Capstone Project-2

Bike Sharing Demand Prediction

By- Parth Sharma

•A General Walk Through:

- About Bike Sharing.
- Determining the problem statement.
- Steps Involved.
- Dataset summary.
- Exploratory Data Analysis.
- Models and Model Selection.
- Comparison
- Observations
- Conclusion.

- # ABOUT BIKE SHARING

  - A **Bike-sharing system,** is a shared transport service in which bicycles are made available for shared use to individuals on a short-term basis for a price.
  - Many bike share systems allow people to borrow a bike from a "dock" and return it at another dock belonging to the same system.
  - The user enters payment information, and the computer unlocks a bike. The user returns the bike by placing it in the dock, which locks it in place.

depositphotos

- ## DETERMINATION OF 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 bike count required at each hour for the stable supply of rental bikes.

- **Data Processing Steps**

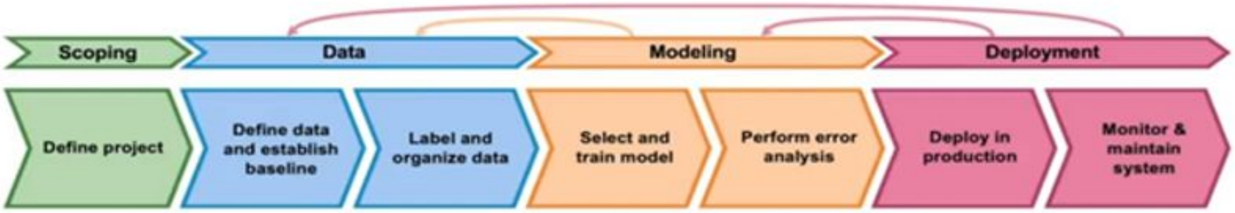
  - **Data Understanding**: This step involves the process of understanding the data set available, getting to know about the dependent variable and independent variables, their data type, have an overview of the statistical features of the numerical features in our dataset.
  - **Data Processing:** Making the data fit for future analysis is the main objective of this step. During this stage, gone through each input variable and studied the type of features make them suitable for the analysis and tried to further break down the features into segments to make our dataset more informative like we did to date column. Also, looked forward to the outliers, null values to keep our data free from any anomaly. Checking for the duplicate values as it can mar the accuracy of our dataset.

  - **EDA:** EDA or Exploratory Data Analysis is the critical process of performing the initial investigation on the data. So, through this we have observed certain trends and dependencies and also drawn certain conclusions from the dataset that will be useful for further processing.

Continued…

- ## Data Processing Steps

  - **Model Fitting And Metric Checking**: During this step we have fit our data to different models and tried to fit our data to different models and predicted the output and then calculated the metrics to check which ever model fits the best for further predictions.



- ## DATASET SUMMARY:

The dataset contains weather information like Temperature, Humidity, Wind-speed, Visibility, Dew- point, Solar radiation, Snowfall, Rainfall, the number of bikes rented per hour and date information.

Attribute Information:

- Date : The date of the day, during 365 days from 01/12/2017 to 30/11/2018, formatting in DD/MM/YYYY, type : string, we need to convert into date time format in order to make is informative.
- Rented Bike Count : Number of rented bikes per hour which our dependent variable and we need to predict that through our model, type : integer. This feature is our target variable.
- Hour: This column represents hour of the day corresponding to which we have the count of bike been rented, starting from 0-23, type : integer, we need to convert it into category data type.
- Temperature(°C): Temperature in Celsius, represents the temperature that happens to at a particular hour on a particular day.

