CAPSTONE PROJECT -2

Bike Sharing Demand Prediction

By

Owes Khan

Points to Discuss:

- Problem Statement
- Data summary
- Data Reading
- Exploratory Data Analysis
- Feature Engineering
- Modelling
- Model Performance Comparison
- Final Model
- Conclusion

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 the bike count required at each hour for the stable supply of rental bikes.

The goal is to build a Machine Learning model to predict the bike-sharing demand using the previously stored data.

Data Summary

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

Attribute Information:

- Date : year-month-day
- Rented Bike count - Count of bikes rented at each hour
- 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)

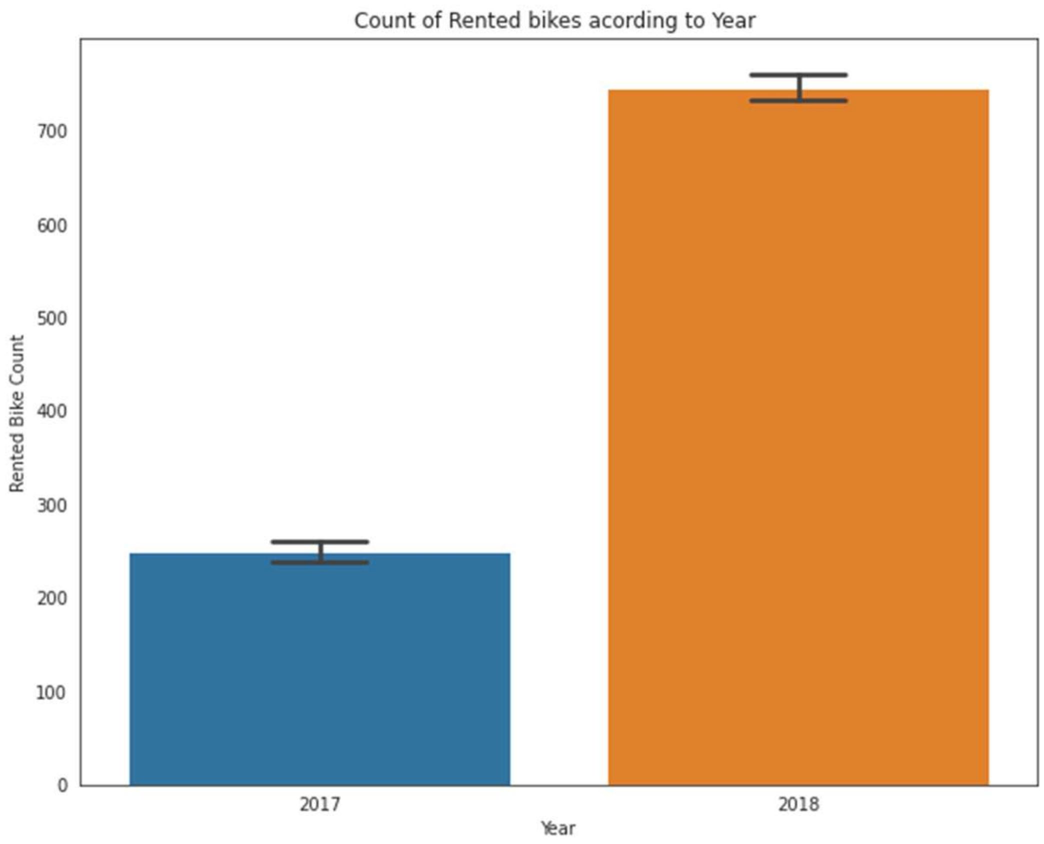
Data Reading

- The dataset contains total fourteen features.
- The dataset has no missing values.
- The dataset has both numerical and categorical features.
- The response variable is “Rented Bike Count” as the job is predict rented bike counts per hour.

```
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Date                       8760 non-null   object
 1   Rented Bike Count          8760 non-null   int64
 2   Hour                       8760 non-null   int64
 3   Temperature(°C)            8760 non-null   float64
 4   Humidity(%)                8760 non-null   int64
 5   Wind speed (m/s)           8760 non-null   float64
 6   Visibility (10m)           8760 non-null   int64
 7   Dew point temperature(°C)  8760 non-null   float64
 8   Solar Radiation (MJ/m2)    8760 non-null   float64
 9   Rainfall(mm)               8760 non-null   float64
 10  Snowfall (cm)              8760 non-null   float64
 11  Seasons                    8760 non-null   object
 12  Holiday                    8760 non-null   object
 13  Functioning Day            8760 non-null   object
dtypes: float64(6), int64(4), object(4)
```

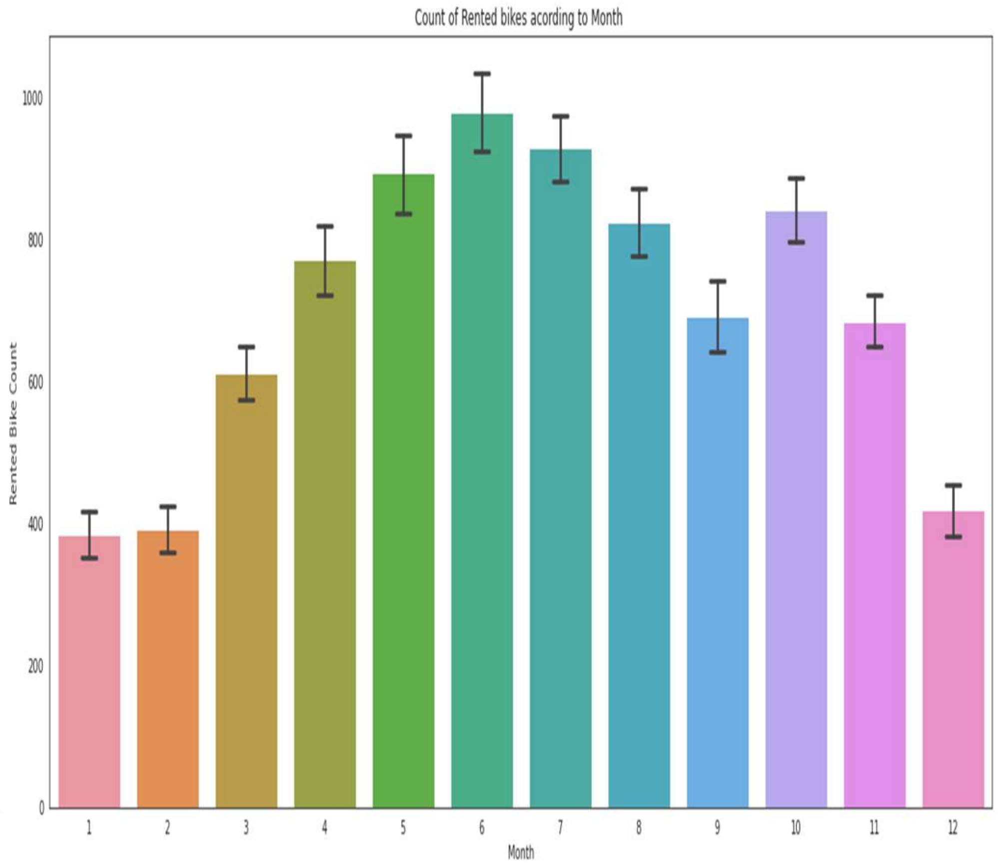
Exploratory Data Analysis

- In year 2017 less rented bike count as compare to 2018 that means more number of people using the bikes as compare to previous year and it become popular day by day .



