Capstone Project

Bike Sharing Demand Prediction

By
Saugata Deb
Shreyash Movale
Ankit Patil
Naga Sai Kiran

Objective

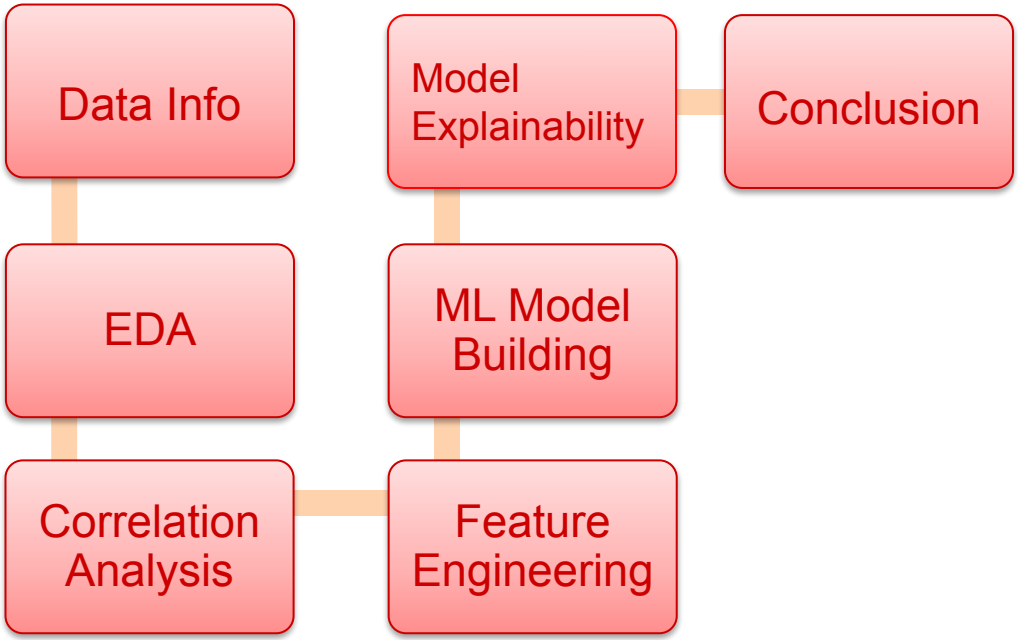
- Rental bike system plays a crucial part in public transport to increase the mobility of traffic in any city.
- Bike-sharing gains a vast range of attention in recent years as part of initiatives to boost the use of cycles, improve the first mile/last mile link to other modes of transportation, and 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. 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.

Methodology

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



Data Insights…

- The data set has 13 variables of which Rented Bike Count is a Dependent variable and the rest are independent variables.
- The size of the data is (8760,13) i.e., we have 8760 rows with 13 columns
- None of the data have null values so we don't have to clean data.
- Data Set is a mixture of categorical and numerical data so we have to arrange and encode the data before feeding it to the ML model.

```
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)
```

Feature Description:-

Date : Date feature which is **str** type is needed to convert it into Datetime format DD/MM/YYYY.

Rented Bike Count : Number of bike rented which is our Dependent variable according to our problem statement which is **int** type.

Hour: Hour feature which is in 24 hour format which tells us number bike rented per hour is **int** type.

Temperature(°C): Temperature feature which is in celsius scale**(°C)** is **Float** type.

Humidity(%): Feature humidity in air (%) which is **int** type.

Wind speed (m/s) : Wind Speed feature which is in (m/s) is **float** type.

Visibility (10m): Visibility feature which is in 10m, is **int** type.

Feature Description:-

Dew point temperature(°C): Dew point Temperature in **(°C)** which tells us temperature at the start of the day is **Float** type.

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

Rainfall(mm): Rainfall feature in mm which indicates 1 mm of rainfall which is equal to 1 litre of water per metre square is **Float** type.

Snowfall (cm): Snowfall in cm is **Float** type.

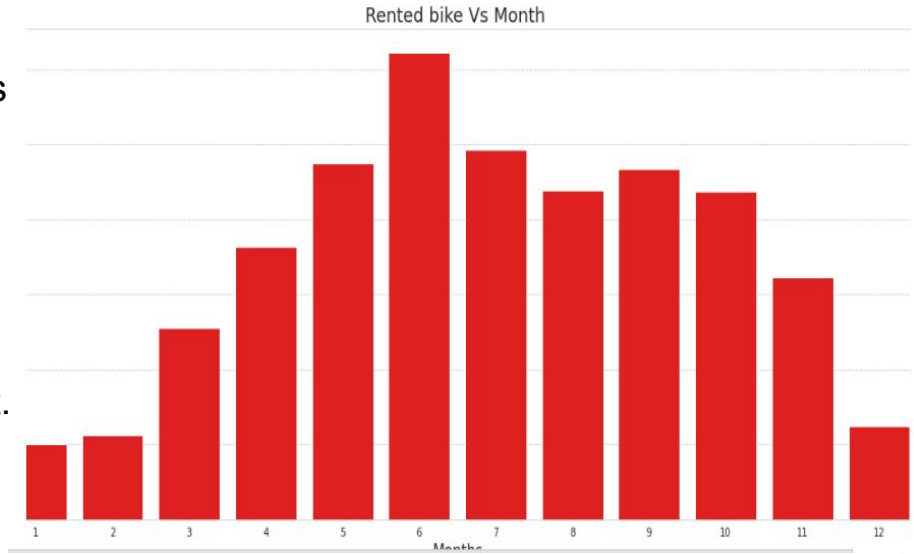
Seasons: Season, in this feature four seasons are present in data is **str** type.

Holiday: whether no holiday or holiday can be retrieved from this feature is **str** type.

Functioning Day: Whether the day is Functioning Day or not can be retrieved from this feature is **str** type.