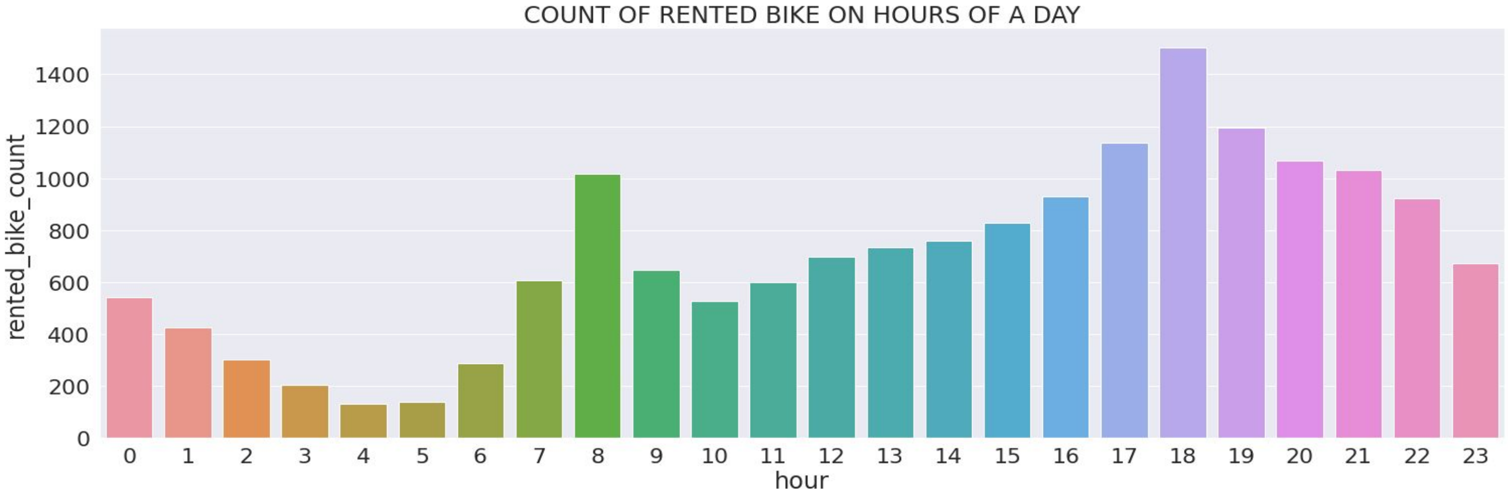
Continued…

- Humidity(%): Humidity in the air in air i.e. the moisture content present in the air, this can affect directly the count for bikes being rented as it affects the weather condition outside.
- Wind speed (m/s) : Speed of the wind in m/s, type : Float, this can affect directly the count for bikes being rented as it affects the weather condition outside.
- Visibility (m): Visibility in m, type : integer, This is measure that represents the degree of clearness or the greatest distance at which prominent objects can be identified with the naked eye.
- Dew point temperature(°C): Temperature at the beginning of the day, type : Float
- Solar Radiation (MJ/m2): Sun contribution, type : Float, this can be regarded as the radiation emitted by the sun, it can prominently affect our target variable.
- Rainfall(mm): Amount of raining in mm, type : Float.
- Snowfall (cm): Amount of snowing in cm, type : Float.
- Seasons: Season of the year, type : string, there are only 4 season's in data.
- Holiday: If the day is holiday period or not, type: string.
- Functioning Day: If the day is a Functioning Day or not, type : string.

EXPLORATORY DATA ANALYSIS

• COUNT OF RENTED BIKE ON HOURS OF A DAY

- In general people used rented bikes during their commuting hours i.e. from 7am to 9am in morning and 5pm to 7pm in the evening.

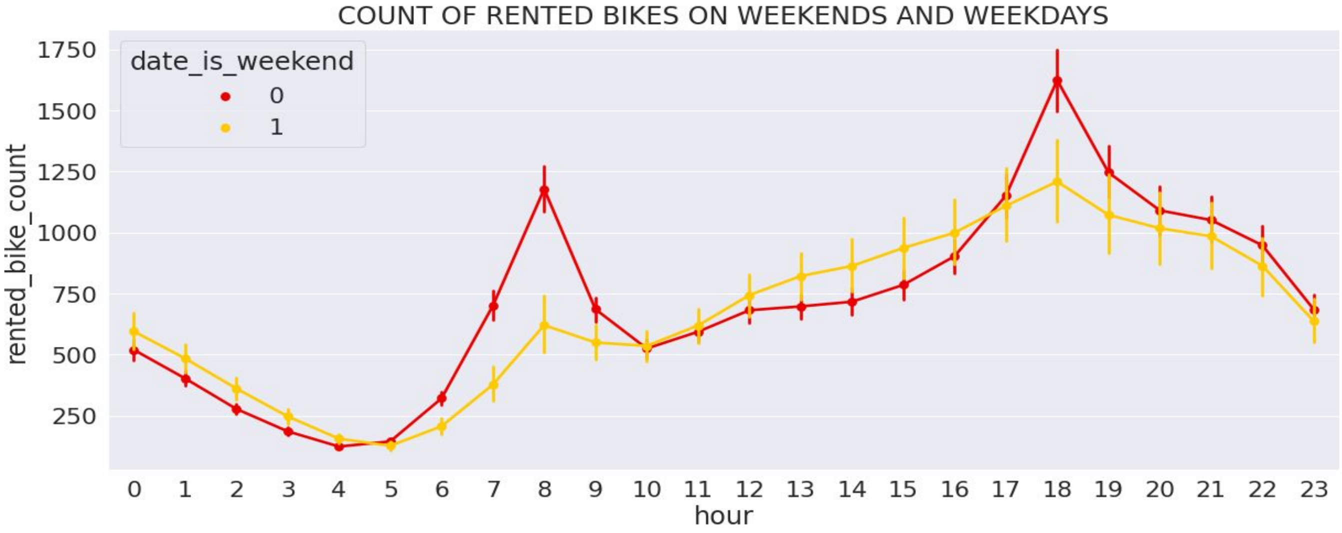


• RENTED BIKES DEMAND ON WEEKENDS AND WEEKDAYS

- The week days which represent in red colour show that the demand of the bike higher because of the office.
- Peak time 7 am to 9 am and 5 pm to 7 pm and it show that the demand of rented bikes are very low early morning but when the evening start from 4 pm to 8 pm the demand slightly high but shows the decreasing trend and that is due obvious reasons.

