Capstone Project

Bike Sharing Demand Prediction

By

Ankit Patil
Chukkapalli Naga Sai
Saugata Deb
Shreyash Movale

Table of Contents

Introduction

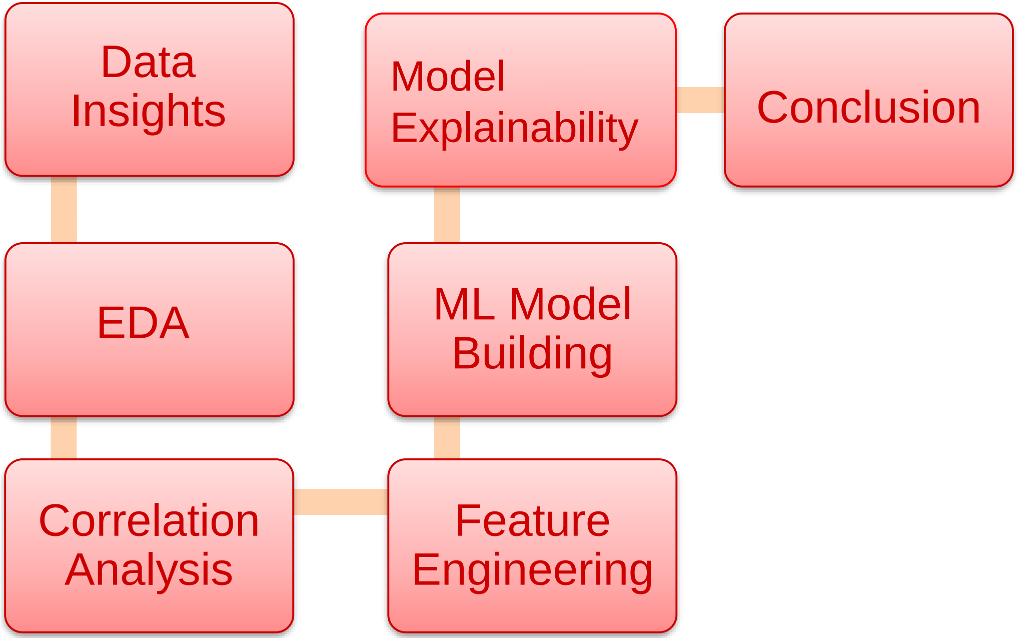
- A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free.
- Bike-sharing companies have gained a vast range of attention in recent years as part of initiatives to boost the use of cycles, improve the first mile/last mile link to other modes of transportation, and minimize the negative effect of transport activities on the environment.
- The goal is to build a Machine Learning model to predict the bike-sharing demand using the previously stored data.

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**.
- Create an ML Model for Prediction of Bike Count required at each hour

Process Flow

The process from getting the data to drawing the conclusion is as follows:



Data Insights:

Date: DD/MM/YYYY **str**

Rented Bike Count : **int**

Hour: **int**

Temperature(°C): **Float**

Humidity: int

Wind speed (m/s) : **float**

Visibility (10m): **int**

Dew point temperature(°C): **Float**

Solar Radiation (MJ/m2): UV radiation is **Float**

Rainfall(mm): **Float** t

Snowfall (cm): **Float**

Seasons: **str**

Holiday: **str**

Functioning Day: **str**

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

- **The data set has 14 variables of which Rented Bike Count is a Dependent variable and the rest are independent variables.**
- **No null value present in the data**

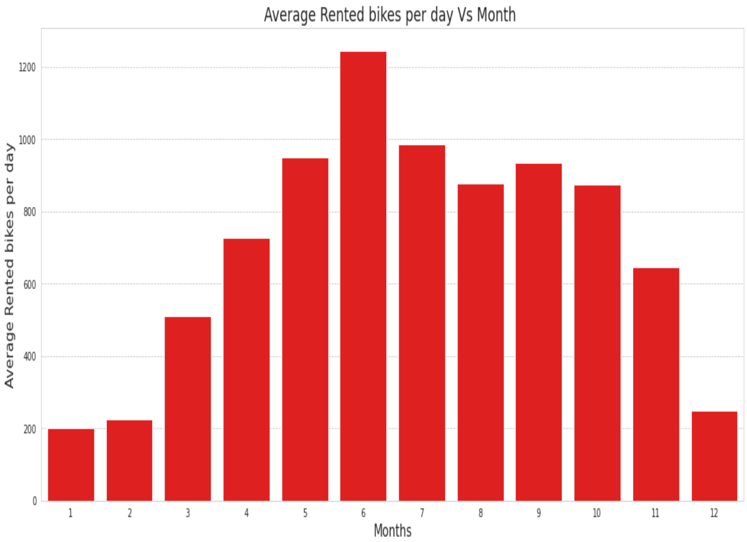
Data Insights:

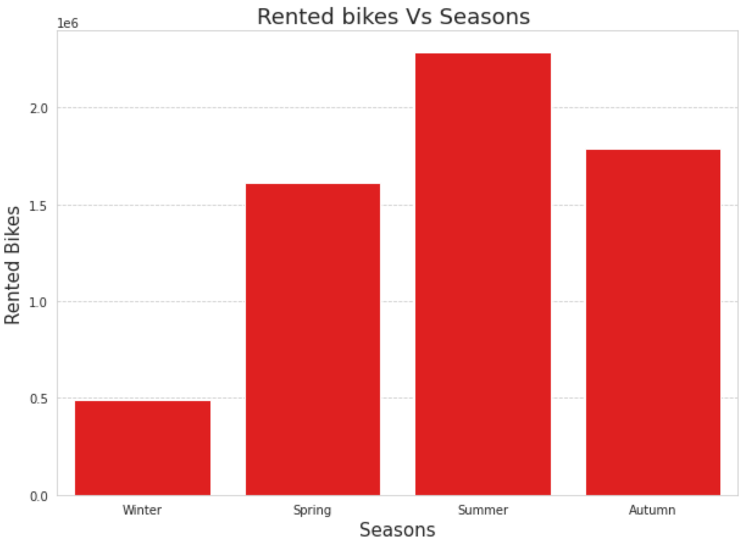
- This Dataset has 8760 Row and 14 Columns.
- There are 4 categorical features, ‘Hours’, ‘Seasons’, ‘Holiday’, & ‘functioning Day’.
- From date string we extract features like day, year and month
- No missing or duplicates values.