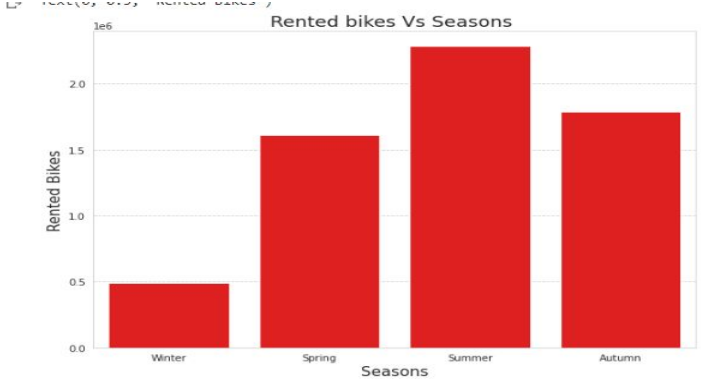
Exploratory Data Analysis

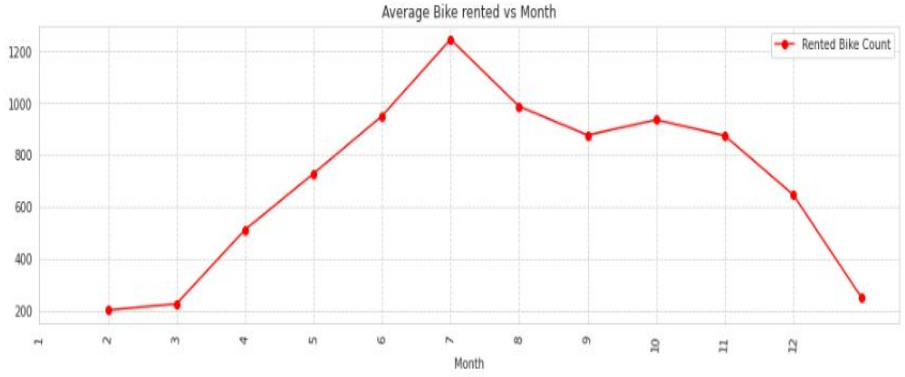
- This is a bar graph between rented bike count per month.
- Months are extracted from the date column and then plotted against the rented bike count.
- Here we can see that in the highest bike was rented in the month of June while lowest bike was rented in the month of January.
- From this we can assume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



Exploratory Data Analysis

- This is the Line graph to show the average bike rented per month in order to find correct data of bike being rented.
- From this visualisation in **Average Bike Rented vs Month** we can clearly see that Average Bike rented in **July** was highest around **1250** and Average Bike Rented in the month of **February** was the Lowest with just **200** average bike.

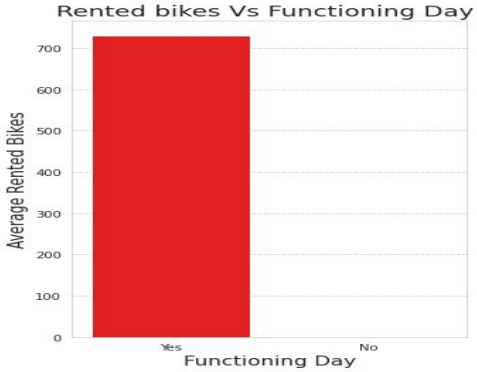


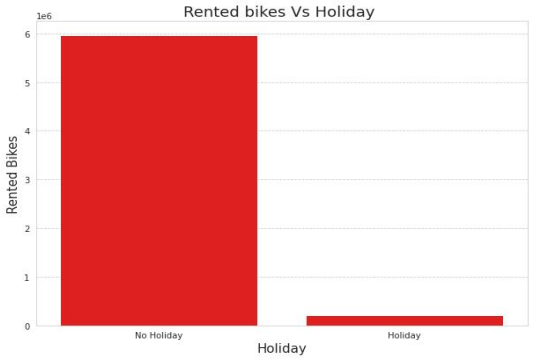


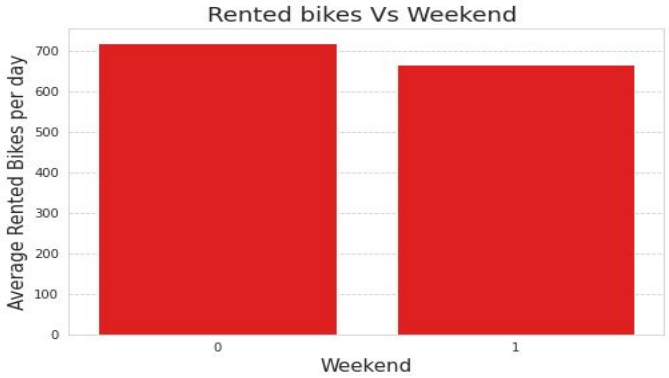
- From this we can conclude our assumption what we have assumed in the previous slide that average bike rented during summer season was highest while in winter average bike rented was lowest.

Exploratory Data Analysis

- Analysis of Rented bikes count with respect to Functioning day, Holiday has been done which shows an almost similar result.
- Also we can see that bike was rented only on non Functioning day and where there is no holiday bike was rented highest than holiday. We can also conclude from here that people tends to commute more to earn their livelihood than on holidays.
- 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. Here we can see that bike rented on weekdays are slightly higher than weekends.





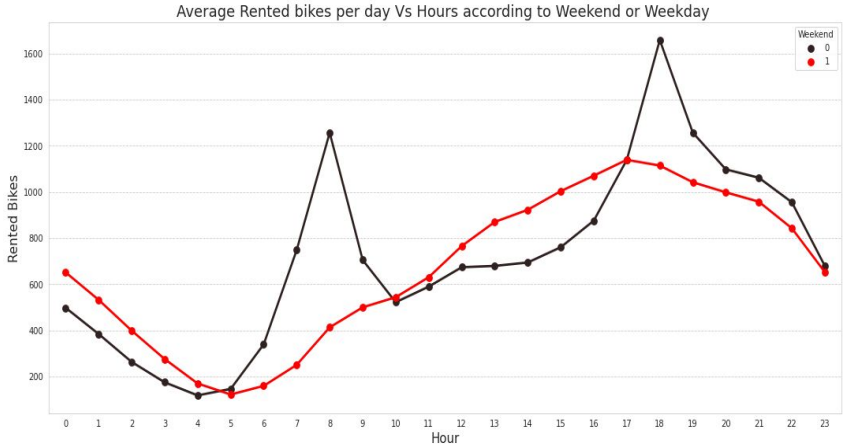


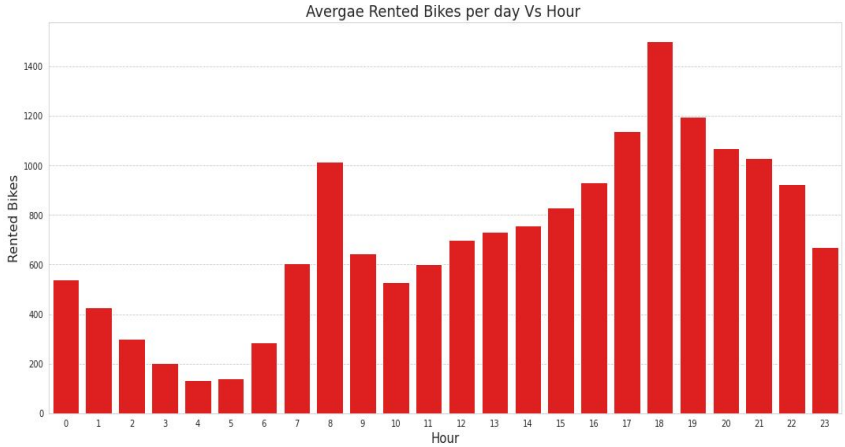
Exploratory Data Analysis

- In Average Bike Rented vs Hour we can clearly see that at 6:00 PM average number of bike rented by the people was 1550. While at 00.00 or at midnight average bike rented was lowest with just around 550 bikes which were on weekdays.