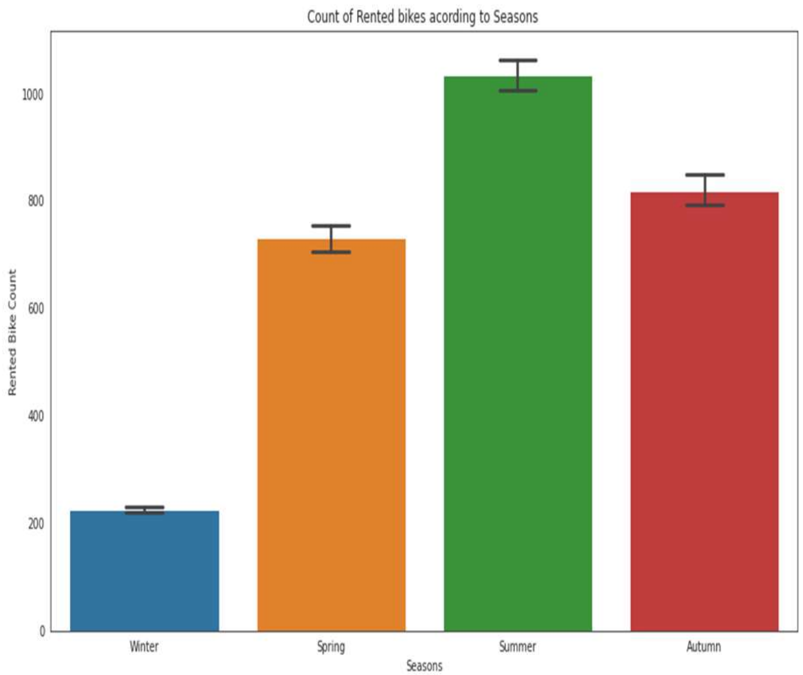
Exploratory Data Analysis

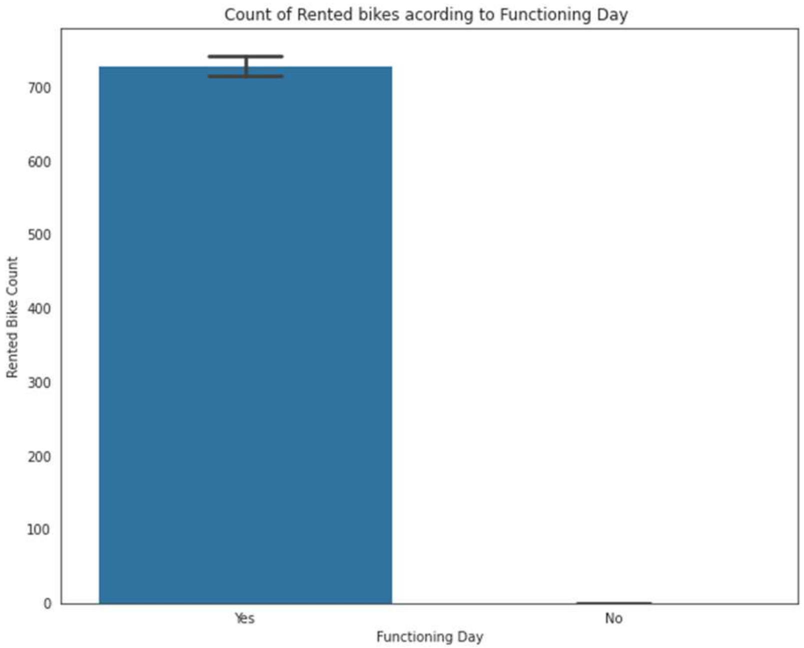
- This is a bar graph between rented bike count per month.
- Months are extracted from the date column and then plotted against the rented bike count.
- Here we can see that in the highest bike was rented in the month of June while lowest bike was rented in the month of January.
- From this we can assume that people tend to rent more bikes during summer season than in winter season.
- In next slide we will see the seasonal bike renting through Visualization so to prove our assumption.



Exploratory Data Analysis

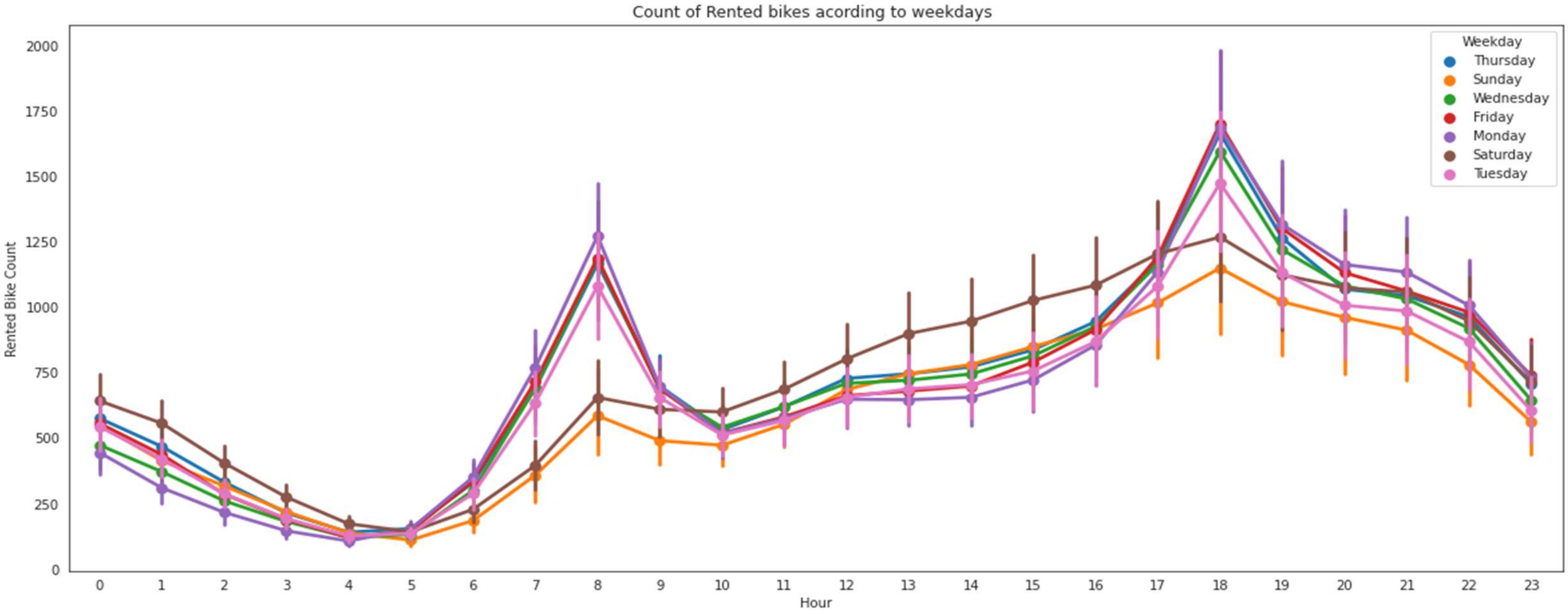
- From this we can conclude our assumption what we have assumed in the previous slide that bike rented during summer season was highest while in winter bike rented was lowest.





- No rented bike count on Non functioning day.