	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Seasons	Holiday	Functioning Day
0	01/12/2017	254	0	-5.2	37	2.2	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
1	01/12/2017	204	1	-5.5	38	0.8	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
2	01/12/2017	173	2	-6.0	39	1.0	2000	-17.7	0.0	0.0	0.0	Winter	No Holiday	Yes
3	01/12/2017	107	3	-6.2	40	0.9	2000	-17.6	0.0	0.0	0.0	Winter	No Holiday	Yes
4	01/12/2017	78	4	-6.0	36	2.3	2000	-18.6	0.0	0.0	0.0	Winter	No Holiday	Yes

Exploratory Data Analysis

- Months are extracted from the date column and then plotted against the average rented bike count.
- Season-based average rented bike analyses are shown in the second figure.

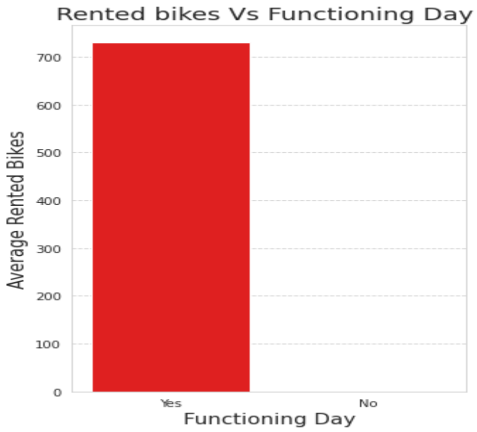


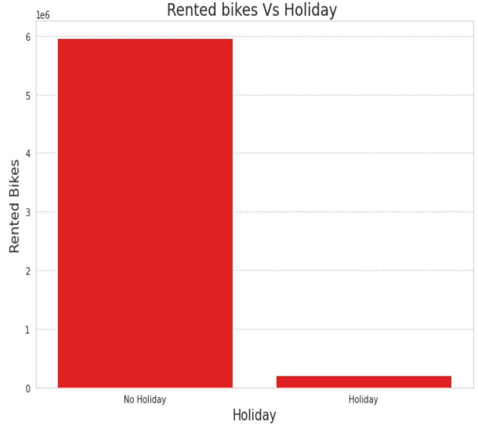


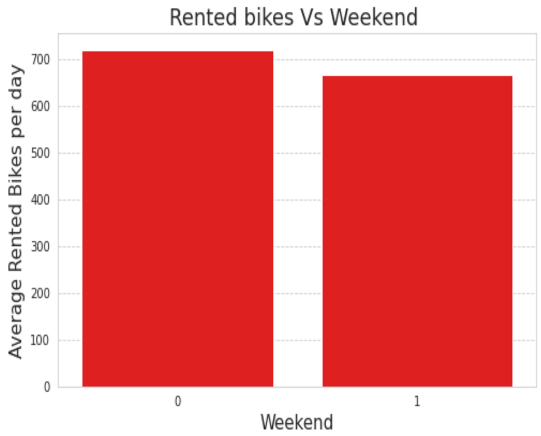
- Summer has seen highest number of rented bikes whereas number of rented bukes was least in winter.

Exploratory Data Analysis

- Analysis of Rented bikes count with respect to Functioning day, Holiday has been done which shows an almost similar result.
- The Date column has been further split into Weekdays and Weekend columns which shows an approximate equal average of rented bike counts on both the sub-categories.





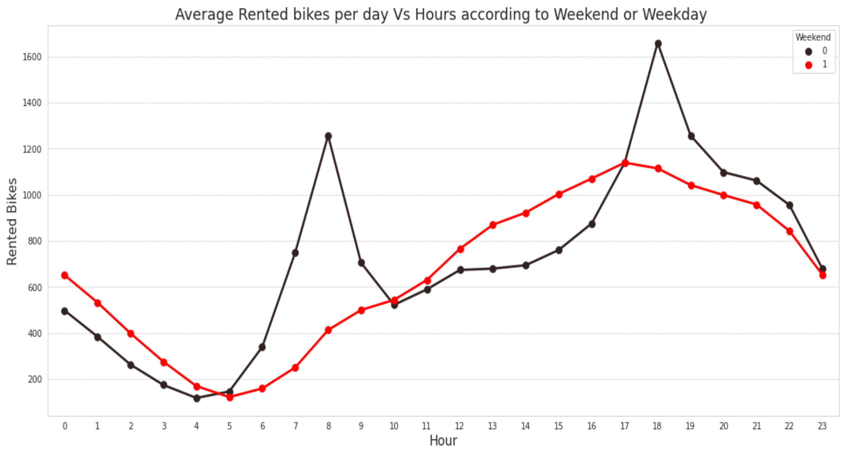


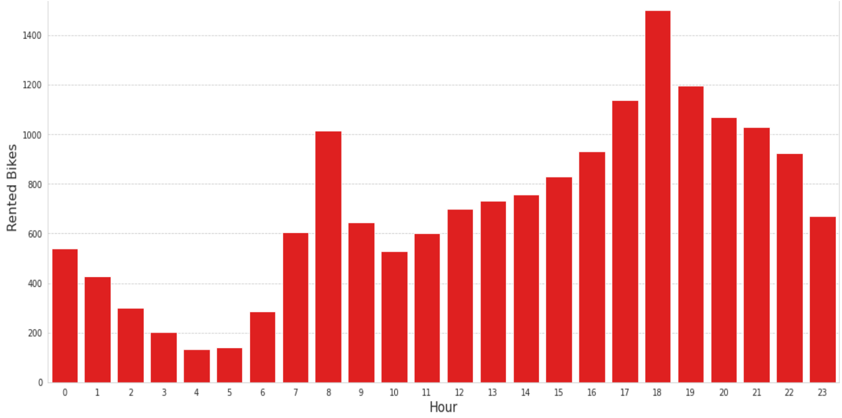
- Bikes were only rented on functioning day
- Very small number of bikes were rented on holiday as compared to non-holiday
- Weekdays have higher number of Average rented bikes than Weekends