- In Average Bike Rented vs Hour we can also see that at 5:00 PM average number of bike rented by the people was around 1150. While at 00.00 or at midnight average bike rented was lowest with just around 650 bikes which were on weekend.

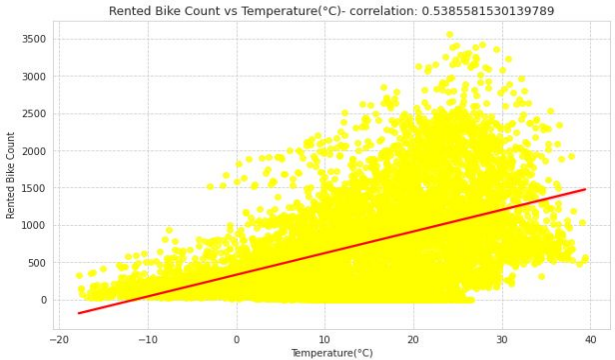
- The plot shows that for weekends the rented bike counts remain in saddle condition while for weekdays it shows a peak at 8:00 AM and 6:00 PM which may be the result of working-class traffic while the trend in weekend pattern corresponds to probably tourists who typically are casual users who rent/drop off bikes uniformly during the day and tour the city.

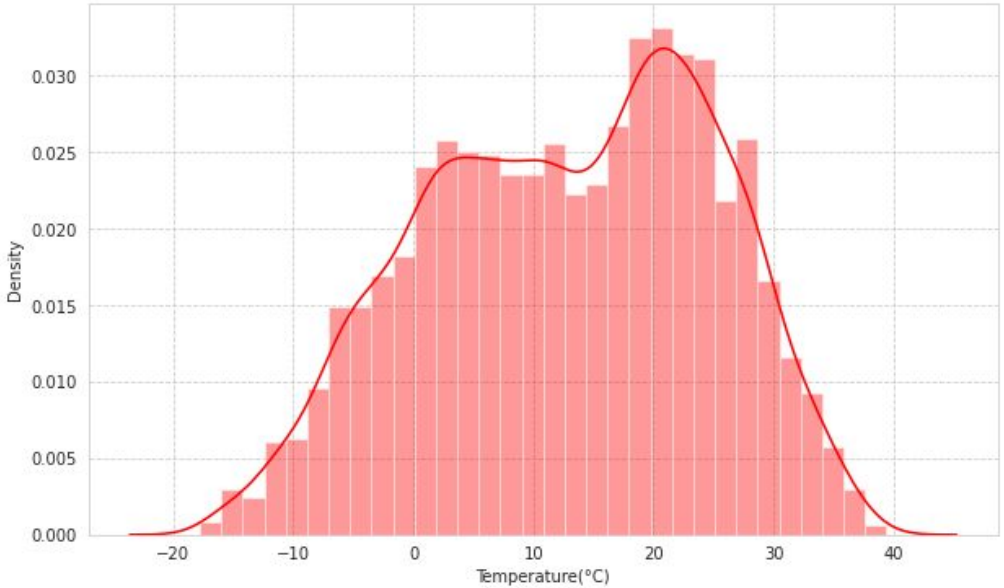




EDA on Numerical Data

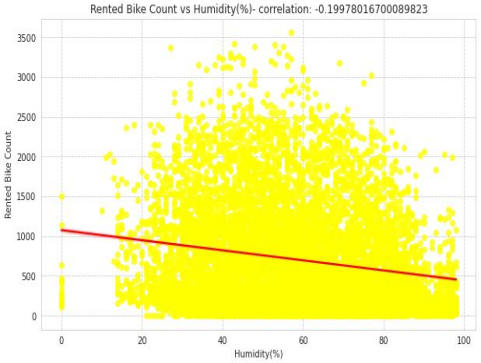
The Temperature of Seoul shows an average range of 0°C to 30 °C. The regression plot for temperature versus rented bike count shows that the Rented Bike Count is linearly proportional to the temperature although it will go to decrease if the temperature rises more than bearable.





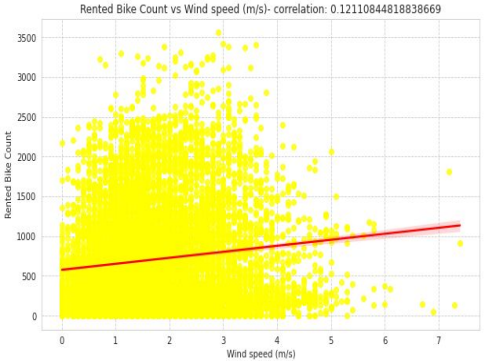
Temperature Based

Regression plots of Humidity, Wind speed & Visibility

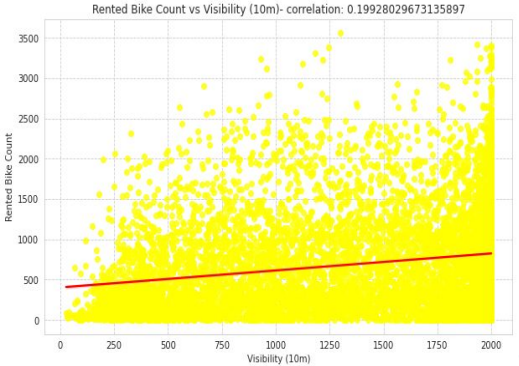


Humidity

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.