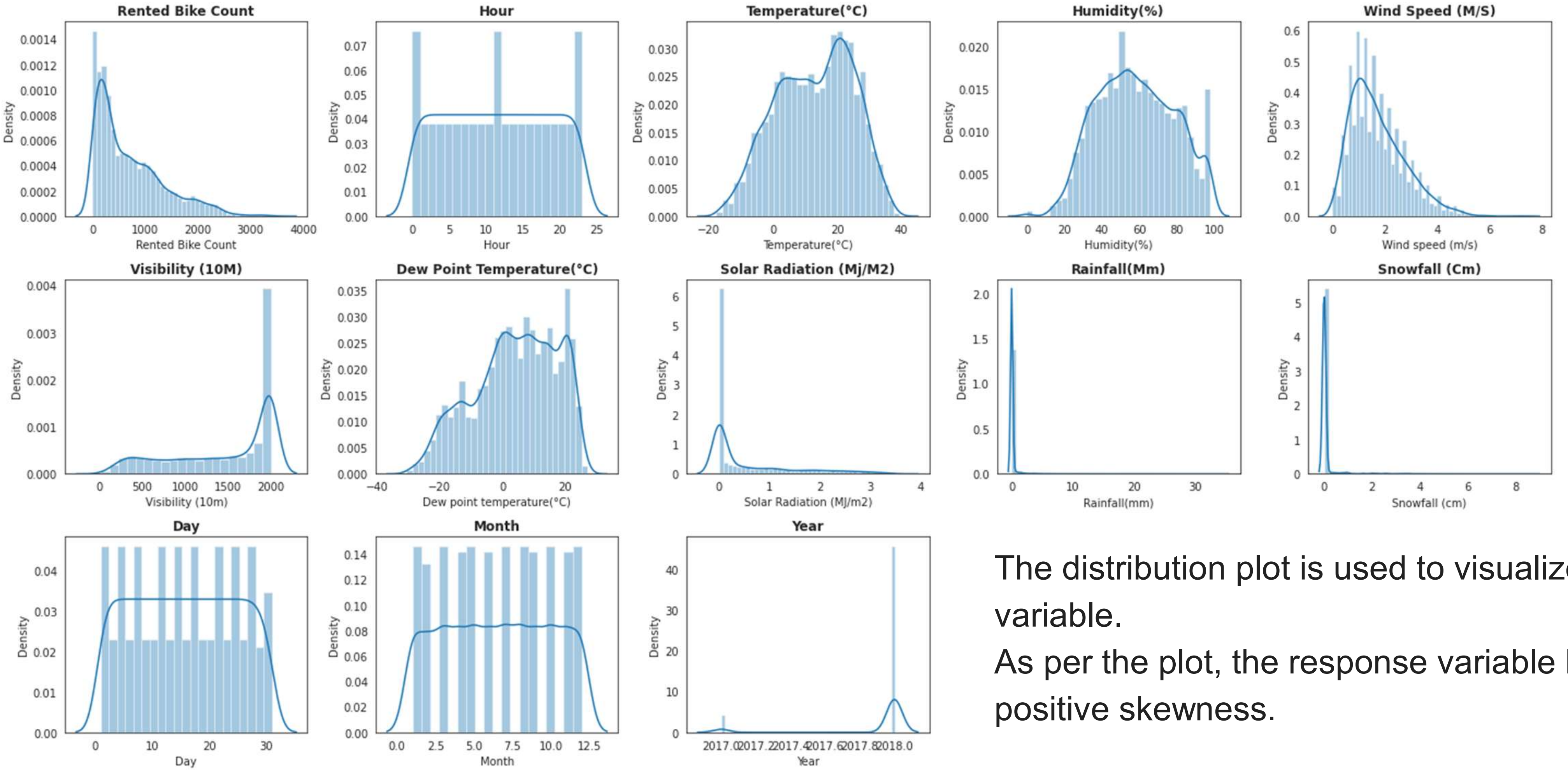
Exploratory Data Analysis



- People generally use more number of rented bikes during from 7 AM - 9 AM and 5 PM- 8 PM as it is office start and end time.
- Least numbers of bike are rented on Sunday as its holiday.

Feature Engineering

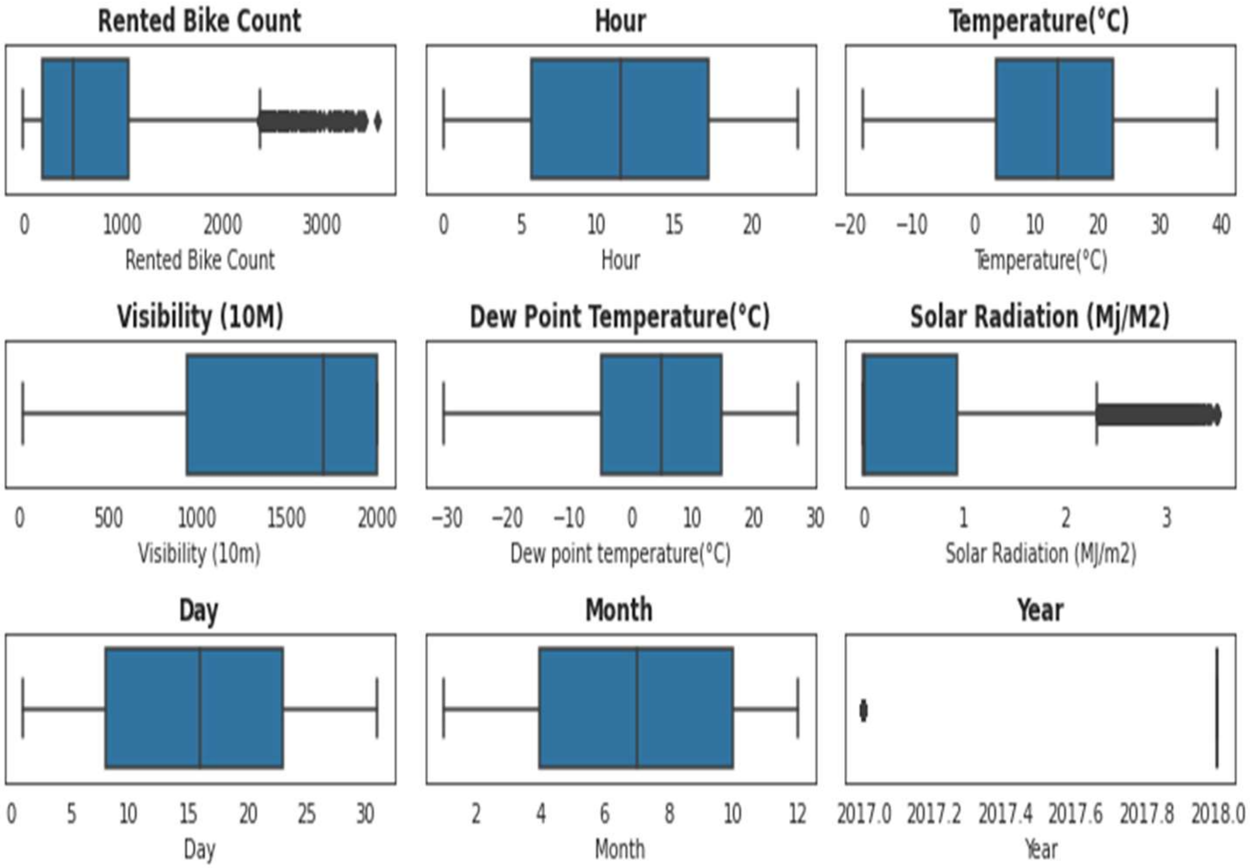
- **Distribution Graph**



The distribution plot is used to visualize the variable.
As per the plot, the response variable has a positive skewness.

Feature Engineering

- **Boxplot**

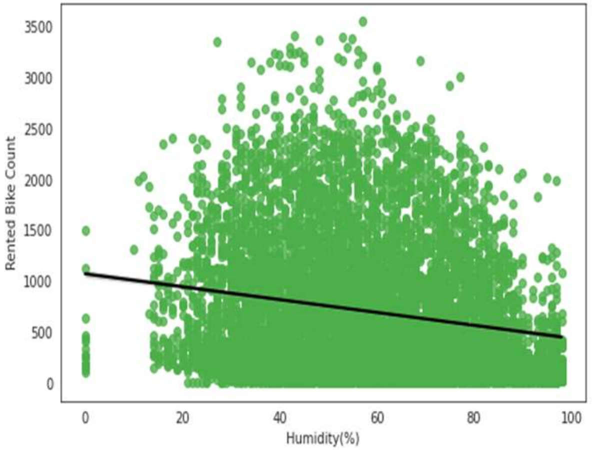


The box plot is used to visualize the 5 number summary (min,25%,50%,75%,max).

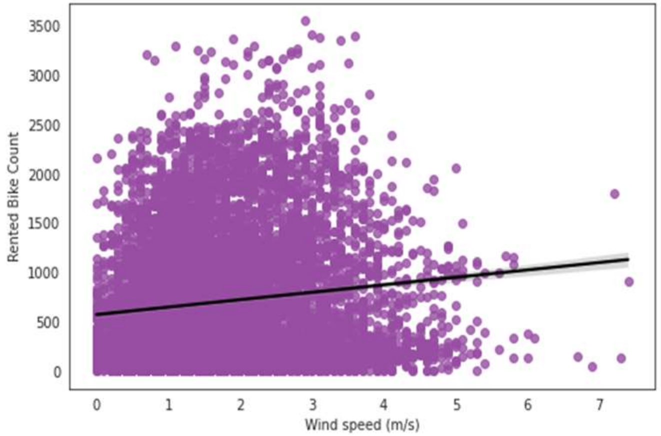
Also we can find out the outliers using boxplot plot.

