Capstone Project

Bike Sharing Demand Prediction

By
Sourav Karmakar

Objective

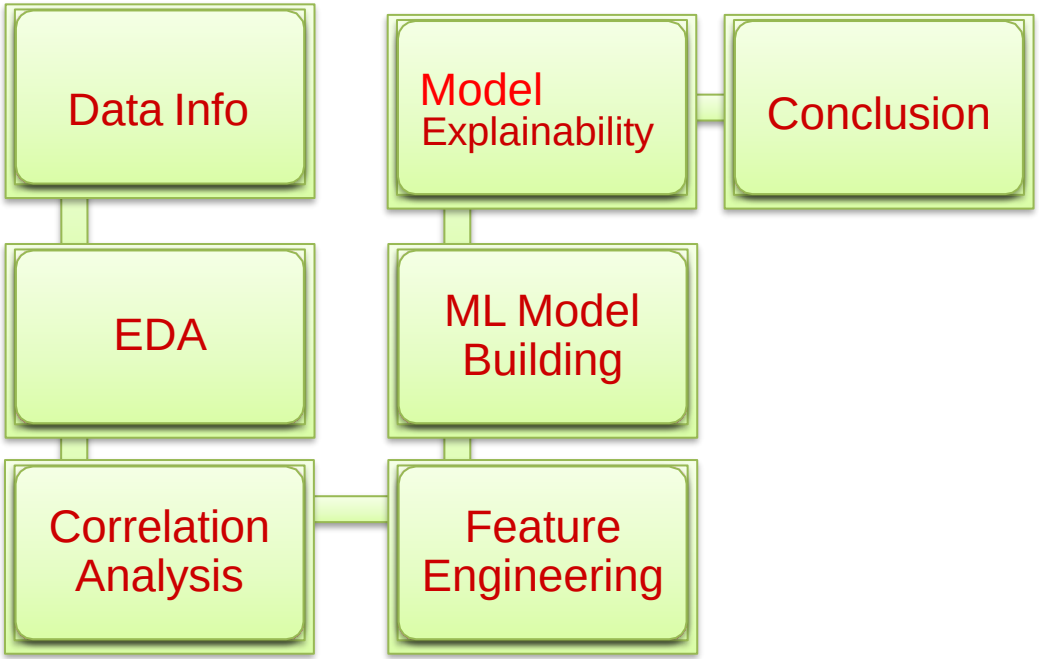
- Help in public transport and increase the mobility of traffic in any city.
- Sustainable development of country
- Meet Zero Carbon emission goal of countries.
- Pollution free, Environment friendly transport medium.
- Easily accessible, affordable to any age person
- The goal is to build a Machine Learning model to predict the bike-sharing demand using the previously stored data.

Problem Statement

Many urban cities introduce rental bikes for good accessibility purposes. It is important to make the rental bike available to the public at the right time as it makes our life easier and faster. 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.

Methodology

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



Data Insights…

- The data set has 14 variables of which Rented Bike Count is a Dependent variable and the rest are independent variables.
- The size of the data is (8760,14) i.e., we have 8760 rows with 14 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.

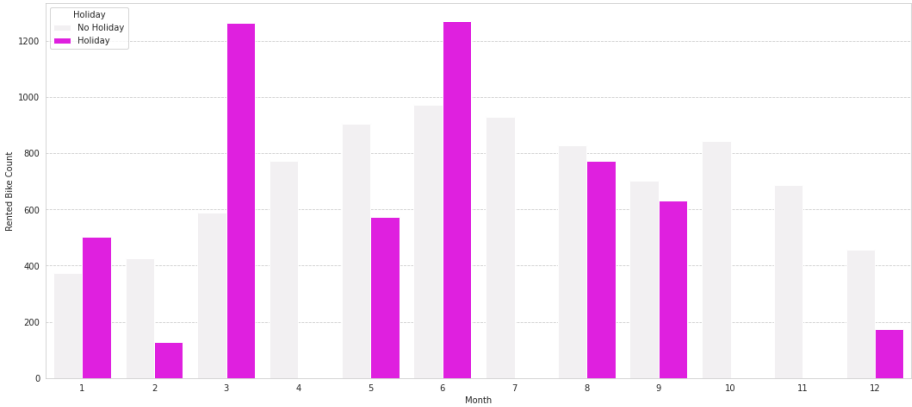
Seasons: Season, in this feature four seasons are present in data is **str** type.

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.

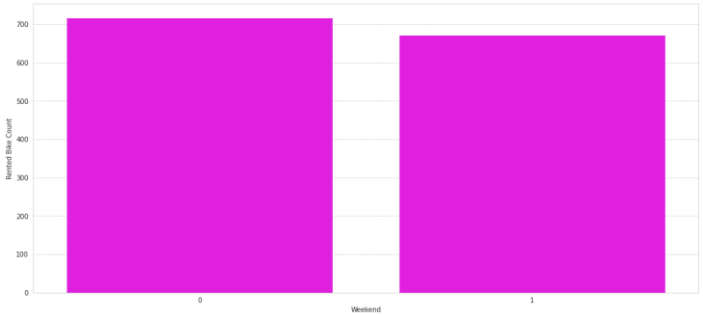
Exploratory Data Analysis

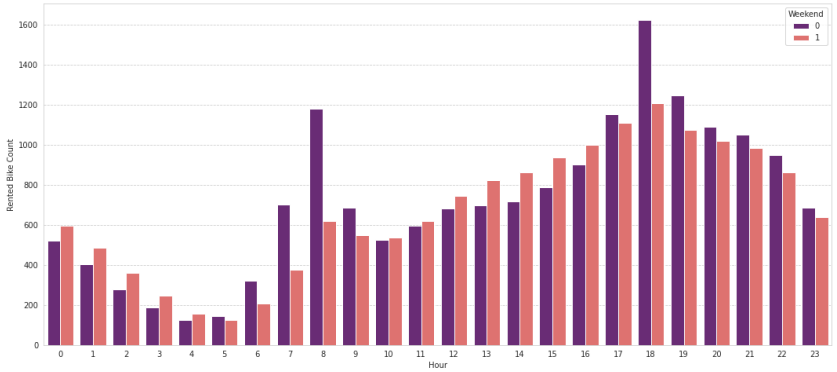
- In June maximum bike rented near 1000. and January, February enjoys less rented bike demand near 400.
- March and June enjoys more bike sharing demand in holiday than non-holidays
- February, December have less bike sharing demand for both in holidays and non-holidays.
- Above graph shows April, October, December months have nearly zero bike sharing demand in holidays.