Exploratory Data Analysis

Lets see the line plots of Rented bikes vs Hour for weekday and weekend

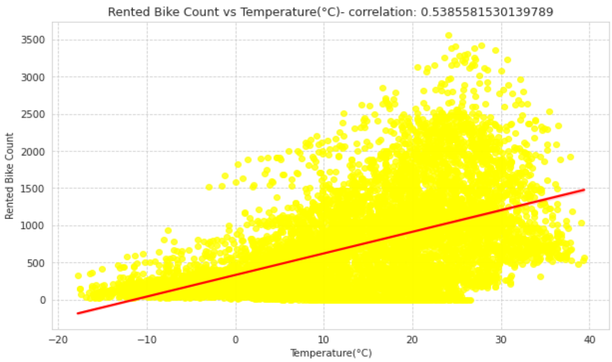
- **The plot shows that for weekends the rented bike counts remain in saddle condition while for weekdays it shows a peak at 8:00 AM and 6:00 PM which may be the result of working-class traffic**

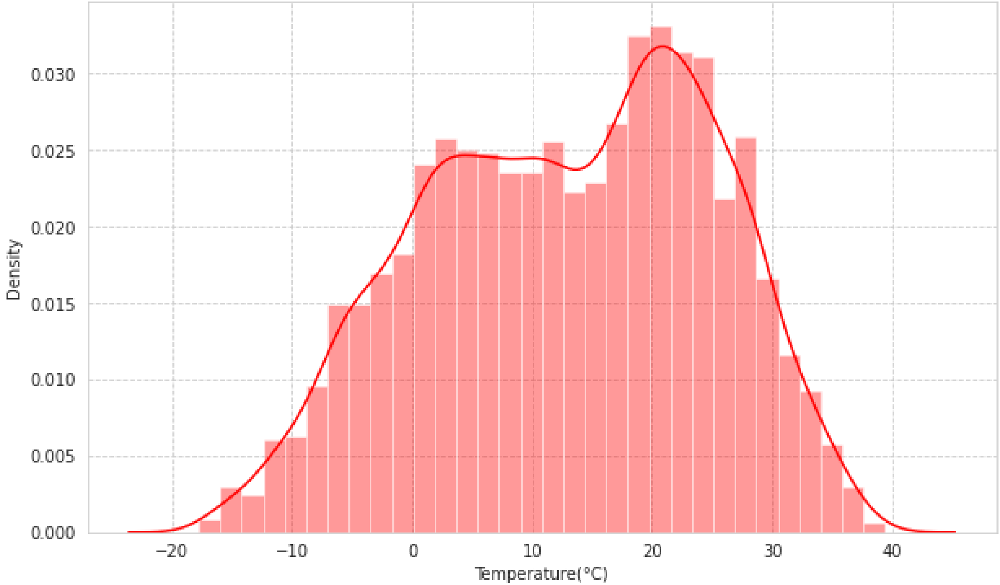




EDA on Numerical Data

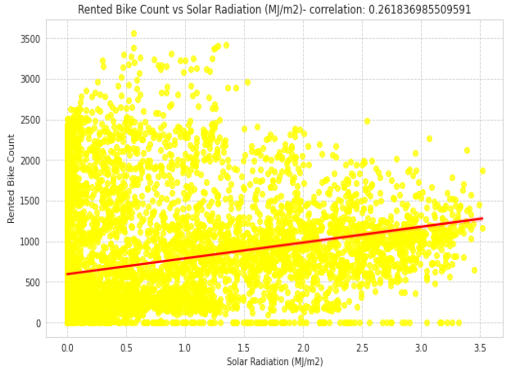
The Temperature of Seoul shows an average range of 0°C to 30 °C. The regression plot for temperature versus rented bike count shows that the Rented Bike Count is linearly proportional to the temperature although it will go to decrease if the temperature rises more than bearable.