Regression plots of Humidity, Wind speed & Visibility

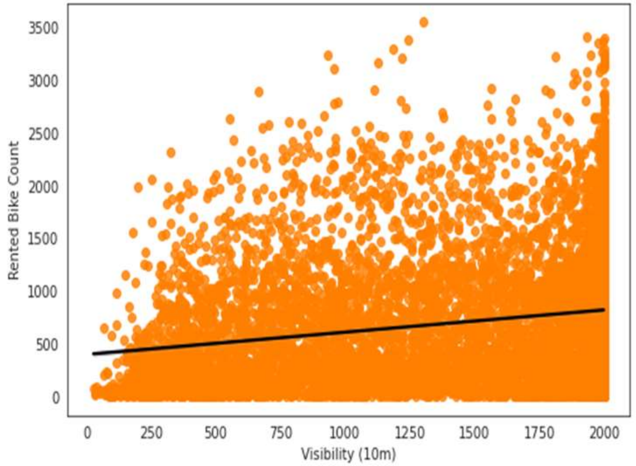
Rented Bike counts are having negative correlation or number of bike rented is decreasing with increase in humidity while we can see positive correlation of Rented bike with Wind Speed and Visibility.



Humidity

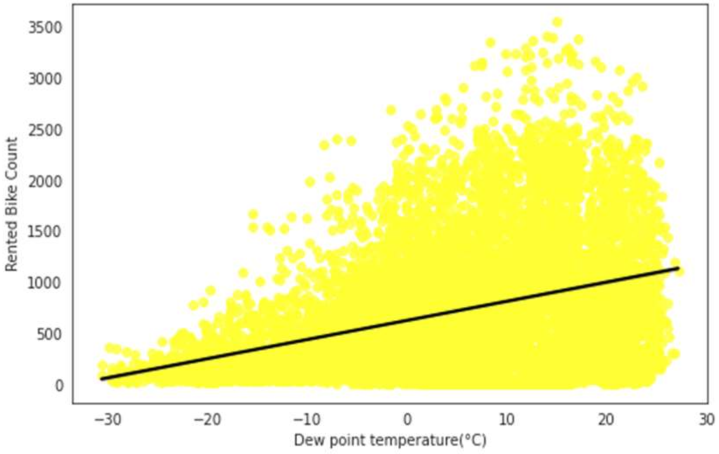


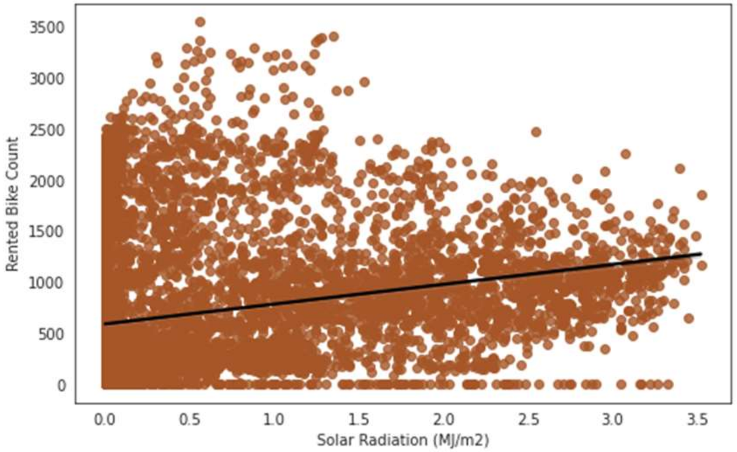
Wind Speed

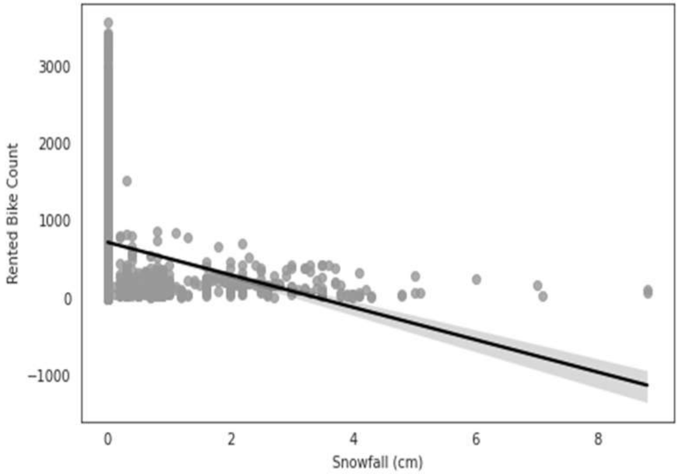


Visibility

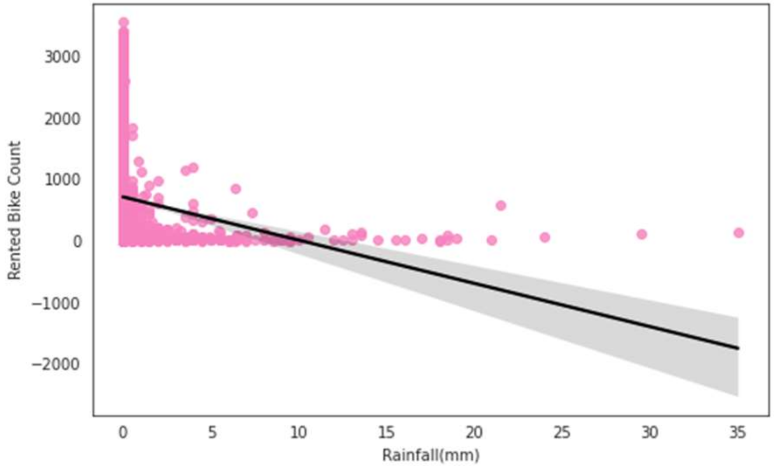
Regression plots of DPT, Solar Radiation, Snowfall & Rainfall