Exploratory Data Analysis

- Although weekdays enjoys more rented bike demand than weekend but difference is not big.

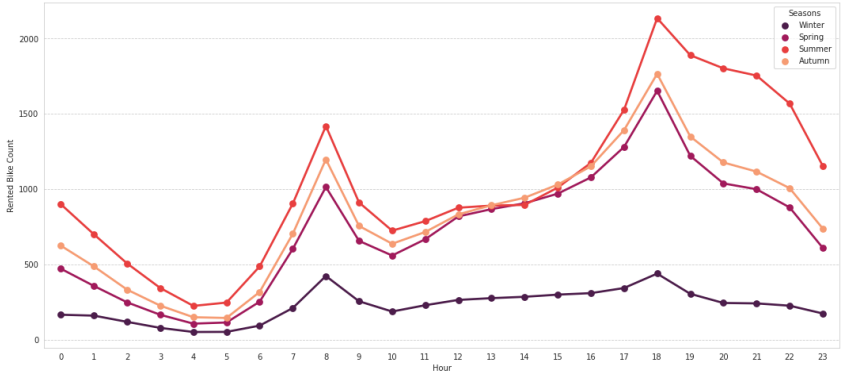


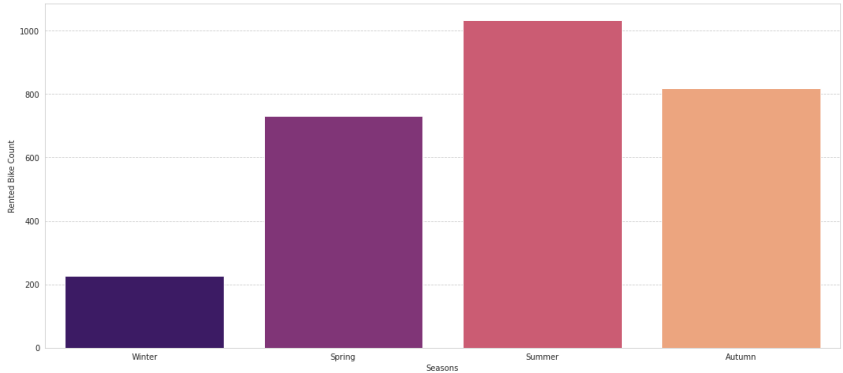


- Hour-18 or 6pm shows maximum rented bike demand in both weekdays(above 1600) and weekend (above 1200).
- Hour 4 & Hour 5 (means 4 and 5 am) shows very less rented bike demand in both weekend and weekdays.
- Hour 8 (8 am) shows good rented bike demand in weekdays(near 1200) but in weekend 8am hour has not so good bike sharing demand(near 600)

Exploratory Data Analysis

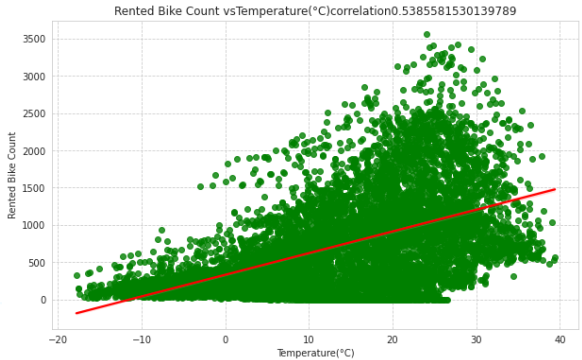
- Summer season has best bike sharing demand(over 1000) followed by autumn, spring and winter(200).
- winter has overall less demand than any other season.
- Hour-18(6pm) and Hour-8(8am) are two best peak time in any season when bike sharing is in high-demand.
- there are not so much difference in demand(near 1000) from Hour-12 to Hour-16(12pm-4pm) among summer, spring, autumn season.
- For every season Hour-4 and 5(4 & 5 am) shows low demand in bike sharing.
- After Hour-10(10am) bike sharing demand is increasing up to Hour-18(6pm) then it is decreasing.