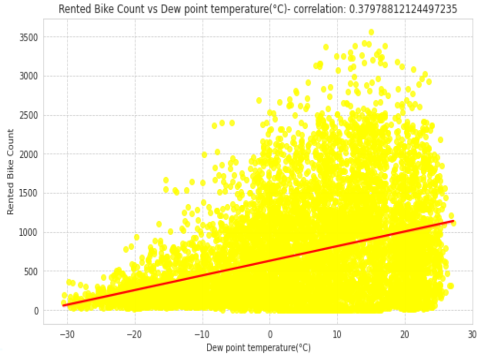
Temperature Based

Relationship b/w Rented Bike Count and Independent Variables

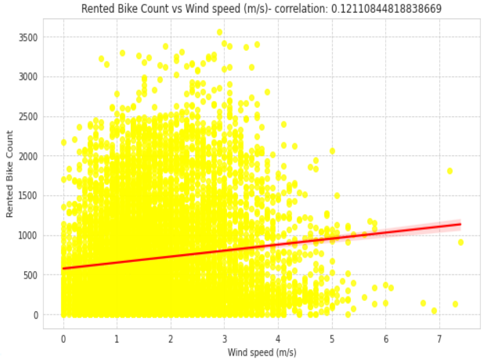


Solar Radiation

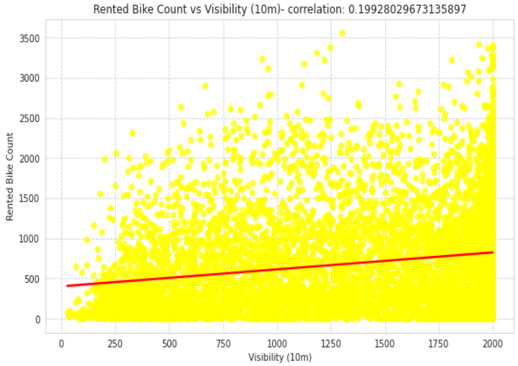
Rented Bike counts are positively correlated with features Dew Point Temperature, Solar Radiation, Wind speed, Visibility



Dew Point Temperature



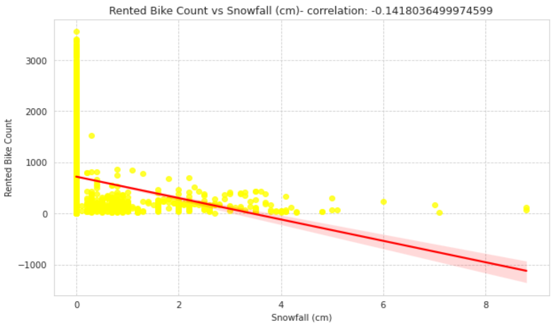
Wind Speed



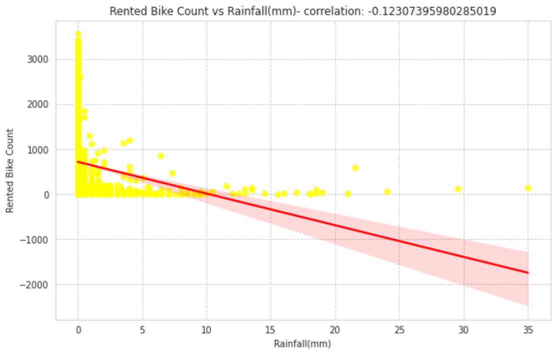
Visibility

Relationship b/w Rented Bike Count and Independent Variables

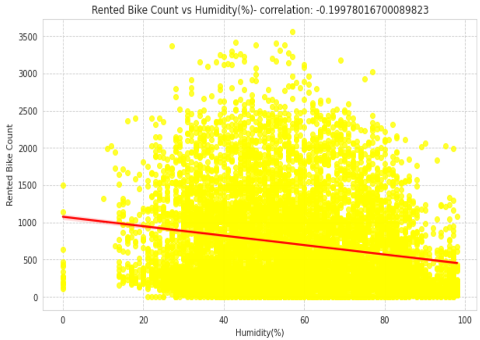
The rented bike counts are negatively correlated with Humidity, Snowfall, Rainfall



Snowfall



Rainfall



Humidity

Correlation Analysis (Before Treatment)

- The correlation matrix shows very high multicollinearity in temperature and dew point temperature.
- So either both the features should be combined into one or one of the features must be dropped and based on VIF (Variance Inflation factor)

	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Month	Weekend
Rented Bike Count	1.000000	0.410257	0.538558	-0.199780	0.121108	0.199280	0.379788	0.261837	-0.123074	-0.141804	0.133514	-0.036467
Hour	0.410257	1.000000	0.124114	-0.241644	0.285197	0.098753	0.003054	0.145131	0.008715	-0.021516	0.000000	-0.000000
Temperature(°C)	0.538558	0.124114	1.000000	0.159371	-0.036252	0.034794	0.912798	0.353505	0.050282	-0.218405	0.216183	0.007214
Humidity(%)	-0.199780	-0.241644	0.159371	1.000000	-0.336683	-0.543090	0.536894	-0.461919	0.236397	0.108183	0.139875	-0.016951
Wind speed (m/s)	0.121108	0.285197	-0.036252	-0.336683	1.000000	0.171507	-0.176486	0.332274	-0.019674	-0.003554	-0.156710	-0.022227
Visibility (10m)	0.199280	0.098753	0.034794	-0.543090	0.171507	1.000000	-0.176630	0.149738	-0.167629	-0.121695	0.064874	-0.026762
Dew point temperature(°C)	0.379788	0.003054	0.912798	0.536894	-0.176486	-0.176630	1.000000	0.094381	0.125597	-0.150887	0.242552	-0.006990
Solar Radiation (MJ/m2)	0.261837	0.145131	0.353505	-0.461919	0.332274	0.149738	0.094381	1.000000	-0.074290	-0.072301	-0.031595	0.012975
Rainfall(mm)	-0.123074	0.008715	0.050282	0.236397	-0.019674	-0.167629	0.125597	-0.074290	1.000000	0.008500	0.011958	-0.014151
Snowfall (cm)	-0.141804	-0.021516	-0.218405	0.108183	-0.003554	-0.121695	-0.150887	-0.072301	0.008500	1.000000	0.053121	-0.006759
Month	0.133514	0.000000	0.216183	0.139875	-0.156710	0.064874	0.242552	-0.031595	0.011958	0.053121	1.000000	0.012839
Weekend	-0.036467	-0.000000	0.007214	-0.016951	-0.022227	-0.026762	-0.006990	0.012975	-0.014151	-0.006759	0.012839	1.000000

Variance Inflation Factor

variables	VIF
Hour	4.418398
Temperature(°C)	33.984042
Humidity(%)	5.617480
Wind speed (m/s)	4.809775
Visibility (10m)	9.106191
Dew point temperature(°C)	17.505235
Solar Radiation (MJ/m2)	2.882383
Rainfall(mm)	1.081868
Snowfall (cm)	1.120882
Weekend	1.409388