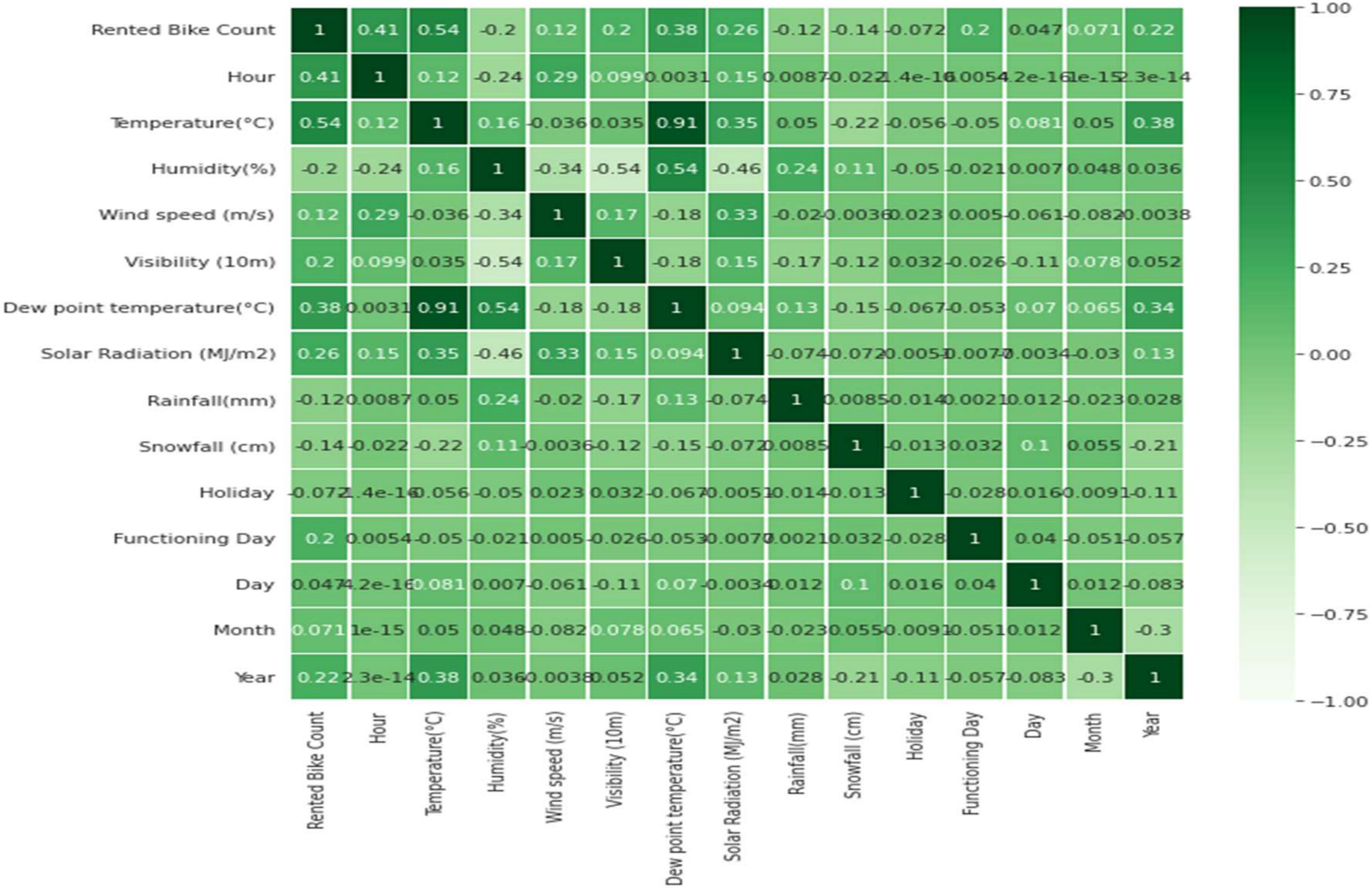


The rented bike counts are showing positive correlation with Dew Point Temperature (DPT) & Solar Radiation or in other words rented bike count is increasing with increase in Temperature (DPT) & Solar Radiation. while we see the negative correlation Snowfall and Rainfall or rented bike count decreasing with the increase in Snowfall and Rainfall.



Correlation Analysis (Before Treatment)

- The correlation matrix shows very high multicollinearity in temperature and dew point temperature.
- So one of the features must have to be dropped based on VIF (Variance Inflation factor)



Variance Inflation Factor

	variables	VIF
0	Humidity(%)	4.878319
1	Visibility (10m)	4.730979
2	Wind speed (m/s)	4.610685
3	Hour	3.922387
4	Temperature(°C)	3.238208
5	Solar Radiation (MJ/m2)	2.247281
6	Snowfall (cm)	1.121043
7	Rainfall(mm)	1.079201
8	Holiday	1.055235

VIF for all features except Dew point temperature

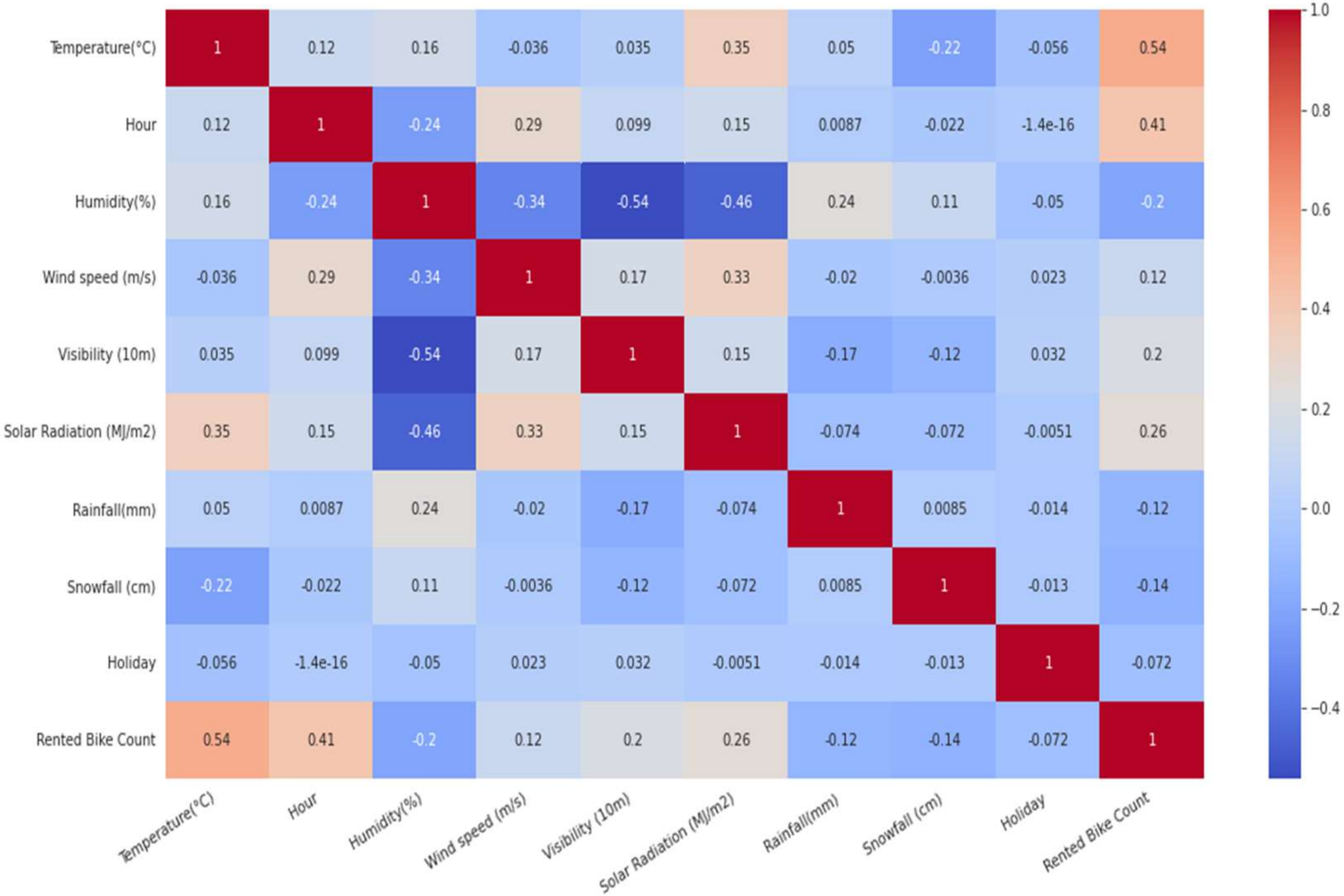
VIFs for features without Dew point Temperature feature:

- VIFs are high for Temperature and Dew Point Temperature when all the features are considered
- When the Dew point temperature feature is not considered for VIFs, all VIFs for other features decreases significantly.
- Therefore, we decided to drop it VIF value is less than 5 so features are less correlated.

Correlation Analysis (After Treatment)

- Correlation plot after dropping the Dew point temperature feature show that there are no more highly correlated parameters present in the dataset.

- We can conclude that, there is no multicollinearity present in the dataset



Data Preparation before Modelling

- The features “Temperature(°C)” and “Dew point temperature(°C)” are collinear so one of the features is dropped to get better the prediction.
- The square root transformation is used for the response variable “Rented Bike Count” to remove the skewness, for linear regression feature should be normally distributed.
- The dataset is split into 80% train and 20% test.
- The Standard Scaler is used to standardize the numerical features.
- The One Hot Encoder is used to encode the categorical features as these features are nominal in nature.
- The final train set has 7008 rows and 25 columns ,and final test set has 1752 rows and 25 columns.