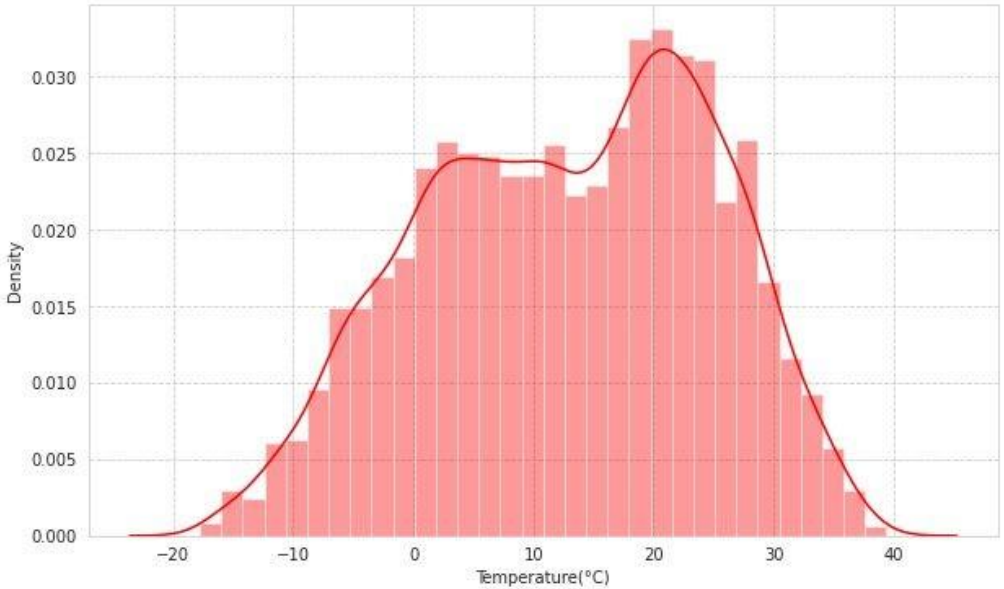




EDA on Numerical Data

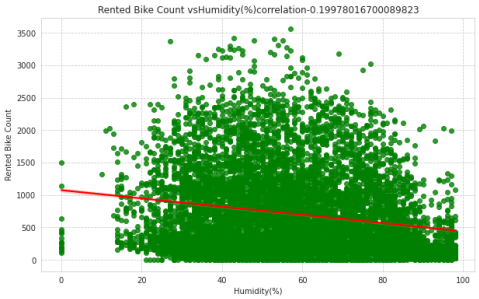
The Temperature of Seoul shows an average range of 0°C to 30 °C. The regression plot for temperature versus rented bike count shows that the Rented Bike Count is linearly proportional to the temperature although it will go to decrease if the temperature rises more than bearable.





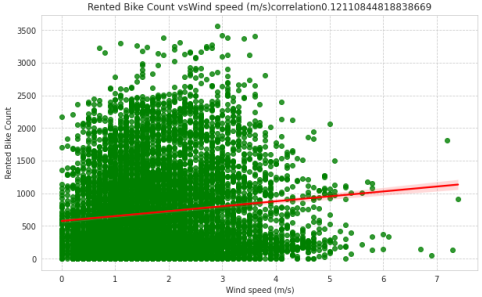
Temperature Based

Regression plots of Humidity, Wind speed & Visibility

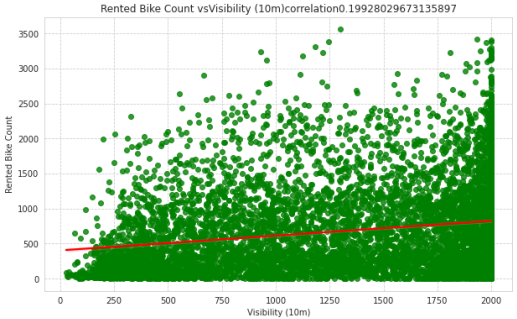


Humidity

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.

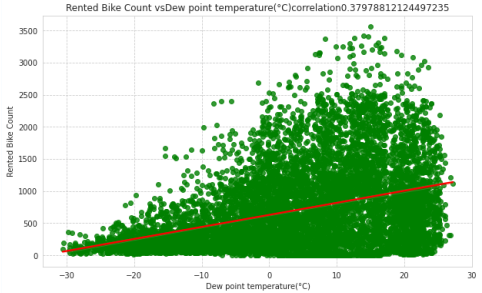


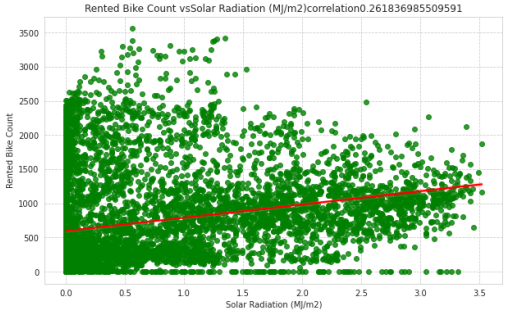
Wind Speed

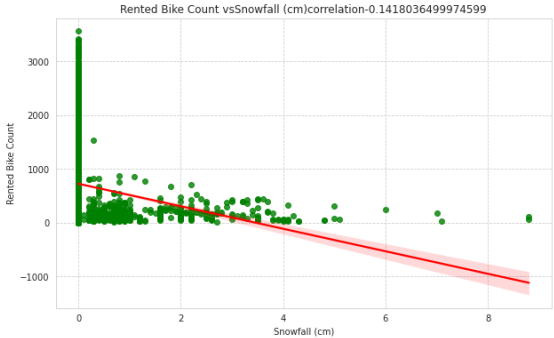


Visibility

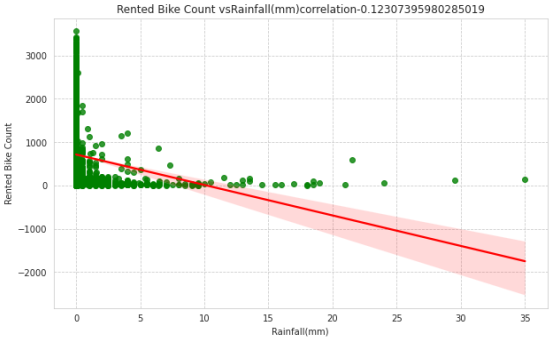
Regression plots of DPT, Solar Radiation, Snowfall & Rainfall







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	Year	Weekend
Rented Bike Count	1.000000	0.410257	0.538558	-0.199780	0.121108	0.199280	0.379788	0.261837	-0.123074	-0.141804	0.070861	0.215162	-0.032259
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	-0.000000
Temperature(°C)	0.538558	0.124114	1.000000	0.159371	-0.036252	0.034794	0.912798	0.353505	0.050282	-0.218405	0.049683	0.377796	-0.012972
Humidity(%)	-0.199780	-0.241644	0.159371	1.000000	-0.336683	-0.543090	0.536894	-0.461919	0.236397	0.108183	0.047798	0.035925	-0.036750
Wind speed (m/s)	0.121108	0.285197	-0.036252	-0.336683	1.000000	0.171507	-0.176486	0.332274	-0.019674	-0.003554	-0.082069	-0.003781	-0.022391
Visibility (10m)	0.199280	0.098753	0.034794	-0.543090	0.171507	1.000000	-0.176630	0.149738	-0.167629	-0.121695	0.077888	0.052381	0.030650
Dew point temperature(°C)	0.379788	0.003054	0.912798	0.536894	-0.176486	-0.176630	1.000000	0.094381	0.125597	-0.150887	0.065101	0.336350	-0.028966
Solar Radiation (MJ/m2)	0.261837	0.145131	0.353505	-0.461919	0.332274	0.149738	0.094381	1.000000	-0.074290	-0.072301	-0.030412	0.128086	0.008271
Rainfall(mm)	-0.123074	0.008715	0.050282	0.236397	-0.019674	-0.167629	0.125597	-0.074290	1.000000	0.008500	-0.022794	0.027522	-0.014280
Snowfall (cm)	-0.141804	-0.021516	-0.218405	0.108183	-0.003554	-0.121695	-0.150887	-0.072301	0.008500	1.000000	0.054758	-0.206418	-0.022557
Month	0.070861	0.000000	0.049683	0.047798	-0.082069	0.077888	0.065101	-0.030412	-0.022794	0.054758	1.000000	-0.295561	0.009174
Year	0.215162	0.000000	0.377796	0.035925	-0.003781	0.052381	0.336350	0.128086	0.027522	-0.206418	-0.295561	1.000000	-0.021590
Weekend	-0.032259	-0.000000	-0.012972	-0.036750	-0.022391	0.030650	-0.028966	0.008271	-0.014280	-0.022557	0.009174	-0.021590	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**