VIF for all features

variables	VIF
Hour	3.855654
Humidity(%)	5.462400
Wind speed (m/s)	4.730040
Visibility (10m)	4.980916
Dew point temperature(°C)	1.663850
Solar Radiation (MJ/m2)	1.925305
Rainfall(mm)	1.080447
Snowfall (cm)	1.111735
Weekend	1.384555

VIF for all features except Temperature

Here is the comparison of VIFs for features with and without Temperature feature:

- **VIFs are high for Temperature and Dew Point Temperature when all the features are considered**
- **When the Temperature feature is not considered for VIFs, all VIFs for other features decreases significantly.**
- **Therefore, we decided to drop Temperature**

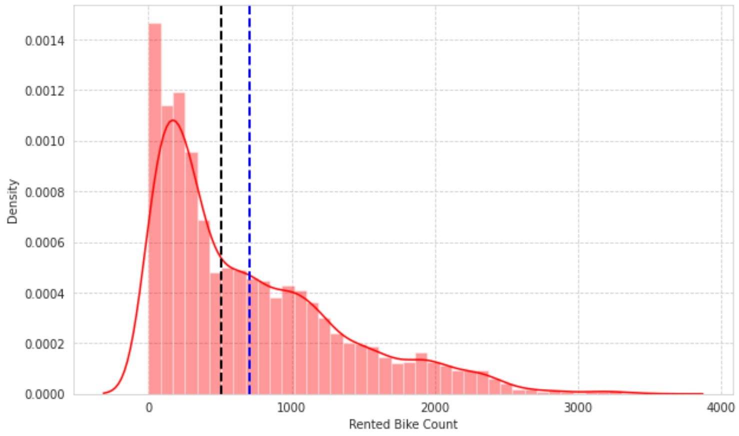
Correlation Analysis (After Treatment)

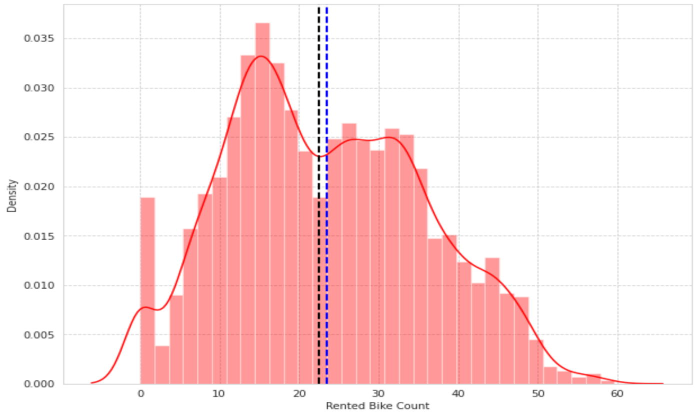
- Correlation plot after dropping the temperature feature show that there are no more highly correlated parameters present in the dataset.
- We can conclude that, there is no multicollinearity present in the dataset

	Rented Bike Count	Hour	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Month	Weekend
Rented Bike Count	1.000000	0.410257	-0.199780	0.121108	0.199280	0.379788	0.261837	-0.123074	-0.141804	0.133514	-0.036467
Hour	0.410257	1.000000	-0.241644	0.285197	0.098753	0.003054	0.145131	0.008715	-0.021516	0.000000	-0.000000
Humidity(%)	-0.199780	-0.241644	1.000000	-0.336683	-0.543090	0.536894	-0.461919	0.236397	0.108183	0.139875	-0.016951
Wind speed (m/s)	0.121108	0.285197	-0.336683	1.000000	0.171507	-0.176486	0.332274	-0.019674	-0.003554	-0.156710	-0.022227
Visibility (10m)	0.199280	0.098753	-0.543090	0.171507	1.000000	-0.176630	0.149738	-0.167629	-0.121695	0.064874	-0.026762
Dew point temperature(°C)	0.379788	0.003054	0.536894	-0.176486	-0.176630	1.000000	0.094381	0.125597	-0.150887	0.242552	-0.006990
Solar Radiation (MJ/m2)	0.261837	0.145131	-0.461919	0.332274	0.149738	0.094381	1.000000	-0.074290	-0.072301	-0.031595	0.012975
Rainfall(mm)	-0.123074	0.008715	0.236397	-0.019674	-0.167629	0.125597	-0.074290	1.000000	0.008500	0.011958	-0.014151
Snowfall (cm)	-0.141804	-0.021516	0.108183	-0.003554	-0.121695	-0.150887	-0.072301	0.008500	1.000000	0.053121	-0.006759
Month	0.133514	0.000000	0.139875	-0.156710	0.064874	0.242552	-0.031595	0.011958	0.053121	1.000000	0.012839
Weekend	-0.036467	-0.000000	-0.016951	-0.022227	-0.026762	-0.006990	0.012975	-0.014151	-0.006759	0.012839	1.000000

Feature Engineering

- One Hot Encoding of categorical feature: Hours, Seasons, and Months.
- The date-time, date, day, and temperature & season columns have been dropped from the data set.
- Ordinal Encoding: Holiday and Functioning day columns.
- Normalization has been done on the dependent variable to deal with skewness of the data and the difference between the rented bike count data plot before and after normalization is shown





Model Building Prerequisites

1. Feature Scaling: We applied standardization in order to standardize the data is a specific range. In our case, we applied **MinMax** scaler on independent features to standardize the data.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2. Test, Train split as :