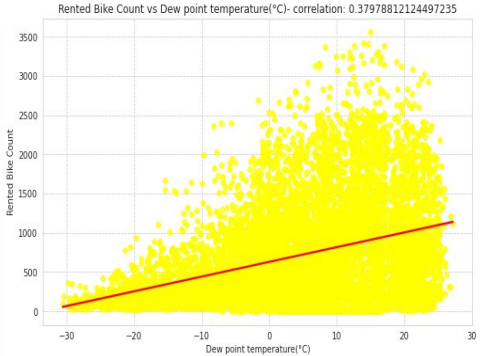


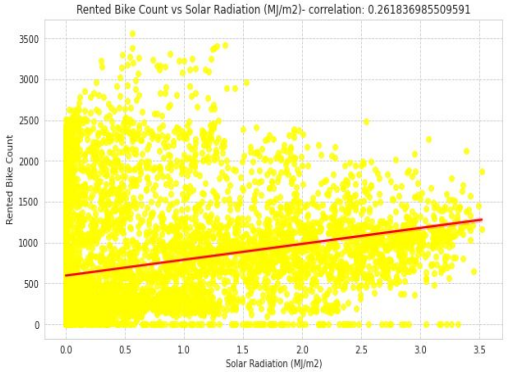
Wind Speed

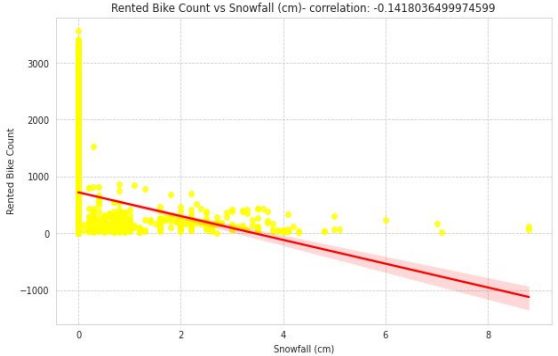


Visibility

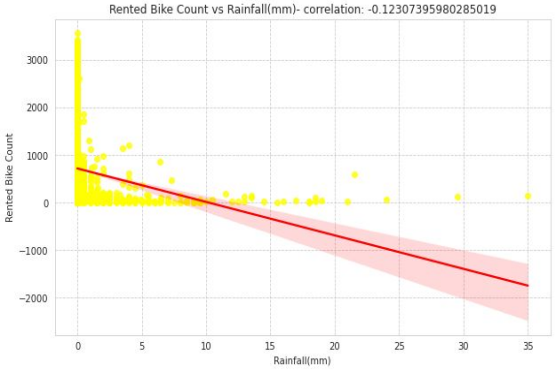
Regression plots of DPT, Solar Radiation, Snowfall & Rainfall







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)

	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	0.139675	-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	0.139675	-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**

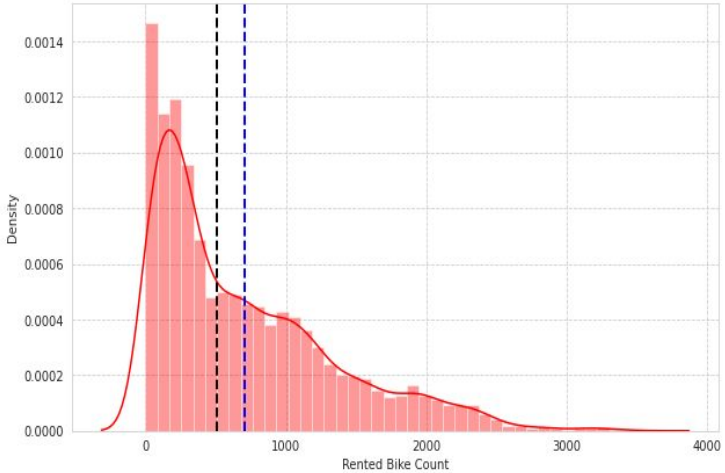
Correlation Analysis (After Treatment)

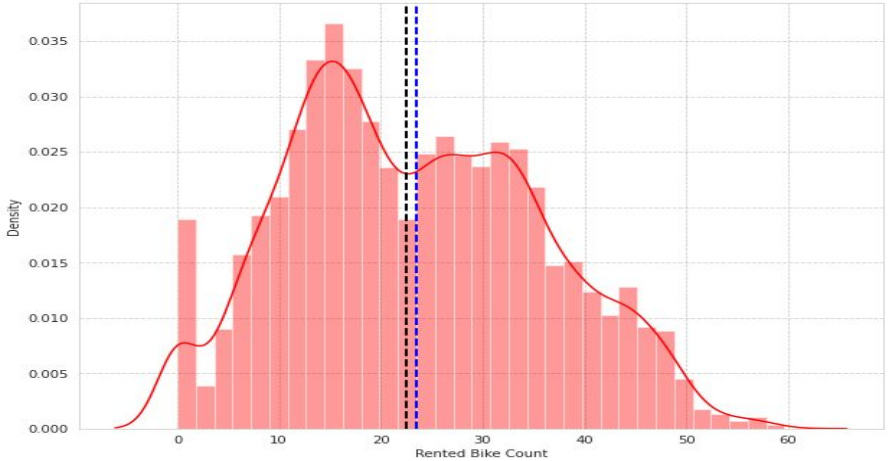
- Correlation plot after dropping the temperature feature show that there are no more highly correlated parameters present in the dataset.
- We can conclude that, there is no multicollinearity present in the dataset

	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

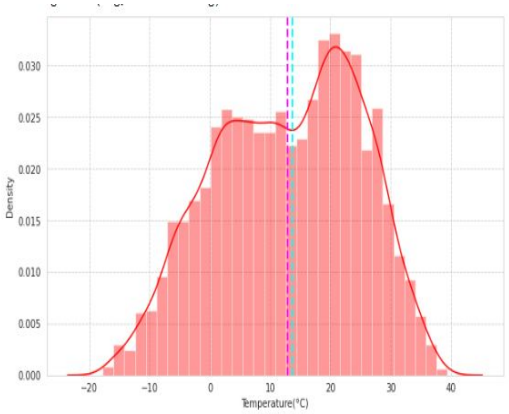
- One Hot Encoding of categorical feature: Hours, Seasons, and Months.
- The date-time, date, day, and temperature & season columns have been dropped from the data set.
- Ordinal Encoding: Holiday and Functioning day columns.
- Normalization has been done on the dependent variable to deal with skewness of the data and the difference between the rented bike count data plot before and after normalization is shown
- In density plot for Rented Bike Count we can see the median and mean lies in range of 500 to 1000 mean is slightly greater than median which means its positively skewed. Similarly we can upon normalizing the bike rented data using sqrt we can see the skewness decreases and showing distribution close to normal distribution.

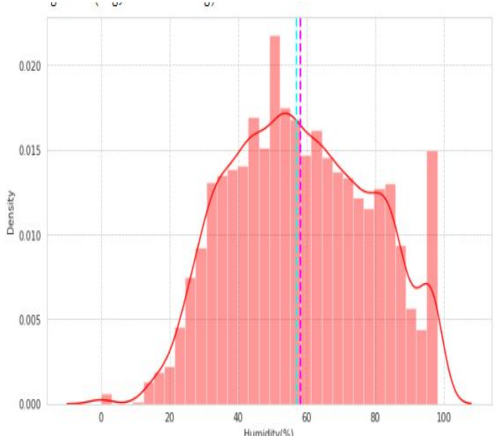


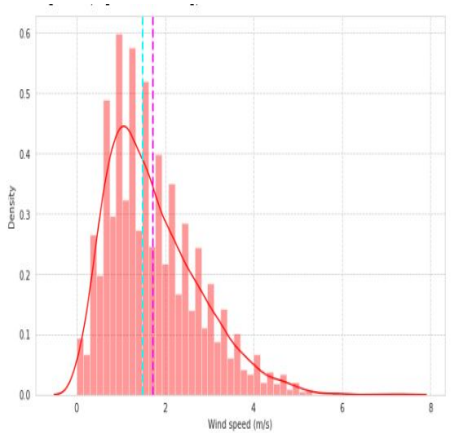


Feature Engineering(Continued..,)

- In density plot for **Temperature** we can see that median is greater than mean we can say to some extent that this is negatively skewed.
- In density plot for **Humidity** we can see that mean is greater than median we can say to some extent that this is positively skewed.
- In density plot for **Wind Speed** we can see that mean is greater than median we can say to some extent that this is positively skewed.