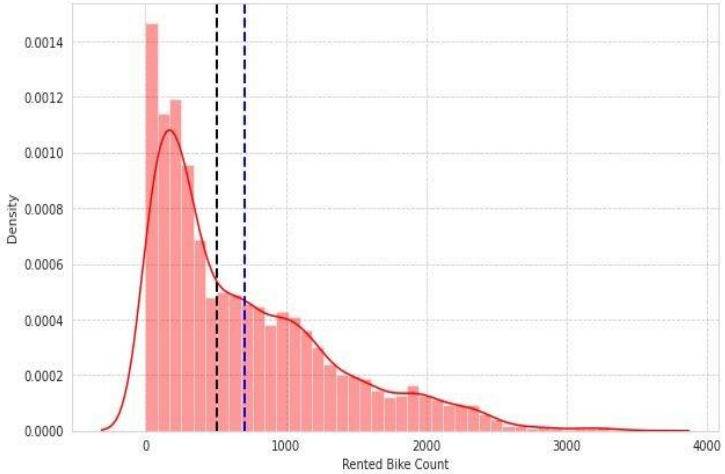
Correlation Analysis (After Treatment)

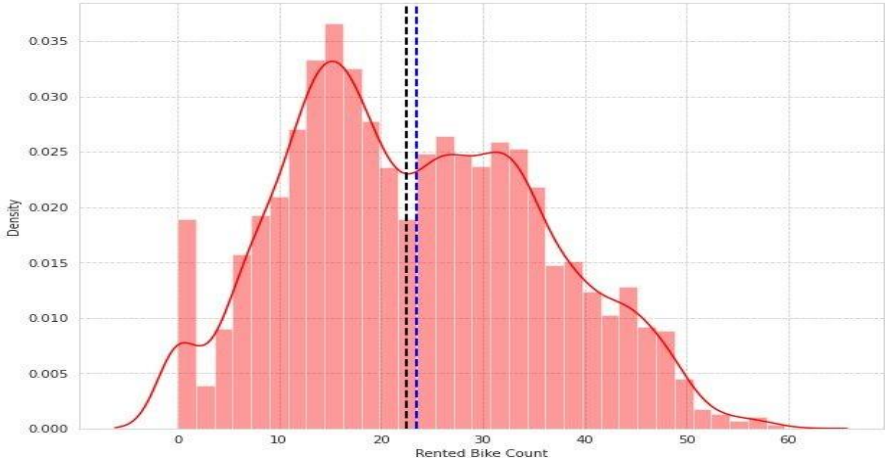
- Correlation plot after dropping the temperature feature show that there are no more highly correlated parameters present in the dataset.
- We can conclude that, there is no multicollinearity present in the dataset

	Rented Bike Count	Hour	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Month	Year	Weekend
Rented Bike Count	1.000000	0.410257	-0.199780	0.121108	0.199280	0.379788	0.261837	-0.123074	-0.141804	0.070861	0.215162	-0.032259
Hour	0.410257	1.000000	-0.241644	0.285197	0.098753	0.003054	0.145131	0.008715	-0.021516	0.000000	0.000000	-0.000000
Humidity(%)	-0.199780	-0.241644	1.000000	-0.336683	-0.543090	0.536894	-0.461919	0.236397	0.108183	0.047798	0.035925	-0.036750
Wind speed (m/s)	0.121108	0.285197	-0.336683	1.000000	0.171507	-0.176486	0.332274	-0.019674	-0.003554	-0.082069	-0.003781	-0.022391
Visibility (10m)	0.199280	0.098753	-0.543090	0.171507	1.000000	-0.176630	0.149738	-0.167629	-0.121695	0.077888	0.052381	0.030650
Dew point temperature(°C)	0.379788	0.003054	0.536894	-0.176486	-0.176630	1.000000	0.094381	0.125597	-0.150887	0.065101	0.336350	-0.028966
Solar Radiation (MJ/m2)	0.261837	0.145131	-0.461919	0.332274	0.149738	0.094381	1.000000	-0.074290	-0.072301	-0.030412	0.128086	0.008271
Rainfall(mm)	-0.123074	0.008715	0.236397	-0.019674	-0.167629	0.125597	-0.074290	1.000000	0.008500	-0.022794	0.027522	-0.014280
Snowfall (cm)	-0.141804	-0.021516	0.108183	-0.003554	-0.121695	-0.150887	-0.072301	0.008500	1.000000	0.054758	-0.206418	-0.022557
Month	0.070861	0.000000	0.047798	-0.082069	0.077888	0.065101	-0.030412	-0.022794	0.054758	1.000000	-0.295561	0.009174
Year	0.215162	0.000000	0.035925	-0.003781	0.052381	0.336350	0.128086	0.027522	-0.206418	-0.295561	1.000000	-0.021590
Weekend	-0.032259	-0.000000	-0.036750	-0.022391	0.030650	-0.028966	0.008271	-0.014280	-0.022557	0.009174	-0.021590	1.000000