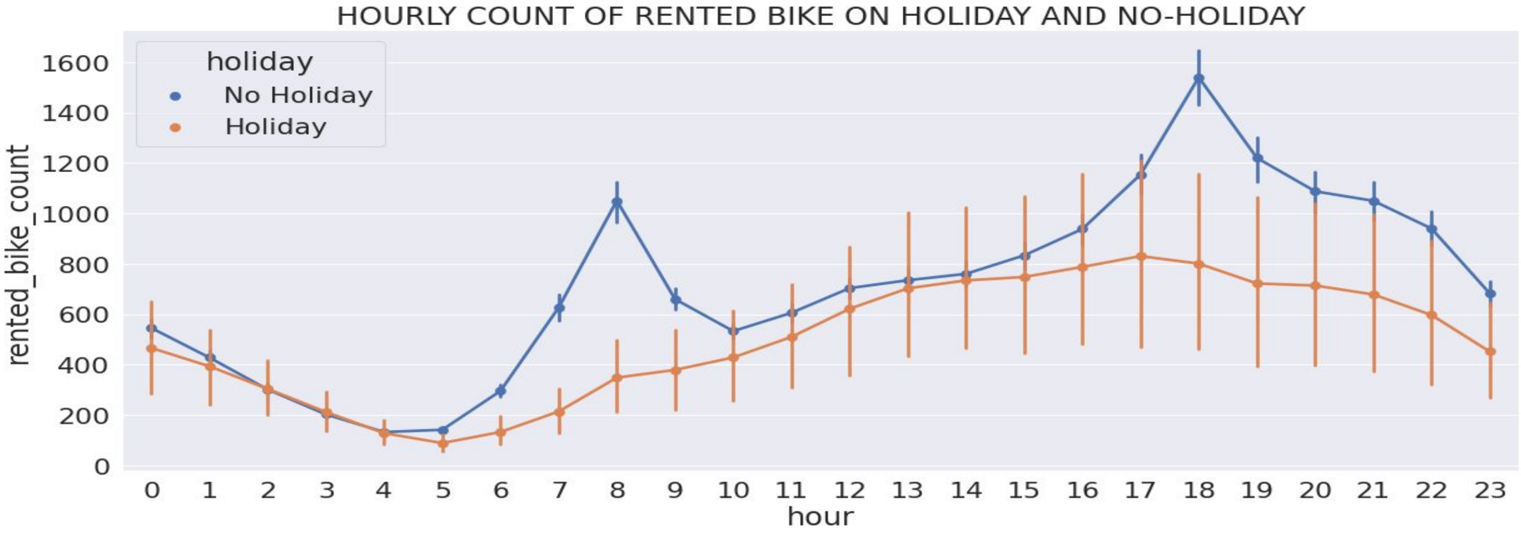
0-Weekday

1-Weekend



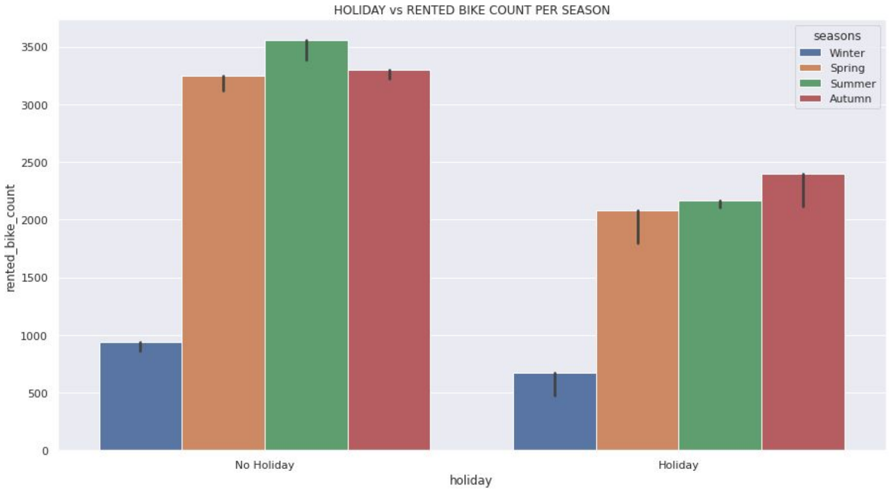
- **HOURLY COUNT OF RENTED BIKE ON HOLIDAY AND NO-HOLIDAY**

- The non holiday days that is represented in blue colour show that the demand of the bike higher on the Non holiday days, it can be assumed that the bike be used by the people commuting to their offices and other.
- Whereas on the holidays the rented bikes count is not that low but lesser in comparison to the non-holidays.