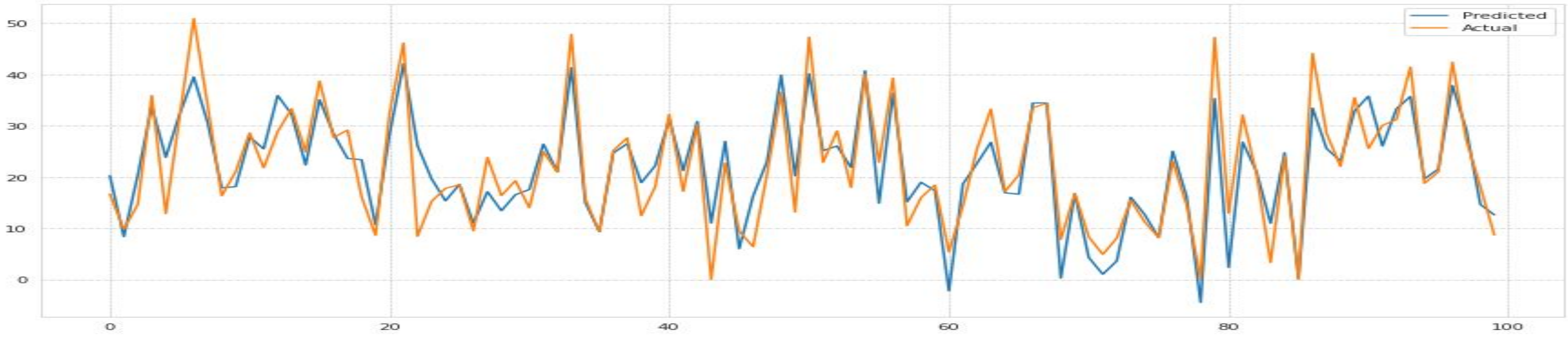
    Training data = 80% of dataset

    Testing data = 20% of dataset

Linear Regression

- Model accuracy is moderate for training as well as test data. Therefore we can conclude that no overfitting is since.
- Since there is no overfitting, we did not go ahead with Regularized linear Regression
- We plotted line graph of actual vs predicted Rented bike count

Training Errors	Testing Errors
MSE: 34.443723451189115	MSE: 34.12057506681097
MAE: 4.436644249627593	MAE: 4.365698635890322
R2: 0.779	R2: 0.774