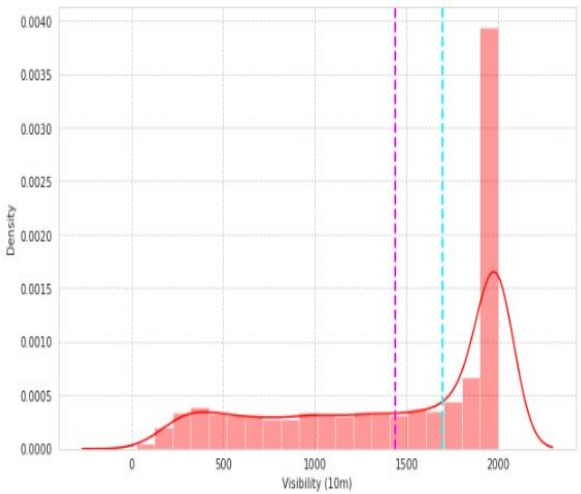


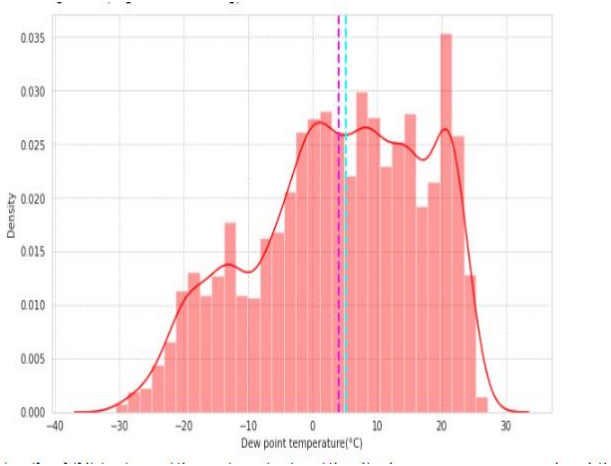


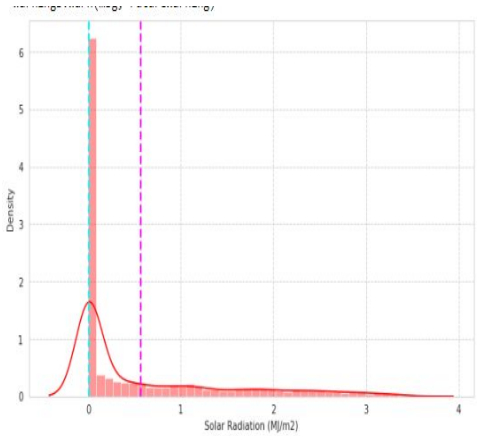


Feature Engineering(Continued…)

- In density plot for **Visibility** we can see that median is greater than mean we can say to some extent that this is negatively skewed.
- In density plot for **Dew Point Temperature** we can see that median is greater than mean we can say to some extent that this is negatively skewed.
- In density plot for **Solar Radiation** we can see that mean is greater than median we can say that this is positively skewed.





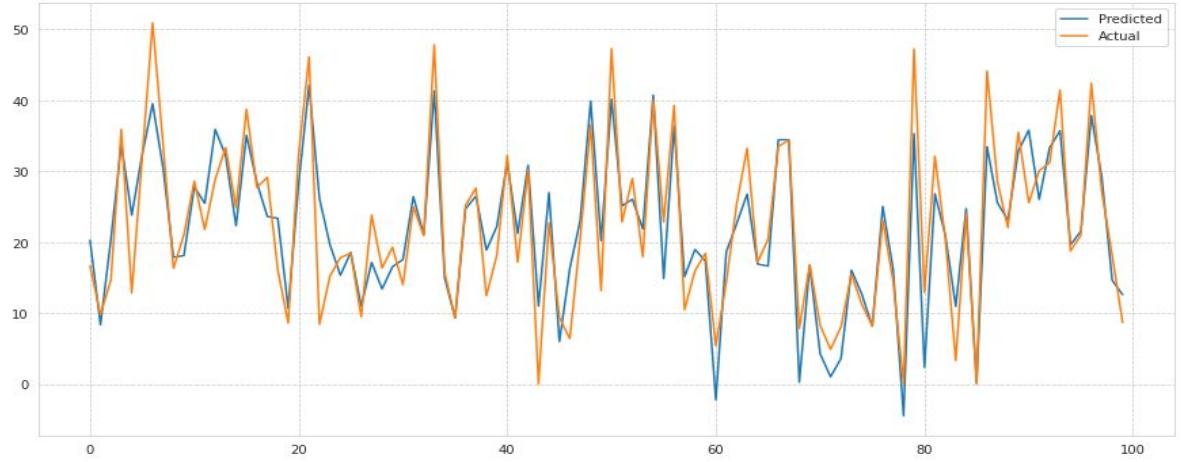


Linear Regression

- Model accuracy is moderate for training as well as test data. Therefore we can conclude that no overfitting.
- 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.