HOLIDAY vs. RENTED BIKE COUNT ACROSS ALL THE SEASONS

- Here we can observe that irrespective of the season the rented bikes are more in demand on the non-holiday.
- Also on a Noholiday the demand of bikes is high in the Summer season whereas Autumn Season tops the list on a Holiday.

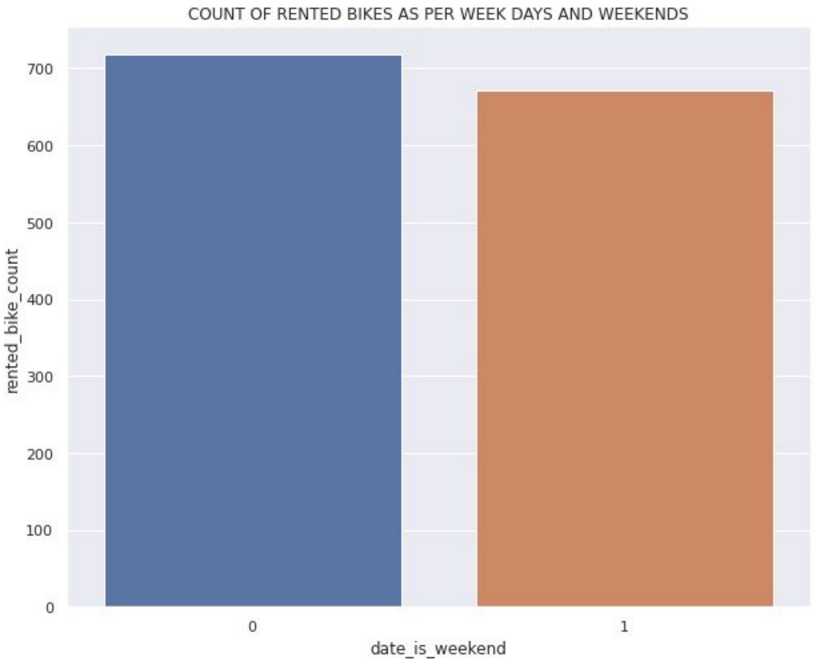


COMPARISON OF BIKES RENTED ON WEEKDAYS AND WEEKENDS

- Here we can observe that the average number of the bikes rented on weekdays is comparatively high.