Feature Engineering

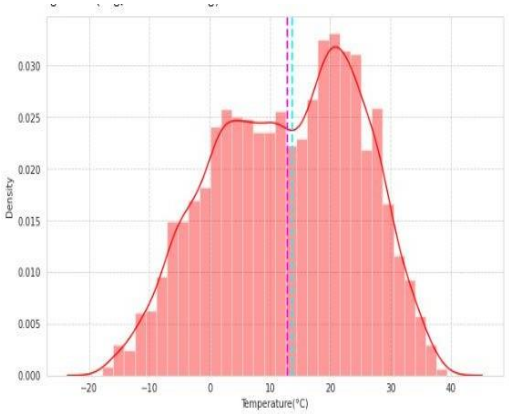
- One Hot Encoding of categorical feature: Hours, Seasons, and Months.
- The date-time, date, day, and temperature & season columns have been dropped from the data set.
- Ordinal Encoding: Holiday and Functioning day columns.
- Normalization has been done on the dependent variable to deal with skewness 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.

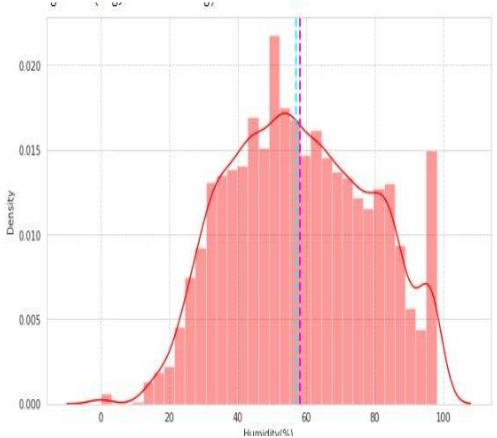


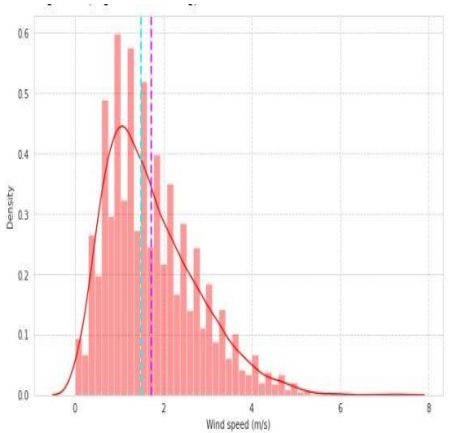


Feature Engineering(Continued..,)

- 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.

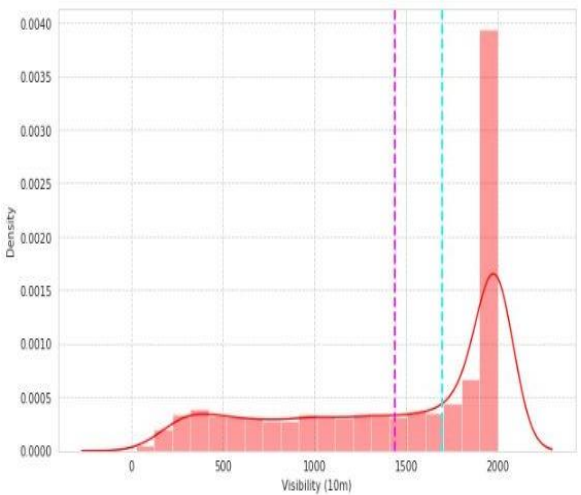


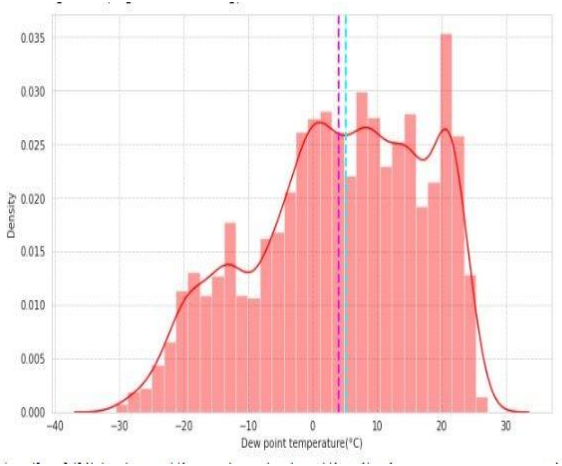


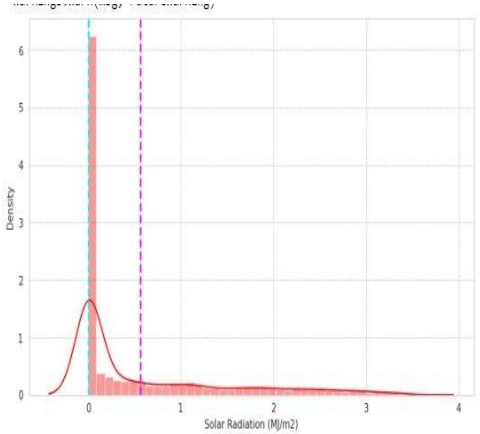


Feature Engineering(Continued…)