Linear Regression

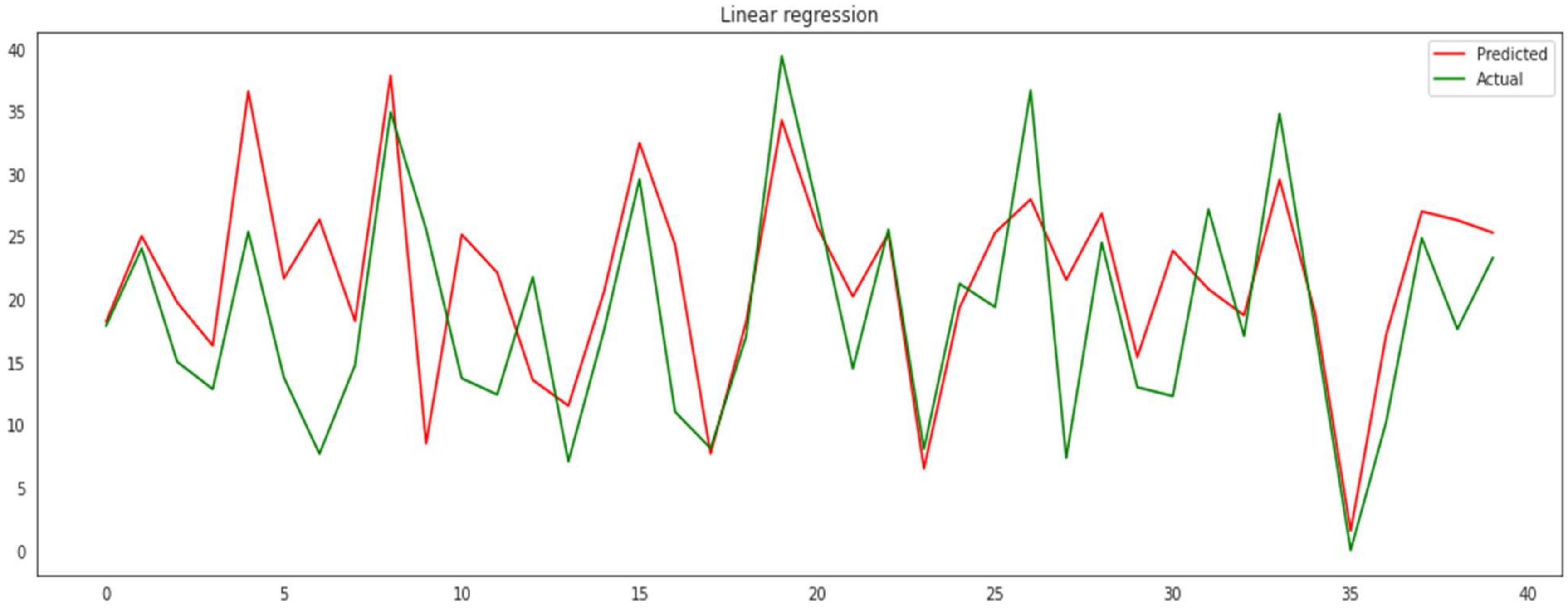
- Model accuracy is less for training as well as test data. Therefore we can conclude that under fitting.
- Since there is under fitting, we did not go ahead with Regularized linear Regression
- We plotted line graph of actual vs predicted Rented bike count.

Test Data

```
MSE : 54.014258575701604
RMSE : 7.349439337507427
MAE : 5.673049509794896
R2 : 0.6570223996444029
Adjusted R2 : 0.6520545896740148
```

Train Data

```
MSE : 53.19174406412738
RMSE : 7.293267036392359
MAE : 5.60103870631177
R2 : 0.6553446003715145
Adjusted R2 : 0.6503524885576605
```



Polynomial Regression

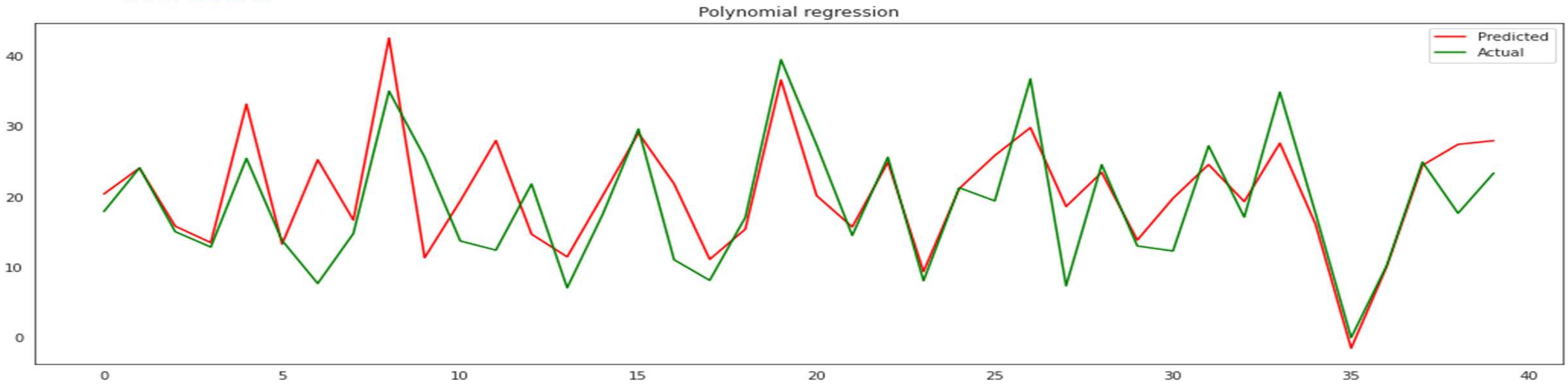
- Model accuracy is improved for training as well as test data as compared to the Linear Regression model.
- MSE and MAE have reduced significantly for polynomial Regression
- R^2 for both training and test data is higher indicating the model is fit well on both the datasets
- We plotted a line graph of actual vs predicted Rented bike count

Test Data

MSE : 37.21894444151347
RMSE : 6.1007331068908
MAE : 4.4913074860447
R2 : 0.7636686203064712
Adjusted R2 : 0.7602455122576078

Train Data

MSE : 33.04928084056067
RMSE : 5.748850392953418
MAE : 4.262401952789473
R2 : 0.7858574992050441
Adjusted R2 : 0.7827557827972377



Decision Tree Regressor

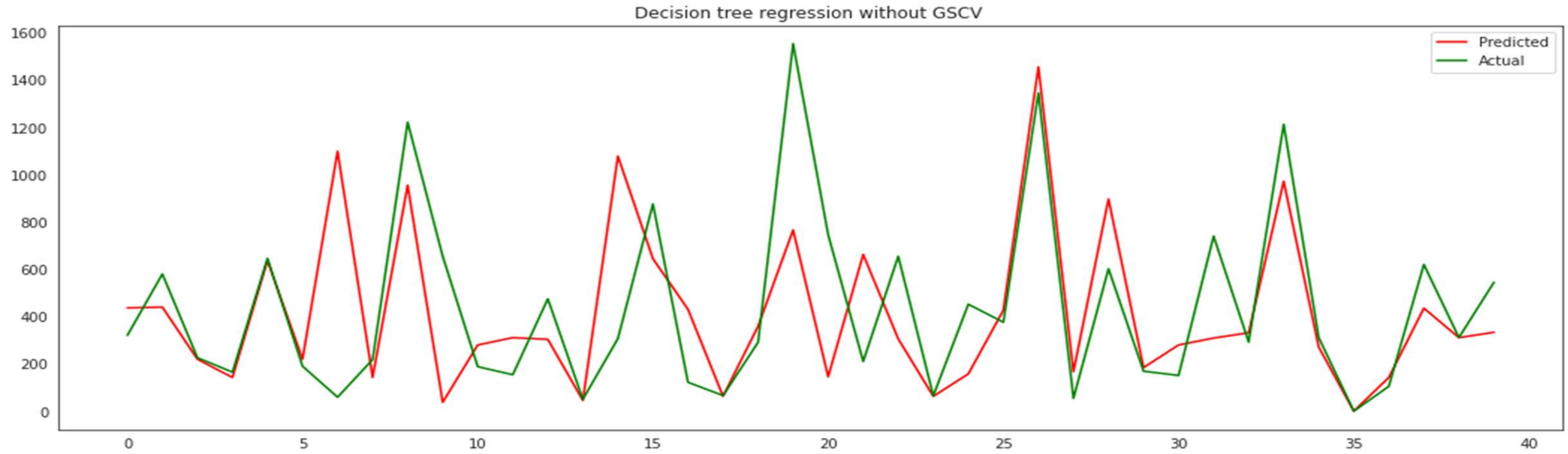
- Parameters: max depth = 15, max-leaf nodes = 1000, max_features=10
- R^2 for training is good but for test data is bad that indicating the model is underfitting.
- We plotted a line graph of actual vs predicted Rented bike count .