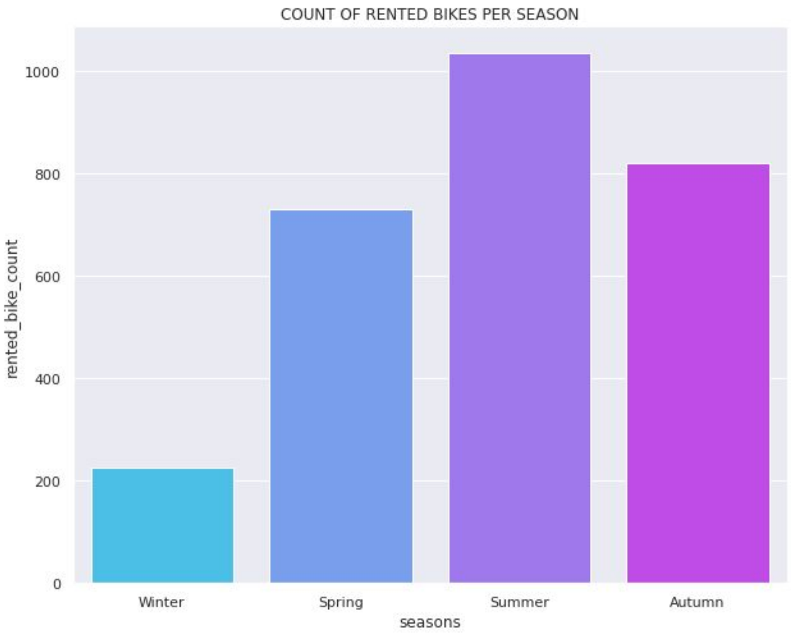
0-Weekday

1-Weekend

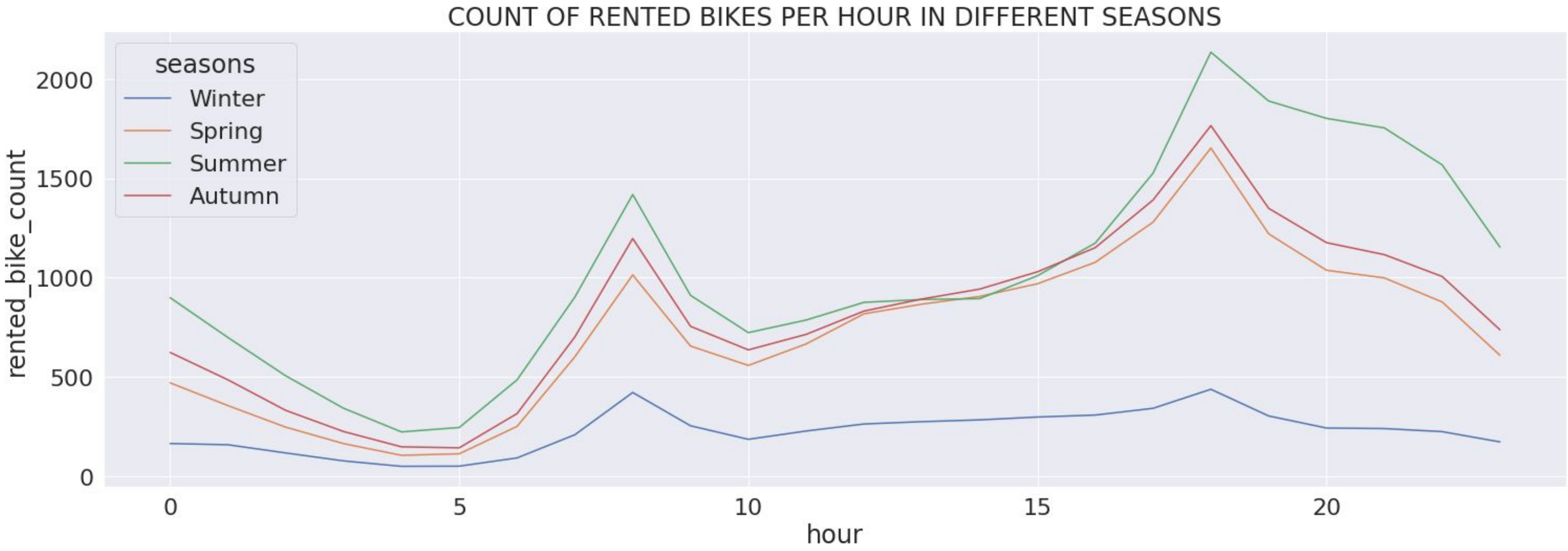


•AVERAGE COUNT OF RENTED BIKE PER SEASON

- Here we an observation that maximum numbers of bikes were rented in the summer season in comparison with the other seasons, it is found that the minimum numbers of bikes were rented in the winter season.