Testing Errors
MSE: 34.12057506681097
MAE: 4.365698635890322
R2: 0.774

Training Errors
MSE: 34.443723451189115
MAE: 4.436644249627593
R2: 0.779

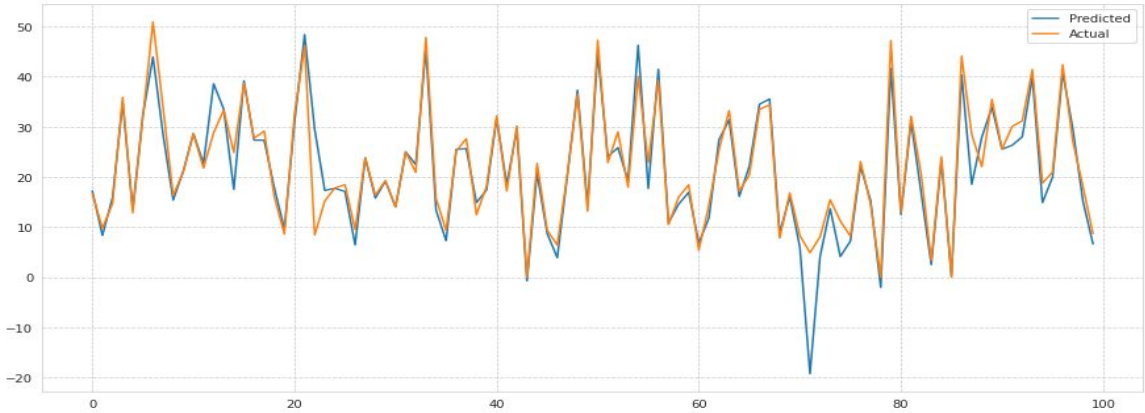


Polynomial Regression

- 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

Testing Errors
MSE: 14.65901362869785
MAE: 2.5455574028490577
R2: 0.9

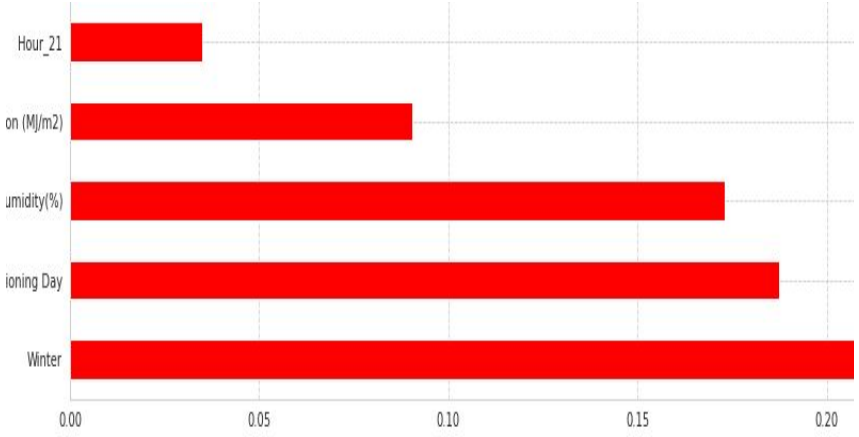
Training Errors
MSE: 11.516976335573187
MAE: 2.25335580563619
R2: 0.93

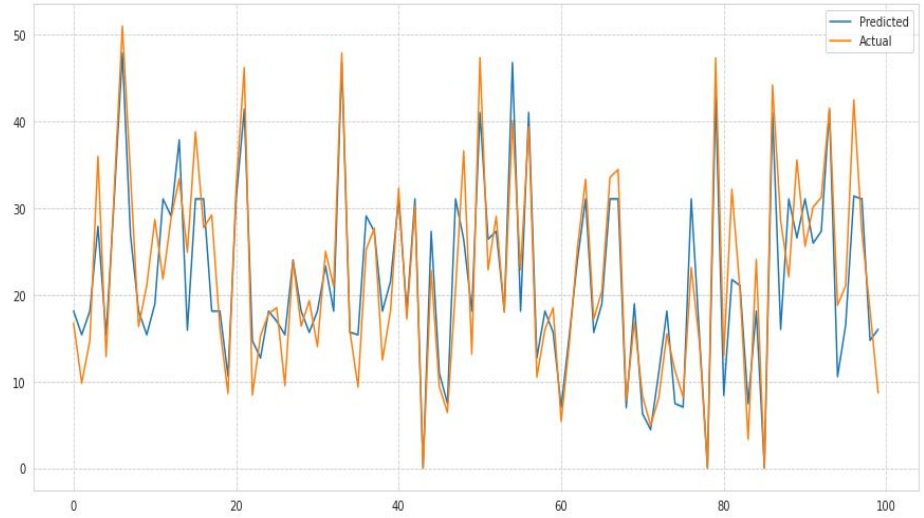


Decision Tree Regressor

- Parameters: max depth = 10, max-leaf nodes = 120
- 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 and feature importance plot for the top 5 features
- Here we can see **Hour_20** is showing least feature importance while **Winter** season is showing highest feature importance in model prediction.

Training Errors	Testing Errors
MSE: 25.793982334841054	MSE: 29.80739508753182
MAE: 3.7210179188275037	MAE: 3.9677881461060367
R2: 0.835	R2: 0.803

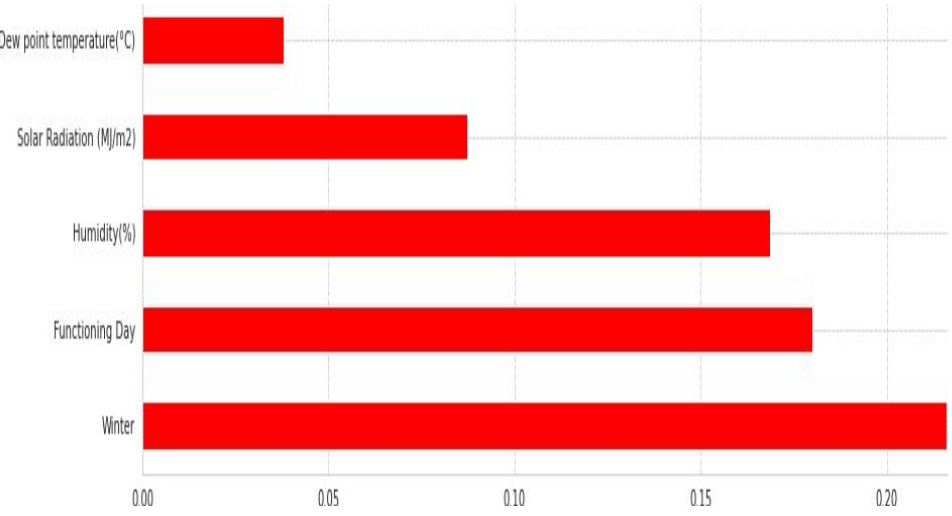


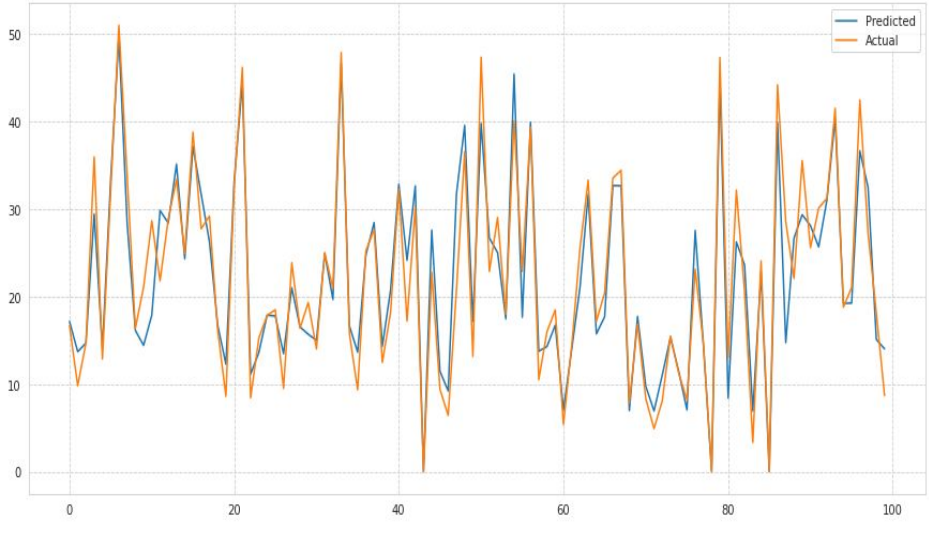


Random Forest Regressor

- Parameters: n_estimators = 180, max depth = 13, max-leaf nodes = 80
- R^2 for both training and test data is moderate indicating the model is fit well on both the datasets
- We plotted a line graph of actual vs predicted Rented bike count and feature importance plot for the top 5 features
- Here we can see Month_3 is showing least feature importance while Winter season is showing highest feature importance in model prediction.