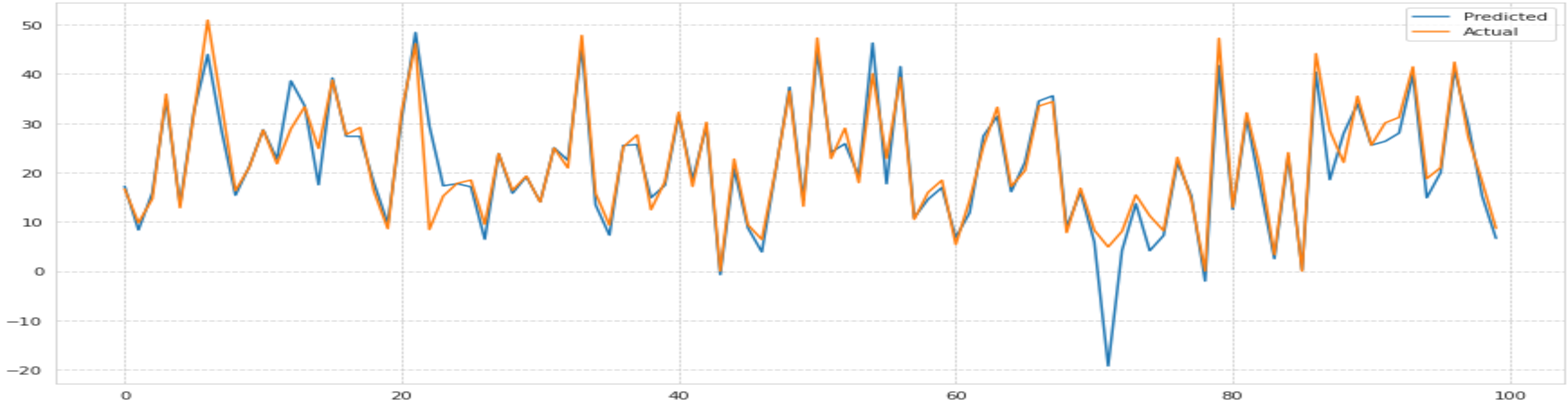


Polynomial Regression

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

Training Errors
MSE: 11.516976335573187
MAE: 2.25335580563619
R2: 0.93

Testing Errors
MSE: 14.65901362869785
MAE: 2.5455574028490577
R2: 0.9

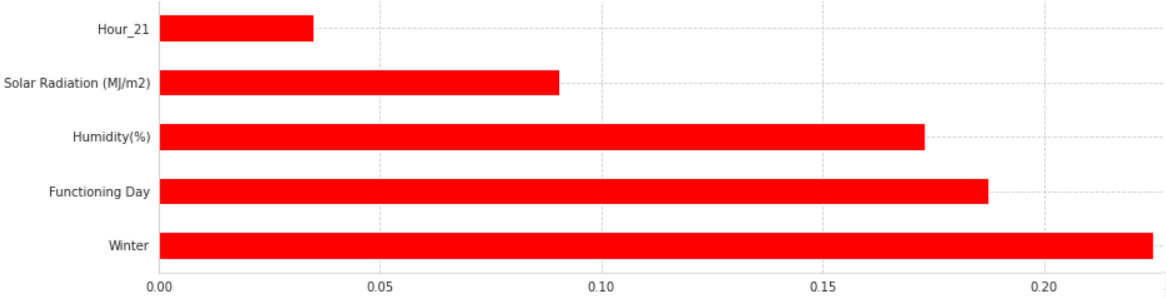


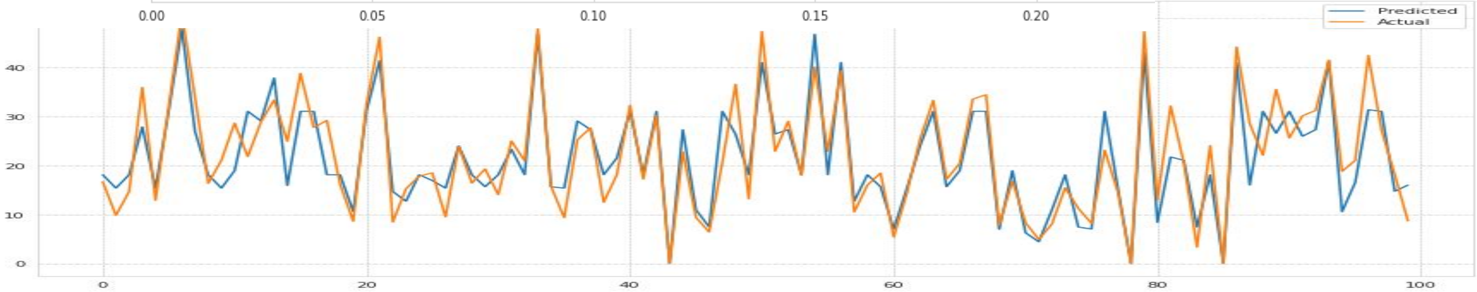
Decision Tree Regressor

- Parameters: max depth = 10, max leaf nodes = 120
- R^2 for both training and test data is moderate indicating model is fit well on both the datasets
- We plotted line graph of actual vs predicted Rented bike count and feature importance plot for top 5 features

```
Training Errors
MSE: 25.793982334841054
MAE: 3.7210179188275037
R2: 0.835
```

```
Testing Errors
MSE: 29.80739508753182
MAE: 3.9677881461060367
R2: 0.803
```



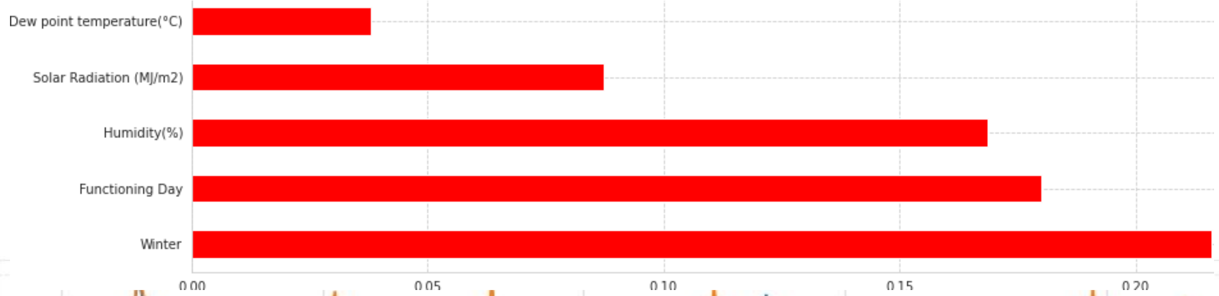


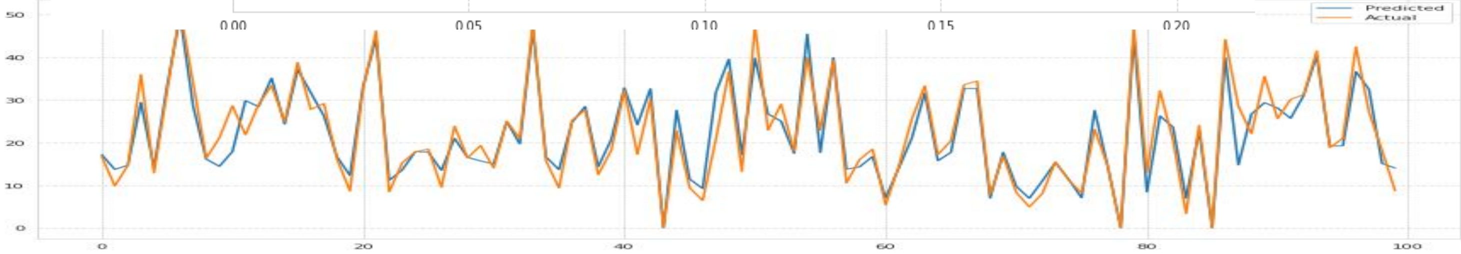
Random Forest Regressor

- Parameters: n_estimators = 180, max depth = 13, max leaf nodes = 80
- R^2 for both training and test data is moderate indicating model is fit well on both the datasets
- We plotted line graph of actual vs predicted Rented bike count and feature importance plot for top 5 features