Test Data

```
MSE : 85618.50511436266
RMSE : 292.6063996469706
MAE : 173.5860744371091
R2 : 0.7954268001864369
Adjusted R2 : 0.7924636889492763
```

Train Data

```
MSE : 20290.85372778228
RMSE : 142.44596774841426
MAE : 85.4867494643508
R2 : 0.9511349915131644
Adjusted R2 : 0.9504272132905857
```



Decision Tree Regressor with GridSearchCV

- bestparams: max depth = 11, max-leaf nodes = 4
- Using Gridsearchcv
- R^2 for training is well as test data is good that indicating the model is well fitted.
- We plotted a line graph of actual vs predicted Rented bike count .

Test Data

```
MSE : 67292.54593662864
RMSE : 259.4080683722629
MAE : 159.38623867278002
R2 : 0.8392140644423842
Adjusted R2 : 0.836885183568143
```

Train Data

```
MSE : 37894.65711268774
RMSE : 194.66550057133324
MAE : 118.01697775648286
R2 : 0.9087410137464139
Adjusted R2 : 0.9074191860196817
```

Random Forest Regressor

- bestParameters: n_estimators = 100, max depth = 90, min_samples_split: 10, min_samples_leaf: 3 using RandomizedSearchCV

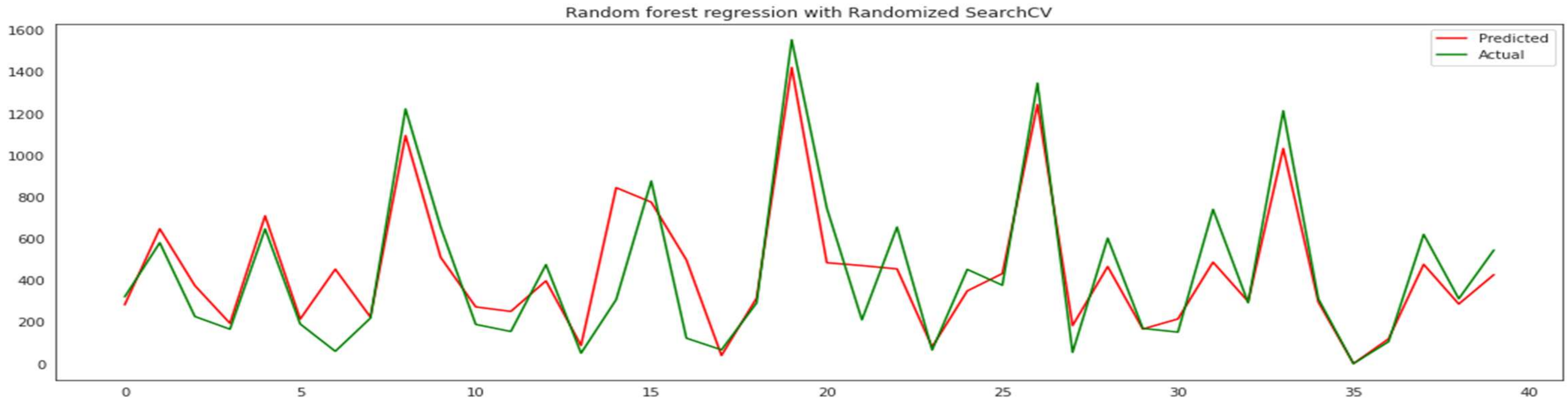
- R^2 for both training and test data is moderate indicating the model is fit well on both the datasets
- We plotted a line graph of actual vs predicted Rented bike count

Test Data

```
MSE : 42786.66487488584
RMSE : 206.84937726492154
MAE : 122.94365296803653
R2 : 0.8977673701366966
Adjusted R2 : 0.8962865962394877
```

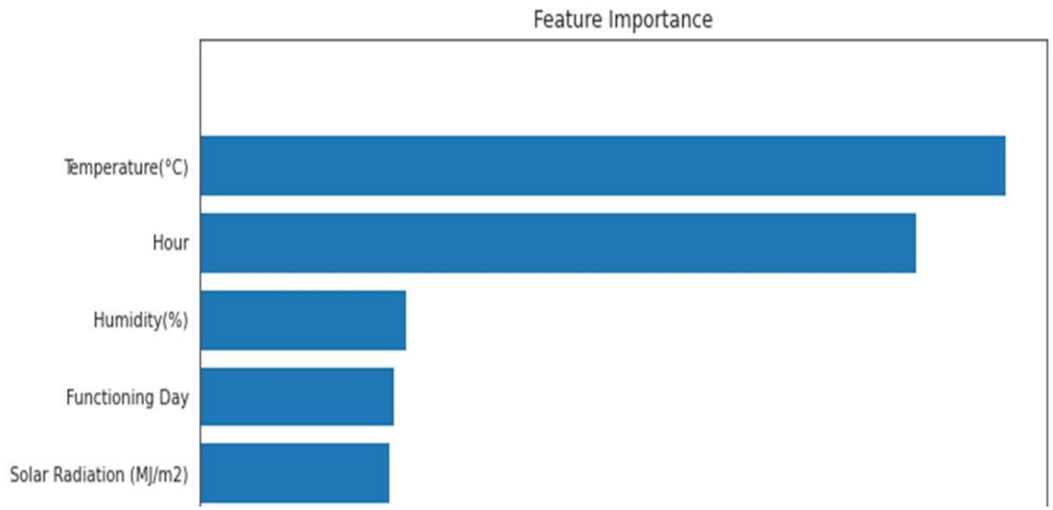
Train Data

```
MSE : 5855.552245390982
RMSE : 76.52158025936855
MAE : 44.96754994292237
R2 : 0.9858984932815139
Adjusted R2 : 0.9856942420254524
```



Gradient Boost with Hyper Parameter Tuning

- Best parameters according to RandomizedSearchCV
- Best_parameters = subsample': 0.8, 'n_estimators': 500, 'max_depth': 6, 'learning_rate': 0.03
- Here we can see Temperature is showing most importance feature then Hour in model prediction.
   -

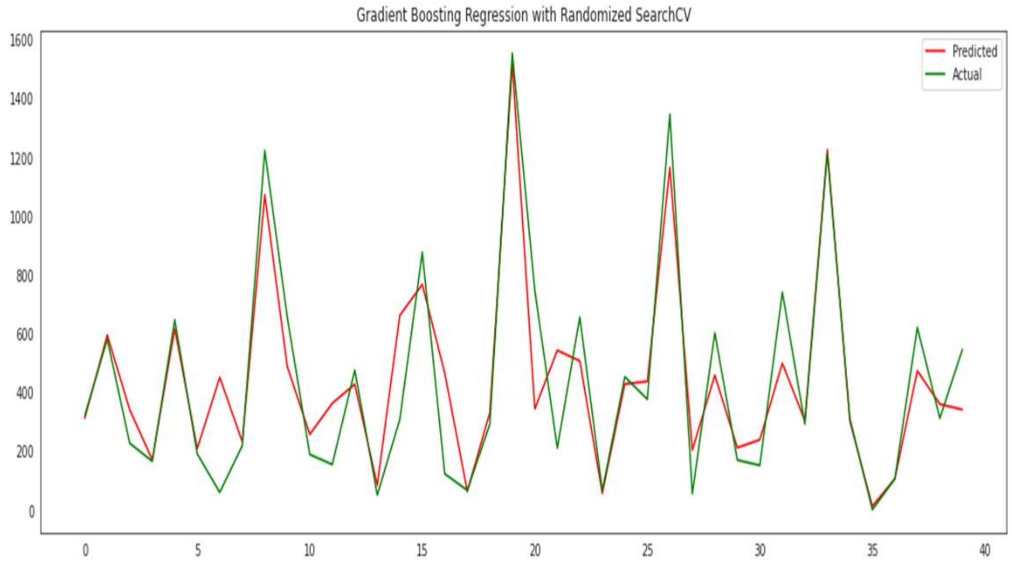


Test Data

MSE : 33298.72775905618
RMSE : 182.4793899569378
MAE : 110.98646123477332
R2 : 0.9204374419024035
Adjusted R2 : 0.9192850294154742