- 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

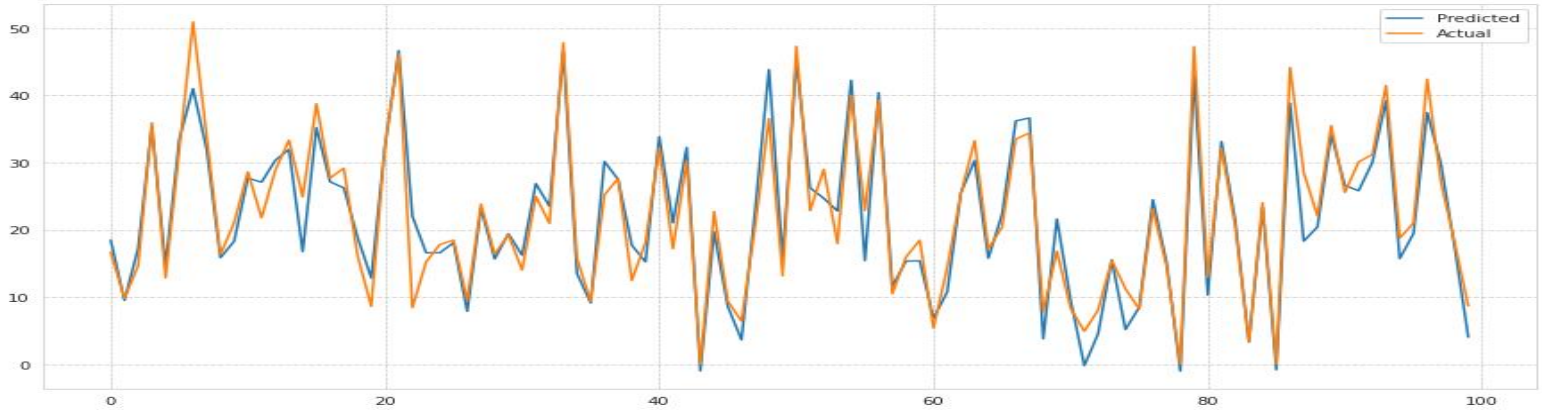
- Model accuracy is moderate for training as well as test data. Therefore we can conclude that no overfitting.
- Since there is no overfitting, we did not go ahead with Regularized linear Regression
- We plotted line graph of actual vs predicted Rented bike count.

```
Trainning Errors
MSE1: 37.352161123013445
MAE: 4.644319757141869
R2: 0.761
Testing Errors
MSE1_test: 36.21780331779815
MAE_test: 4.564676954443265
R2_test: 0.76
```

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

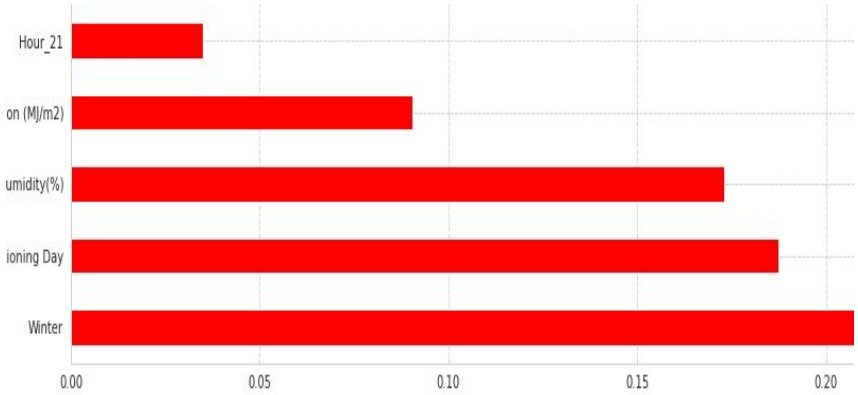
```
Training Errors
MSE: 14.1509673185268
MAE: 2.607308092188883
R2: 0.91


Testing Errors
MSE: 16.795328902139108
MAE: 2.87251476143409
R2: 0.89
```

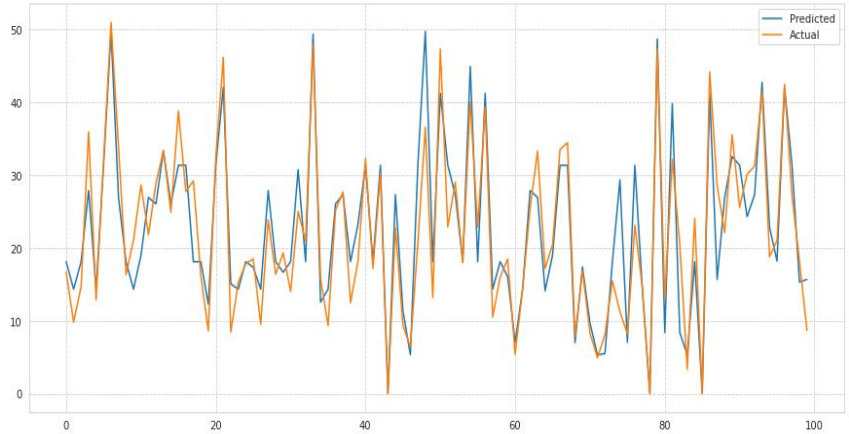


Decision Tree Regressor

- Parameters: max depth = 10, max-leaf nodes = 200
- 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_18** is showing least feature importance while **Winter** season is showing highest feature importance in model prediction.

```
Training Errors
MSE: 24.715448527413255
MAE: 3.6151056157960966
R2: 0.842


Testing Errors
MSE: 30.48214244817496
MAE: 4.005385023326545
R2: 0.798
```