Training Errors	Testing Errors
MSE: 17.41071289914180	MSE: 18.8454619617235
MAE: 3.107133318857307	MAE: 3.154232751221673
R2: 0.888	R2: 0.875

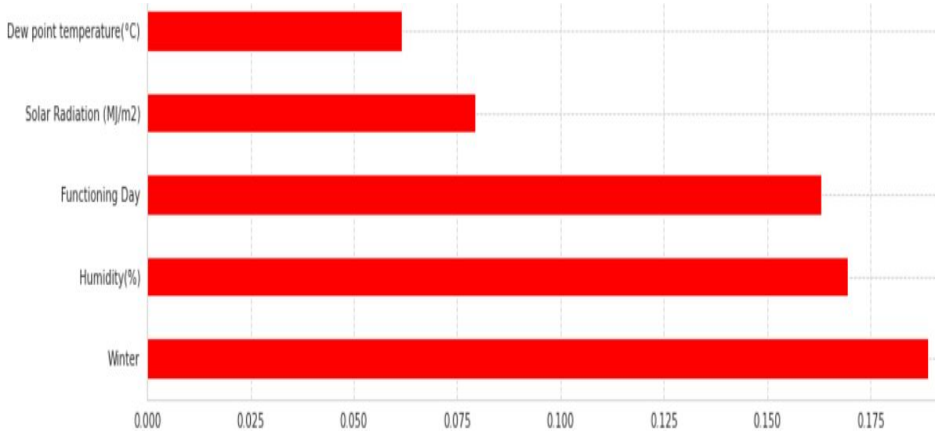


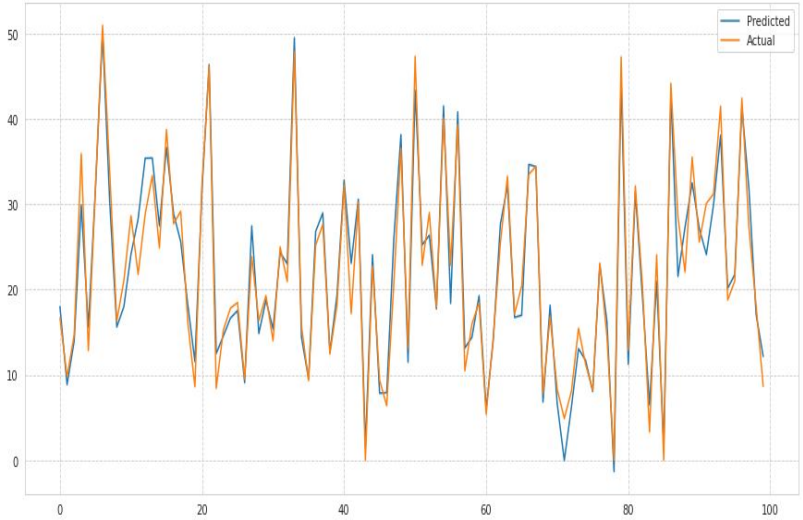


Gradient Boost with Hyper Parameter Tuning

- parameters = 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 Gridsearchcv
- Best_parameters = max_depth=10, min_samples_leaf=40, min_samples_split=50
- Here we can see Hour_19 is showing least feature importance while Winter season is showing highest feature importance in model prediction.

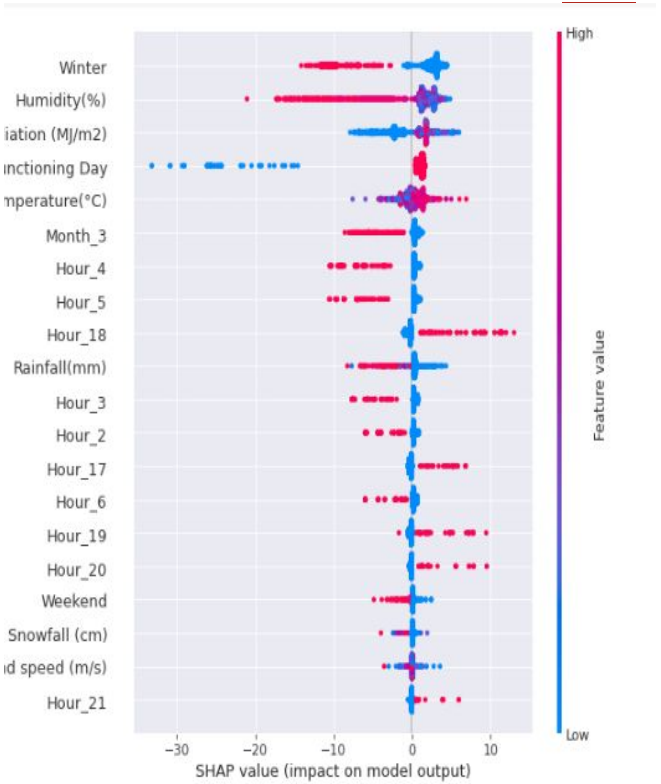
Training Errors	Testing Errors
MSE: 6.502066630324401	MSE: 10.078320275215765
MAE: 1.712901821228588	MAE: 2.167583140792035
R2: 0.958	R2: 0.933