Train Data

MSE : 10440.193318642428
RMSE : 102.17726419630948
MAE : 67.93274217577584
R2 : 0.9748576308338789
Adjusted R2 : 0.9744934597857021



Metrics Selection

The problem is to predict the rented bike count , in which it is very important to know what factors will derive the prediction well. So to use following metric for prediction will be a good choice.

- R2 (R Squared) Score
- Adjusted R2 Score

But we will also look into other metrics like " Mean Squared Error (MSE) " and " Root Mean Squared Error (RMSE) " to keep a check how much error is made in prediction.

- **R2 Score** : The proportion of the variance in the dependent variable that is predictable from the independent variable(s).

- **Adjusted R2 Score** : The percentage of variance in the target field that is explained by the input or inputs. It is a corrected goodness-of-fit (model accuracy) measure for linear models

Modelling

- Linear Regression
- Polynomial Regression
- Decision Tree Regressor without Gridsearchcv
- Decision Tree Regressor with Gridsearchcv
- Random Forest Regressor
- Gradient Boosting Regressor

Model Performance Comparison

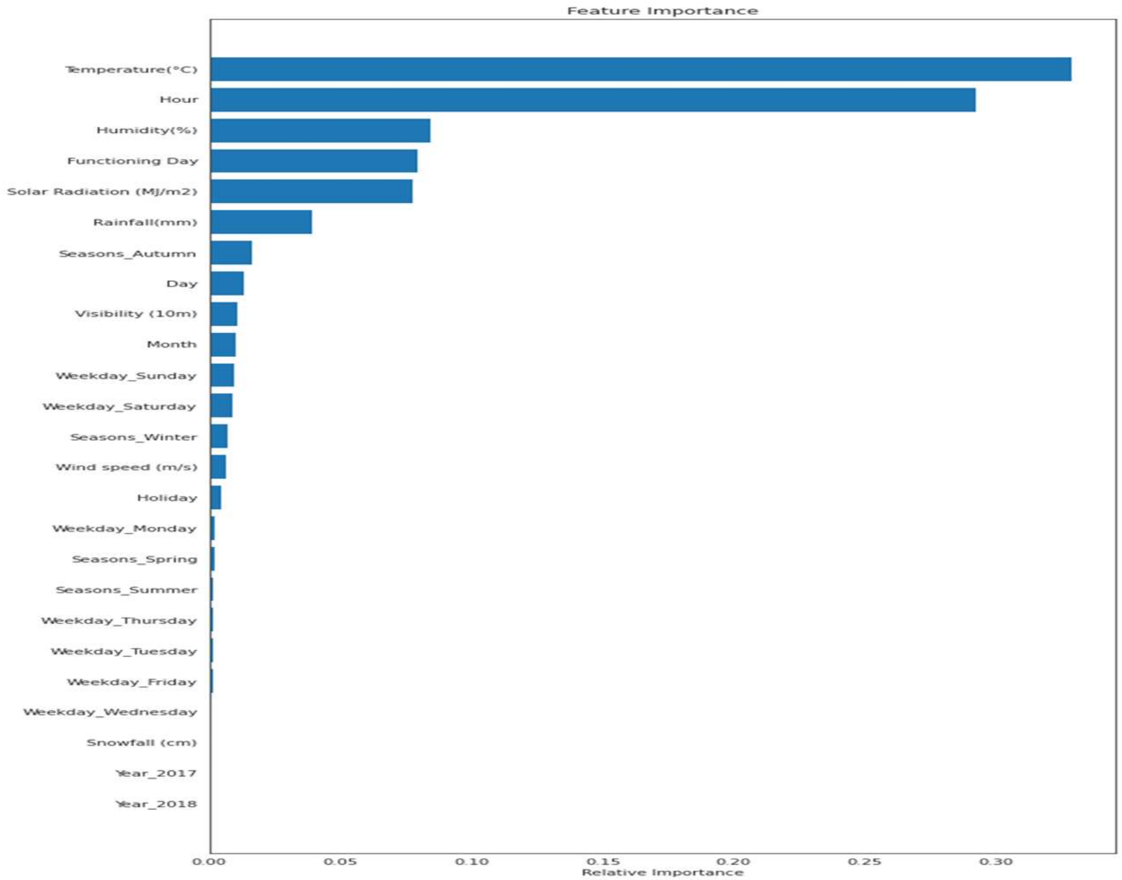
		Model	MAE	MSE	RMSE	R2	Adj_R2
Training set	0	Linear regression	5.601000	53.192000	7.293000	0.655000	0.650000
	1	Polynomial regression	4.262000	33.049000	5.749000	0.786000	0.780000
	2	Decision tree regression without GSCV	85.487000	20290.854000	142.446000	0.951000	0.950000
	3	Decision tree regression with GSCV	118.017000	37894.657000	194.666000	0.909000	0.910000
	4	Random forest regression	44.968000	5855.552000	76.522000	0.986000	0.990000
	5	Gradient Boosting Regression	67.933000	10440.193000	102.177000	0.975000	0.974000
Test set	0	Linear regression	5.673000	54.014000	7.349000	0.657000	0.650000
	1	Polynomial regression	4.491000	37.219000	6.101000	0.764000	0.760000
	2	Decision tree regression without GSCV	173.586000	85618.505000	292.606000	0.795000	0.790000
	3	Decision tree regression with GSCV	159.386000	67292.546000	259.408000	0.839000	0.840000
	4	Random forest regression	122.944000	42786.665000	206.849000	0.898000	0.900000
	5	Gradient Boosting Regression	110.986000	33298.728000	182.479000	0.920000	0.919000

Final Selected Model

The job is to predict the rented bike count in each hour, so for this Random Forest and GBoost performs best with R2 score as . But we also have to look model explainability and feature importance into consideration so that we can derive the important factors for predictions and also the reasons for the same to explain and to improve the business model. The Decision Tree performs good for train data with R2 score as 0.95 but bad with test data R2 score as 0.79. While the training time complexity of ensemble models is greater than of Decision Tree , since decision tree is prone overfitting, to reduce time complexity of ensemble model we use randomized searchcv for hyper parameter tunning which help us to get better R2 score for our ensemble model.

● Feature Importance

- Top features which are important for prediction are used in the plot.
- The “Temperature" feature is the most important factor for the predictions. The “Hour"   and Humidity are the other most important factors. The “Function Day” , "Solar Radiation", and "RainFall“ are next important factors. So these factors are the most important for the predictions, therefore we need to focus on them to improve the business model.



Conclusions

1.People generally use more number of rented bikes during from 7 AM - 9 AM and 5 PM-8 PM r as it is office start and end time.
2.In summer more number of bikes are rented whereas, winter has the lowest count.
3.Least numbers of bike are rented on Sunday as its holiday.
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.we can say that temperature has a highest weightage then Hour and humidity.
7.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.
8.Random forest regressor performs really good when compared to linear regression with high model performance. 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 than 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.
9.Out of all above models Gradient Boosting Regressor gives the highest R2 score of 97.5% for Train Set and 92.0% for Test set and no overfitting is seen.

Thank You