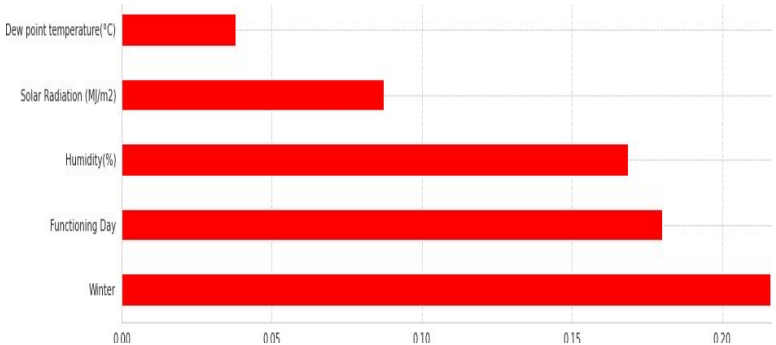


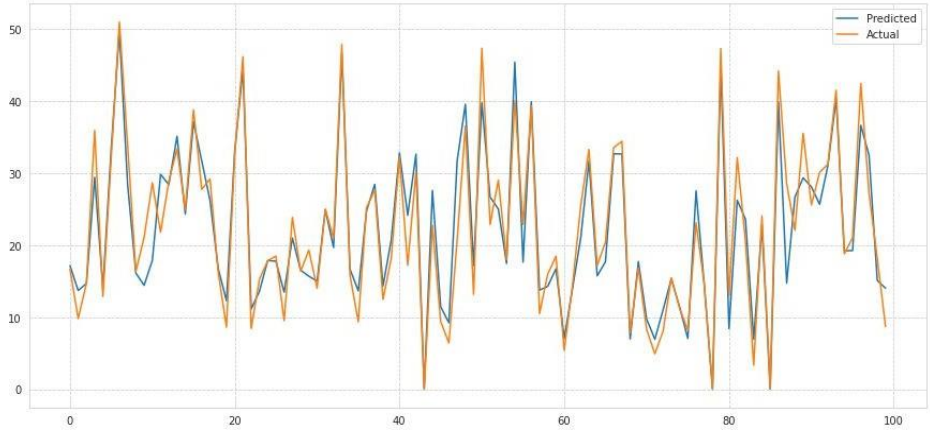


Random Forest Regressor

- Parameters: n_estimators = 180, max depth = 12, max-leaf nodes = 84
- 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 lowest feature importance while Winter season is showing highest feature importance in model prediction.

```
Trainning Errors
MSE: 19.547610124997167
MAE: 3.3107815356966492
r2: 0.875


Testing Errors
MSE: 22.12148543100839
MAE: 3.424461188207054
R2: 0.854
```

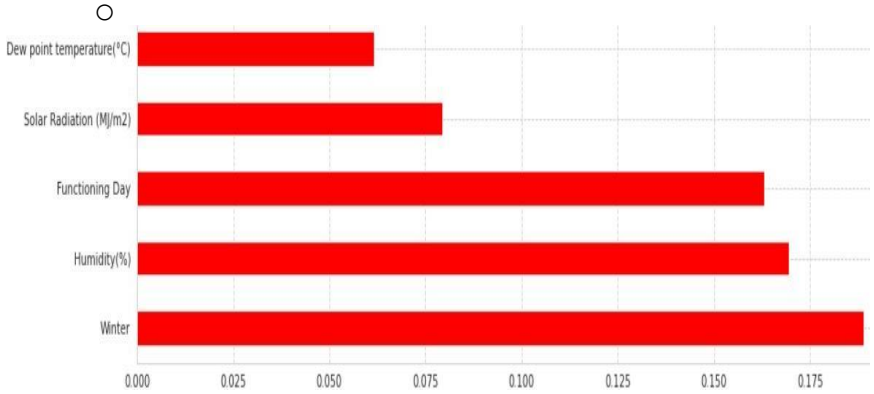


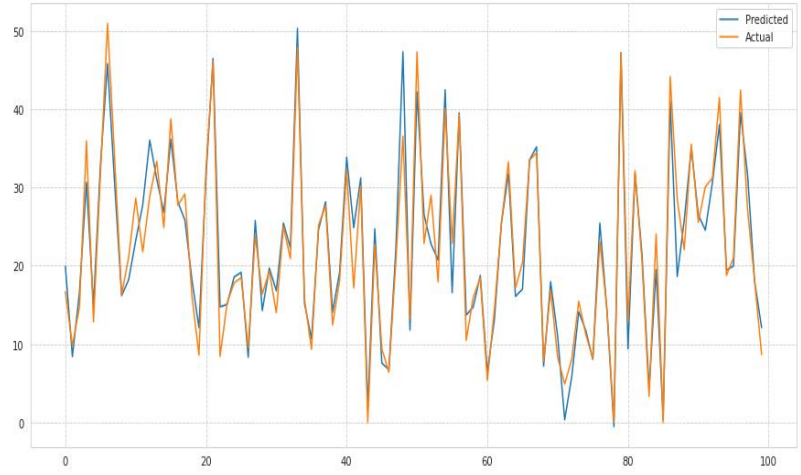


Gradient Boost with Hyper Parameter Tuning

- 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=80
- Here we can see Hour_19 is showing lowest feature importance while Winter season is showing highest feature importance in model prediction.

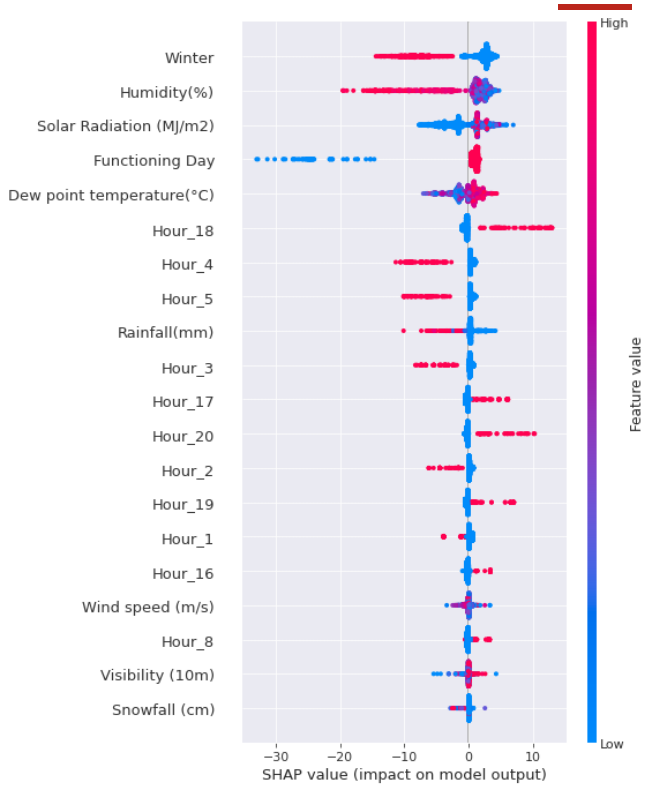
```
Training Errors
MSE: 8.627750557153538
MAE: 2.015903026054293
R2: 0.945

Testing Errors
MSE: 12.733720673058976
MAE: 2.4930521620341244
R2: 0.916
```