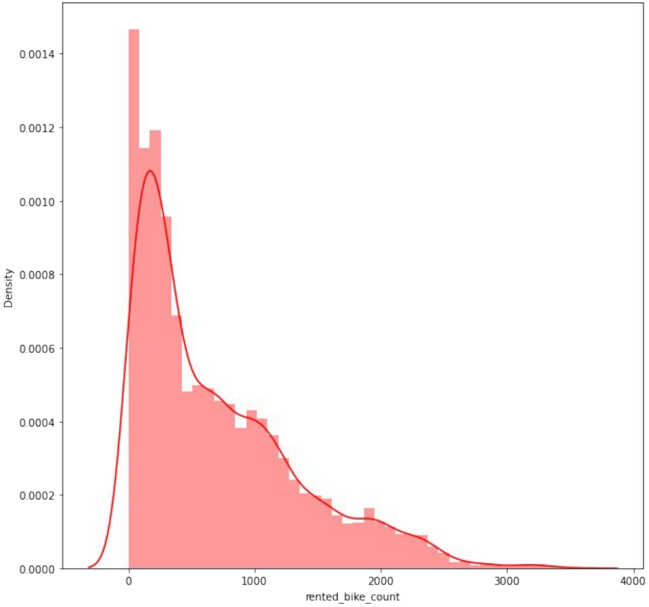
• RENTED BIKES COUNT PER HOUR IN DIFFERENT SEASONS



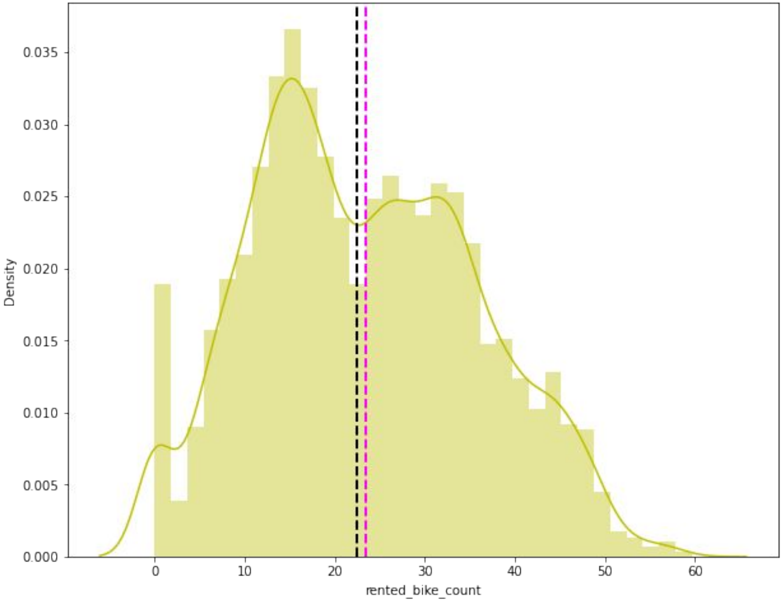
•DISTRIBUTION OF OUR TARGET VARIABLE-RENTED BIKE COUNT

- Plot shows the distribution of the rented bikes count and it is observed to be skewed toward right therefore we need to apply transformation to normalise the values.