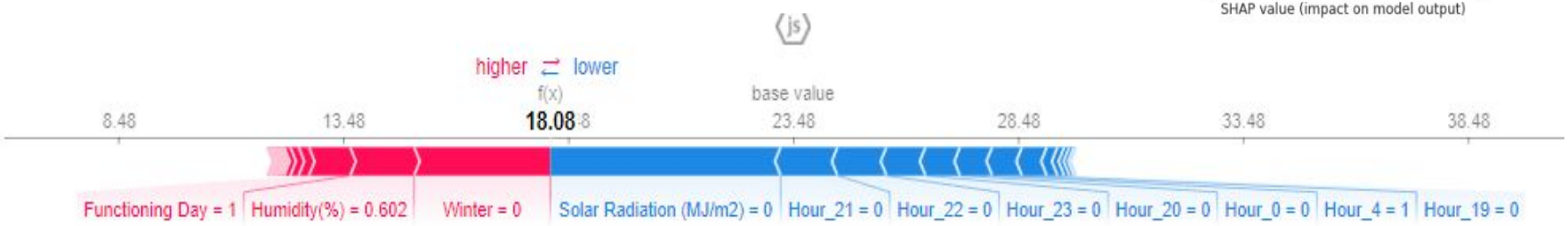




Model Explainability

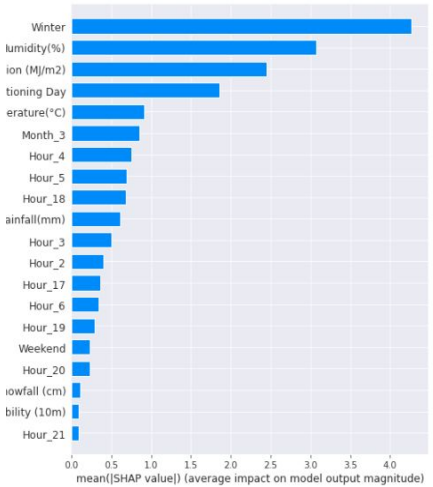
- We have used SHAP JS visualisation to explain feature values and their importance in each models. Here we can see negative feature or blue color block pushes the prediction toward left over base value. Also we can see highest Solar radiation value which causing prediction negative while winter which has the high positive value is causing positive model prediction and it is common for Decision Tree.Random Forest and Gradient Boost models.
- Also we can see from SHAP summary that high **Hour_18** value increasing prediction. Also we can see low **Snowfall** value increasing prediction and it is also common phenomenon in all the models.

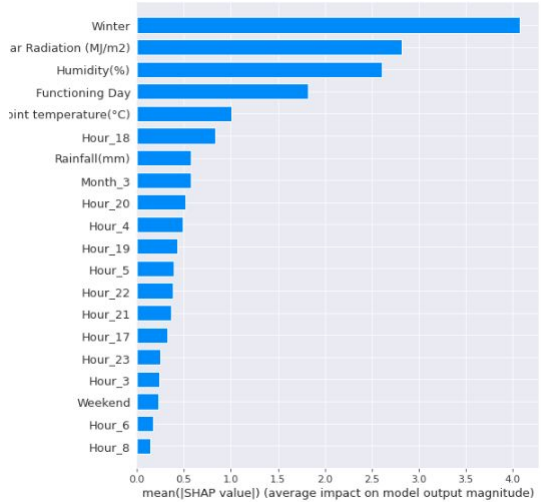


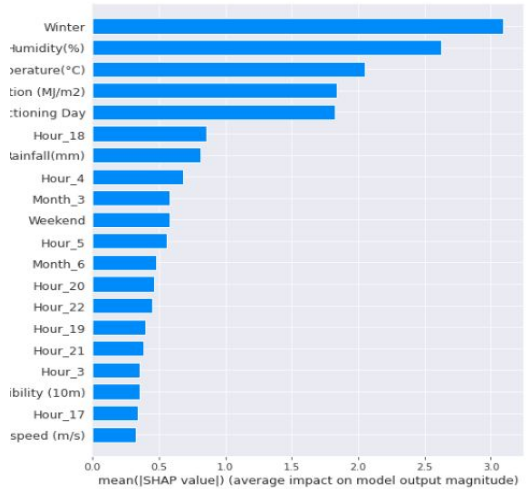


Model Explainability(Continued..)

- Here we have used bar graph to explain mean **SHAP** values in each models.In Decision Tree starting from left we can see **Winter** has the highest feature value while **Hour_21** has the Lowest feature_value
- In Random Forest Model we can see **Winter** has the highest feature value while **Hour_8** has the Lowest shap value.
- In bar graph we can see **Winter** has the highest feature value while **Wind Speed** has the Lowest shap value.
- From we can conclude that Hour_21,Hour_8 and Wind Speed is not contributing in Decision Tree,Random Forest and Gradient Boost in model prediction.







Conclusion

- In the summer season the highest number of bikes were rented as compared to other seasons
- Higher number of Bikes were rented on a weekday as compared to weekends
- Lowest number of bikes were rented in January and after gradually increasing, the highest number of bikes were rented in May
- Bike Rental is at its peak at 6 PM
- Bikes are rented most on a clear day, i.e. where there is no snowfall or rainfall
- In Hour vs Rented Bike Count we can see that during 18:00 Hrs(i.e 6:00 PM) highest number of bike was rented as compared to 5:00 Hrs(i.e 5:00 AM). This means people tends to rent less bikes at early morning.
- In Rainfall vs Rented Bike Count and similarly with Snowfall vs Rented Bike Count we can see that people tend to rent highest number of bikes during 0.00mm of Rainfall or no rainfall and 0.00cm of snowfall or no snowfall as compared to when there is actually rainfall or snowfall. In other words people rent less bikes or no bikes with the increase of rainfall or snowfall.
- In month vs Rented Bike Count we can see that people tends to rent more bike in 6 or june month as compared to less bike during dec or january.From this we can assume that people tends to rent more bikes in summer as compared to winter.
- In weekend vs Rented Bike count we can see that people tends to rent more bike during weekdays as compared to weekends.
- In Average Bike Rented vs Hour we can clearly see that at 6:00 PM average number of bike rented by the people was 1550. While at 00.00 or at midnight average bike rented was lowest with just around 550 bikes.
- In Average Bike Rented vs Month we can clearly see that Average Bike rented in July was highest around 1250 and Average Bike Rented during month of February was the Lowest with just 200 average bike.

Conclusion

- After applying linear regression model, we got R2 score of 0.779 for training data and R2 score of 0.774 for test data, which signifies that model is optimally fit on both training and test data i.e. no overfitting is seen.
- Therefore, for even better fit, we applied polynomial regression model with degree = 2, we got R2 score of 0.933 for training data and 0.90 for test data
- We also tried Tree based classifiers for our data, we applied Decision Tree Regressor, since decision tree is prone to overfit, we gave certain parameters like maximum depth of the tree, maximum leaf nodes etc, with that we we got R2 score of 0.835 for training data and 0.803 for test data which is less than polynomial regression.
- To get better accuracy on tree based model, we applied Random forest with n_estimator as 180 and with maximum depth as 13, with that we got R2 score of 0.888 for training data and 0.875 for test data.
- Finally, we applied Gradient boost with parameters selected after grid search which resulted in highest R2 score of 0.958 for training data and 0.933 for test data with very less mean squared error of 6 and 10 in training as well as in test data.
- Also we can see from SHAP summary that high **Hour_18** value increasing prediction. Also we can see low **Snowfall** value increasing prediction and it is a common phenomenon in all the models.
- Lastly, In bar graph from SHAP we can see **Winter** has the highest feature value while **Wind Speed** has the Lowest shap value.We can conclude that Hour_21,Hour_8 and Wind Speed is not contributing in Decision Tree,Random Forest and Gradient Boost in model prediction.

Thank you