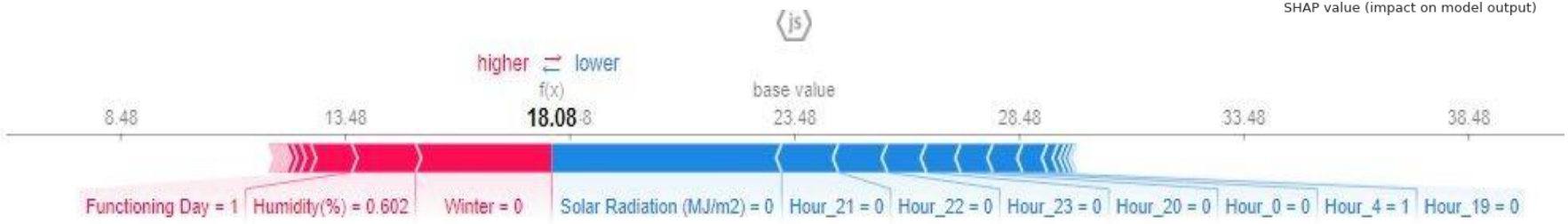




Model Explainability

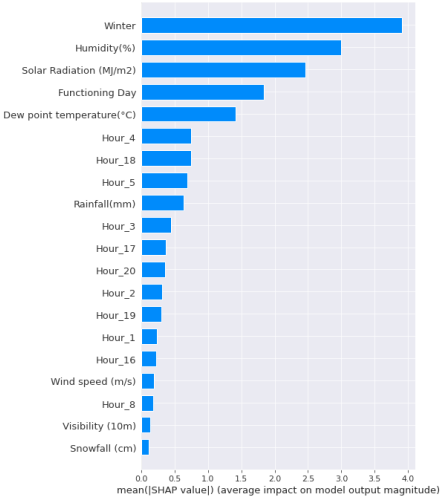
- We have used SHAP JS visualisation to explain feature values and their importance in each models. Here we can see negative feature or blue color block pushes the prediction toward left over base value. Also we can see highest Solar radiation value which causing prediction negative while winter which 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.

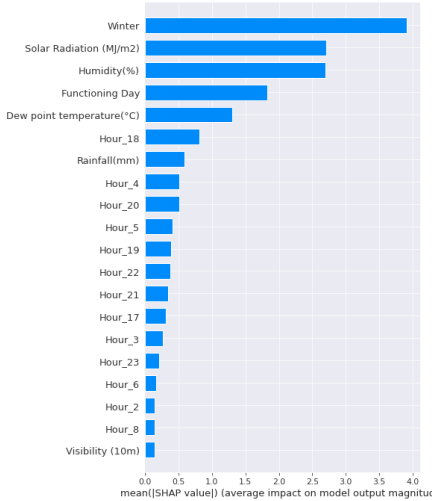


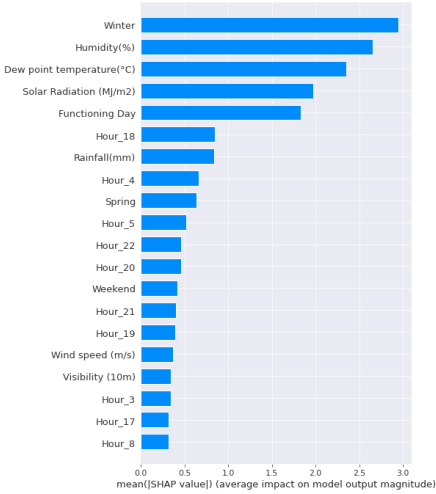


Model Explainability(Continued..)

- 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 **Snowfall** has the Lowest feature_value
- In Random Forest Model we can see **Winter** has the highest feature value while **Visibility** has the Lowest shap value.
- In bar graph we can see **Winter** has the highest feature value while **Hour_8** has the Lowest shap value.
- We can conclude that Hour_8,Visibility and Wind Speed is not contributing in Decision Tree, Random Forest and Gradient Boost in model prediction.







Conclusion

- 1) In June maximum bike rented near 1000. and january, february enjoys less rented bike demand near 400. March and June enjoys more bike sharing demand in holiday than non-holidays February,December have less bike sharing demand for both in holidays and non-holidays. April,october,december months have nearly zero bike sharing demand in holidays.

- 2) weekdays have more rented bike demand than weekend

- 3)In weekdays 6 pm and 8 am but in weekend only 6 pm are peak time of bike sharing demand. 4 and 5 am has lowest bike sharing demand in both weekeend and weekdays

- 4) Summer season enjoys overall best and least bike sharing demand and winter has overall less demand than any other season. There are very less bike sharing demand in morning 4 and 5 am

- 5) From the regression plots we can conclude that the columns

- 'Rainfall', 'Snowfall', 'Humidity' these features are negatively related with the dependent variable 'Rented Bike Count'. This means when Rainfal,snowfal,humidity is higher bike sharing demand is lower.
- 'Temperature', 'Wind_speed','Visibility', 'Dew_point_temperature', 'Solar_Radiation' are positively correlated with the dependent variable 'Rented Bike Demand'. This means if 'Temperature', 'Wind_speed','Visibility', 'Dew_point_temperature', 'Solar_Radiation' are higher or lower then bike sharing demand maybe higher or lower respectively.

Conclusion

- 6) After applying linear regression model, we got r2 score of 0.761 for training dataset and 0.76 for test dataset which defines that model is optimally fit for training and test data i.e no overfitting
- 7) Therefore, for even better fit, we applied polynomial regression model with degree = 2, we got R2 score of 0.91 for training data and 0.89 for test data
- 8) 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 0.842 for training data and 0.798 for test data which is less than polynomial regression.
- 10) To get better accuracy on tree based model, we applied Random forest with n_estimator as 180 and with maximum depth as 12, with that we got R2 score of 0.875 for training data and 0.854 for test data.
- 11) Finally, we applied Gradient boost with parameters selected after grid search which resulted in highest R2 score of 0.945 for training data and 0.916 for test data with very less mean squared error of 8.6 and 12.4 in training as well as in test data.
- Therefore we can say that it gives us optimal result in term of test dataset. It is best for final prediction
- 12) Lastly, In bar graph we can see Winter has the highest feature value.We can conclude that Hour_8,Visibility and Wind Speed is not contributing in Decision Tree,Random Forest and Gradient Boost in model prediction

Thank you