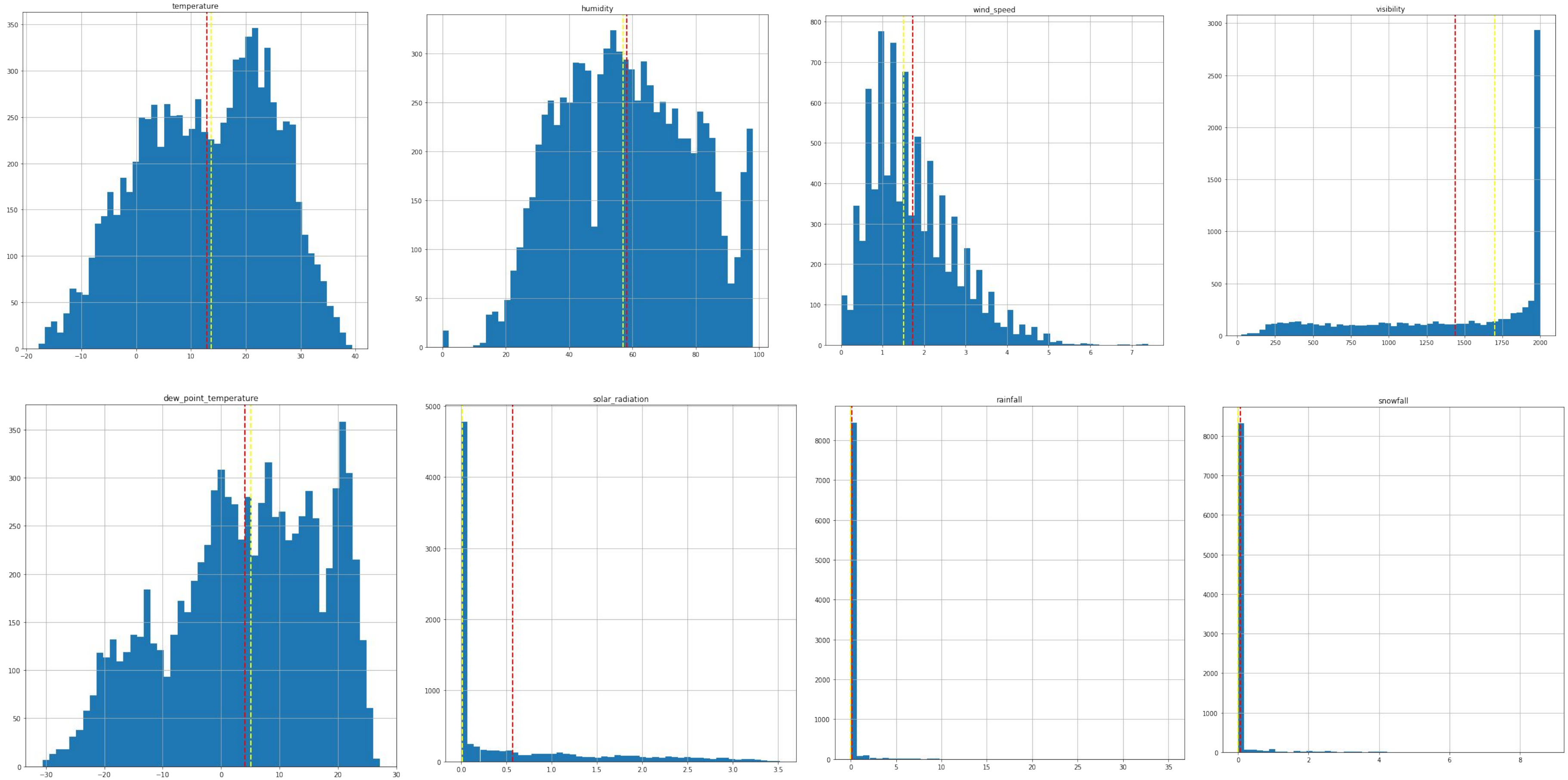
BEFORE TRANSFORMATION



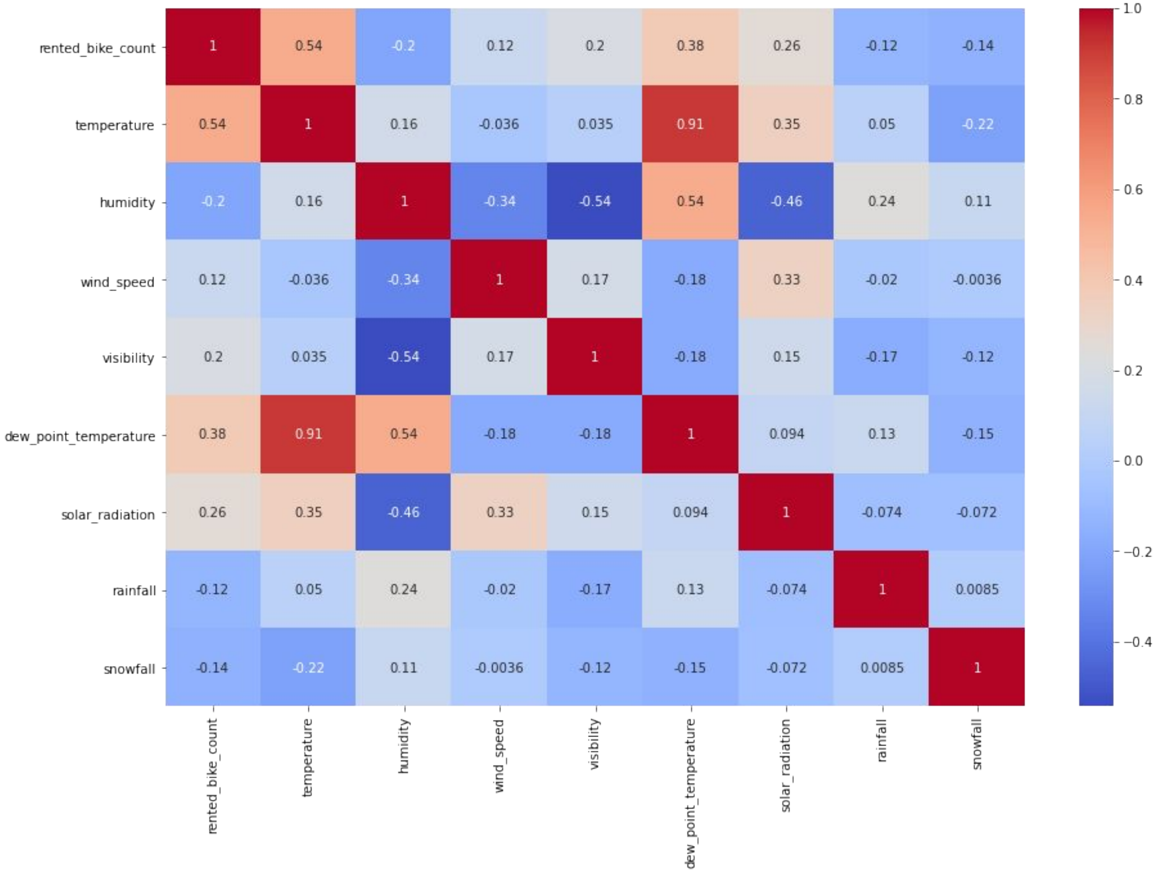
AFTER TRANSFORMATION



- # DISTRIBUTION OF ALL THE INDEPENDENT VARIABLES



•CORRELATION MATRIX



	rented_bike_count	temperature	humidity	wind_speed	visibility	dew_point_temperature	solar_radiation	rainfall	snowfall
rented_bike_count	1	0.54	-0.2	0.12	0.2	0.38	0.26	-0.12	-0.14
temperature	0.54	1	0.16	-0.036	0.035	0.91	0.35	0.05	-0.22
humidity	-0.2	0.16	1	-0.34	-0.54	0.54	-0.46	0.24	0.11
wind_speed	0.12	-0.036	-0.34	1	0.17	-0.18	0.33	-0.02	-0.0036
visibility	0.2	0.035	-0.54	0.17	1	-0.18	0.15	-0.17	-0.12
dew_point_temperature	0.38	0.91	0.54	-0.18	-0.18	1	0.094	0.13	-0.15
solar_radiation	0.26	0.35	-0.46	0.33	0.15	0.094	1	-0.074	-0.072
rainfall	-0.12	0.05	0.24	-0.02	-0.17	0.13	-0.074	1	0.0085
snowfall	-0.14	-0.22	0.11	-0.0036	-0.12	-0.15	-0.072	0.0085	1