Training Errors
MSE: 17.4107128991418
MAE: 3.10713331885730
R2: 0.888

Testing Errors
MSE: 18.8454619617235
MAE: 3.15423275122167
R2: 0.875

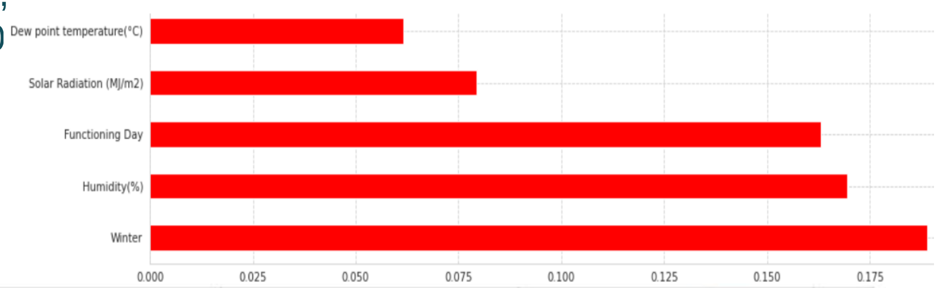


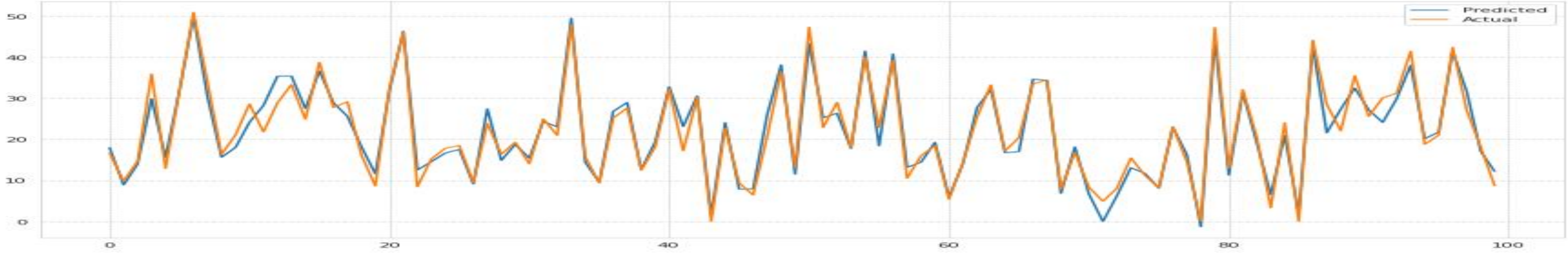


Gradient Boost with Hyper Parameter Tuning

- paramters = n_estimators = [50,80,100],
  max_depth = [4,6,8,10],
  min_samples_split = [50,80,100],
  min_samples_leaf = [40,50]
- Best parameters according to Gridsearch CV
  Best_parameters = max_depth=10,
  min_samples_leaf=40,
  min_samples_split=50

```
Training Errors              Testing Errors
MSE: 6.502066630324401       MSE: 10.078320275215765
MAE: 1.712901821228588       MAE: 2.167583140792035
R2: 0.958                    R2: 0.933
```

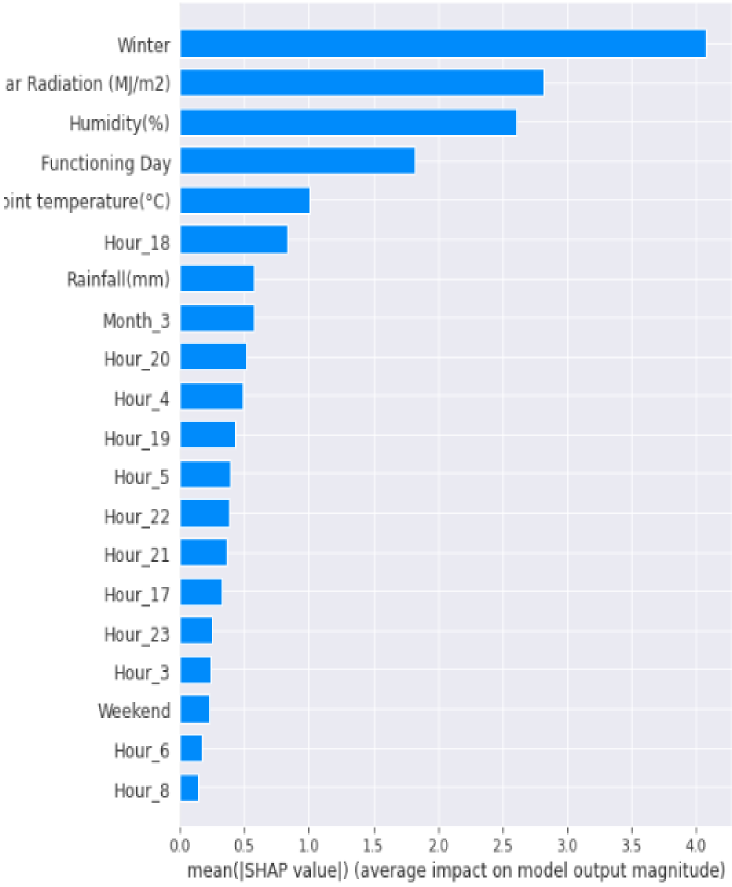




Model Explainability

- Higher value indicates that, that feature impacts highly on the dependent variable and vice versa

- Here we have used bar graph to explain top 20 mean **SHAP** values in Gradient Boost. We can see **Winter** has the highest feature, followed by **Solar Radiation** and **Humidity**



Conclusion

- **EDA**
  - In summer season, highest number of bikes were rented as compared to other seasons
  - Higher number of Bikes were rented on weekday as compared to weekends
  - Lowest number of bikes were rented in January and after gradually increasing, highest number of bikes were rented in the month of June
  - Number of Bikes Rented is at its peak at 6 PM
  - Bikes are rented most on a clear day, i.e. where there is no snowfall or rainfall

Conclusion

- ## Models

	Training set				Test set			
	Model	MAE	MSE	R2_score	Model	MAE	MSE	R2_score
0	Linear regression	4.437	34.444	0.779	Linear regression	4.366	34.121	0.774
1	Polynomial regression	2.253	11.517	0.926	Polynomial regression	2.546	14.659	0.903
2	Decision Tree Regression	3.721	25.794	0.835	Decision Tree Regression	3.968	29.807	0.803
3	Random Forrest	3.107	17.411	0.888	Random Forrest	3.154	18.845	0.875
4	Gradient Boost with GridSearch	1.713	6.502	0.958	Gradient Boost with GridSearch	2.168	10.078	0.933

- Gradient Boost with optimal parameters selected by applying GridSearchCV gives the best fit with highest R^2 score and lowest MSE and MAE followed by polynomial regression
- Although Gradient boost gives slightly better accuracy than polynomial regression, it should also be kept in mind that Polynomial regression requires much lesser computational time.

Thank you