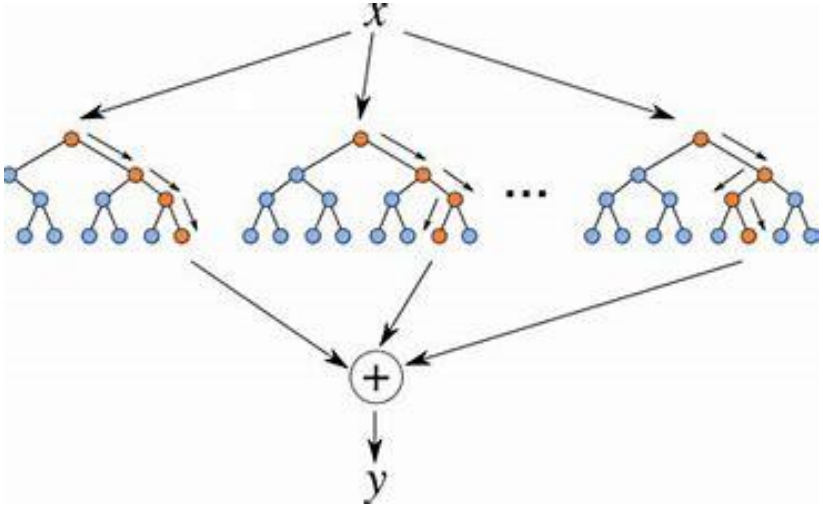
Adjusted r2 - score is used to overcome a problem of **r2-score.**

rmse tells us the average distance between the predicted values from the model and actual values.

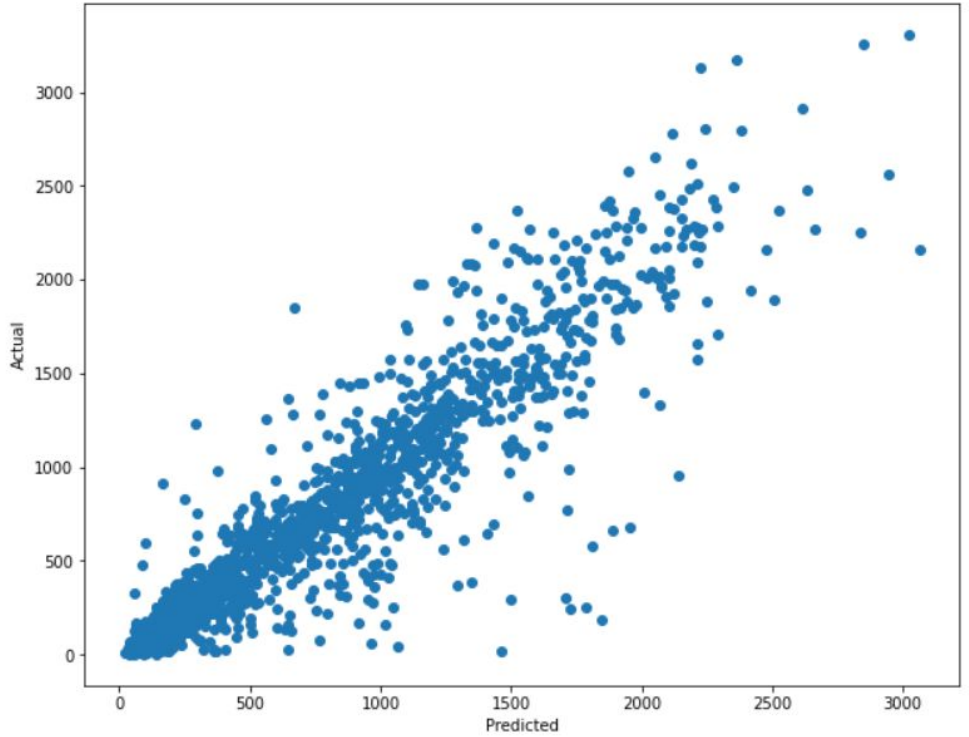


2. Random Forest Regressor

Random forest regressor is an ensemble learning technique. In ensemble learning technique we take same model multiple times or multiple algorithms to put forth a model that's more powerful than the original.

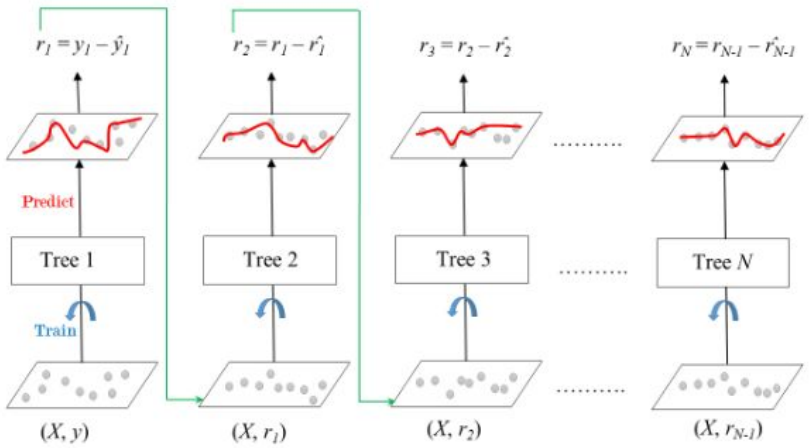



```
R^2 Score : 0.85625370406709
Adj R^2 is 0.8541854120392783
RMSE is: 237.68728972598564
```

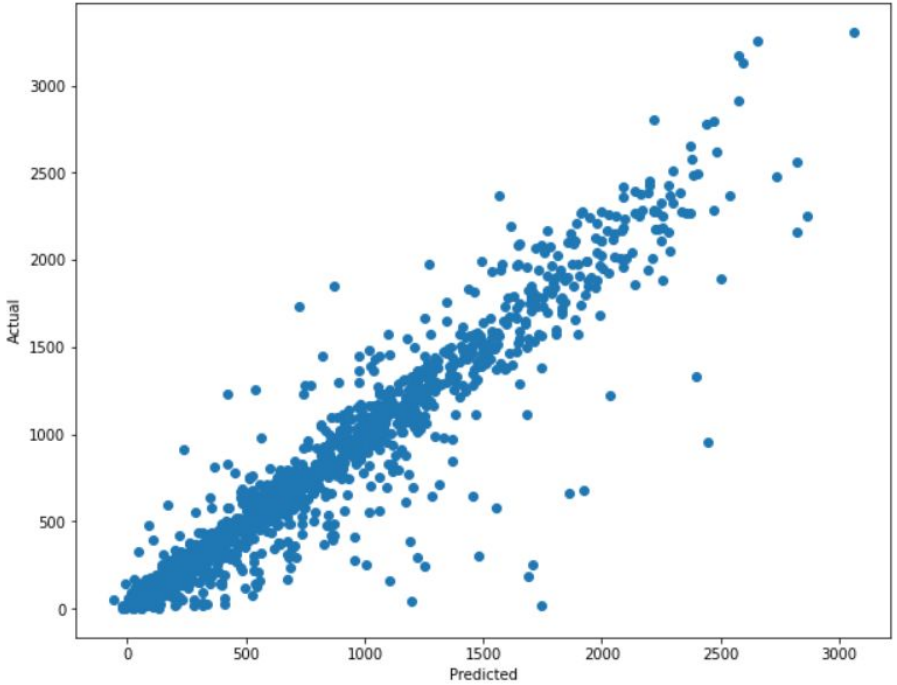


3. Gradient Boosting Regressor

Model is trained sequentially and not parallelly unlike random forest. Here, the error of previous model acts as an input for current model and so on. Week learners are combined to build a one strong model which gives good accuracy and performance.




```
R^2 Score : 0.9104120287568997
Adj R^2 is 0.9091229931994451
RMSE is: 187.64304773474177
```

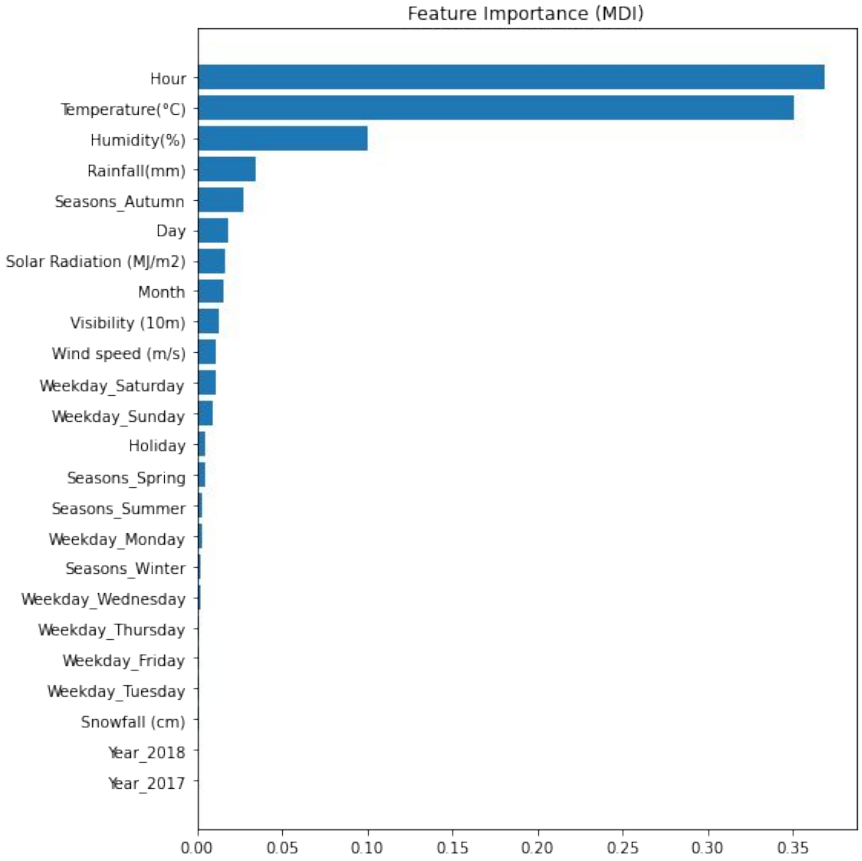


Feature Importance

We can observe that **Hour, Temperature & Humidity** are the most important features for the prediction of our target variable.

Rainfall, Seasons_Autumn, Day, Solar Radiation, Month & Visibility have some impact on the target variable.

Rest are the least important features for the prediction.



Conclusion

1. Most number of bikes are rented during evening time.
2. In summer more number of bikes are rented whereas, winter has the lowest count.
3. Least numbers of bike are rented on 12th of the month.
4. More bikes are rented if the humidity is low and wind-speed is high.
5. Rainfall and snowfall impact the number of bikes rented tremendously with very high downfall.
6. Linear regression is not suitable for our problem as it makes many assumptions and our dataset is prone to it. Thus, linear regression gives us the lowest r2-score and highest rmse.
7. Random forest regressor performs really good when compared to linear regression with high model performance and low rmse. But it's performance is low when compared to gradient boosting regressor. However, time taken for hyperparameter tuning and training the model is much low for random forest regressor then gradient boosting regressor. Thus, there's a tradeoff of accuracy and time in between random forest and gradient boosting regressor. It's up to us and business domain to which algorithm to use.
8. Hour, temperature and solar radiation were the most important features for predicting the count of bikes required.

THANK YOU