•OBSERVATIONS FROM HEAT MAP AND CHECK FOR MULTICOLLINEARITY

On the target variable line the most positively correlated variables are:

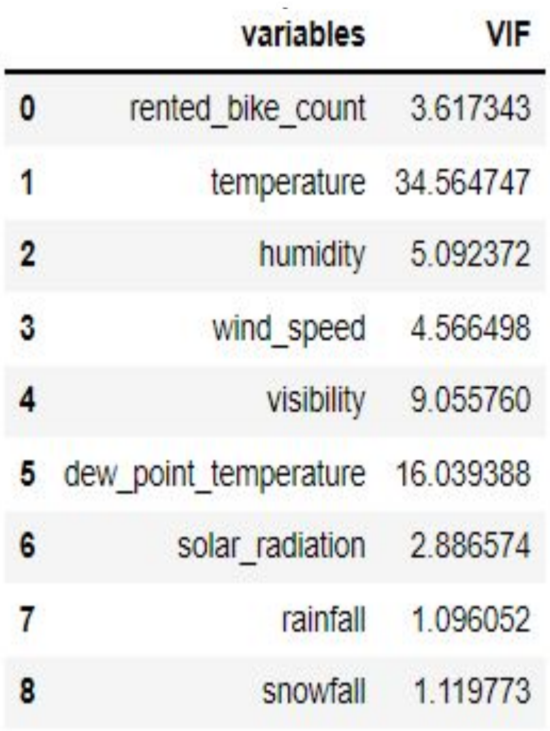
- Temperature
- Wind Speed
- Dew Point Temperature
- Solar Radiation
- Visibility

Most negatively correlated variables are:

- Humidity
- Rainfall
- Snowfall

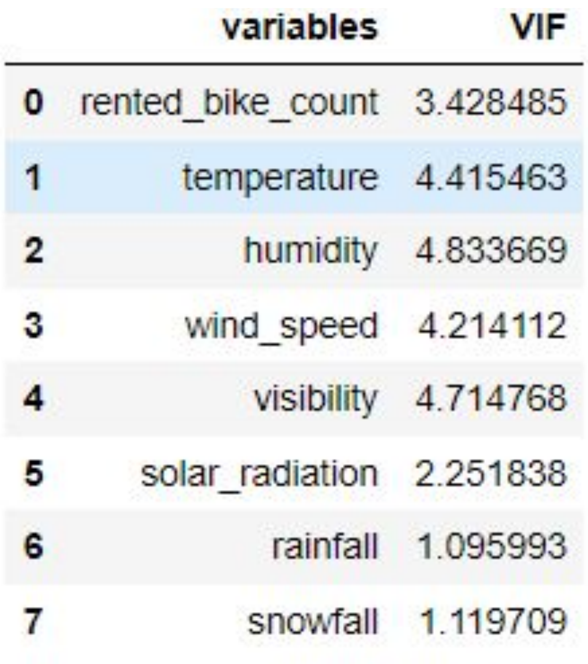
$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

BEFORE TREATING MULICOLLINEARITY



	variables	VIF
0	rented_bike_count	3.617343
1	temperature	34.564747
2	humidity	5.092372
3	wind_speed	4.566498
4	visibility	9.055760
5	dew_point_temperature	16.039388
6	solar_radiation	2.886574
7	rainfall	1.096052
8	snowfall	1.119773

AFTER TREATING MULICOLLINEARITY



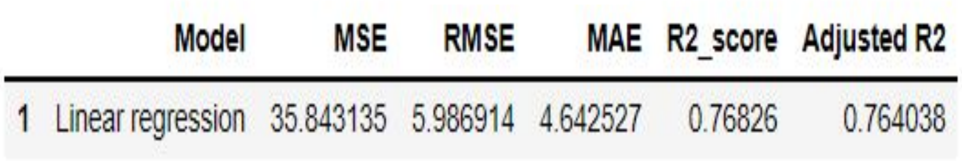
	variables	VIF
0	rented_bike_count	3.428485
1	temperature	4.415463
2	humidity	4.833669
3	wind_speed	4.214112
4	visibility	4.714768
5	solar_radiation	2.251838
6	rainfall	1.095993
7	snowfall	1.119709

MODELS AND MODEL SELECTION

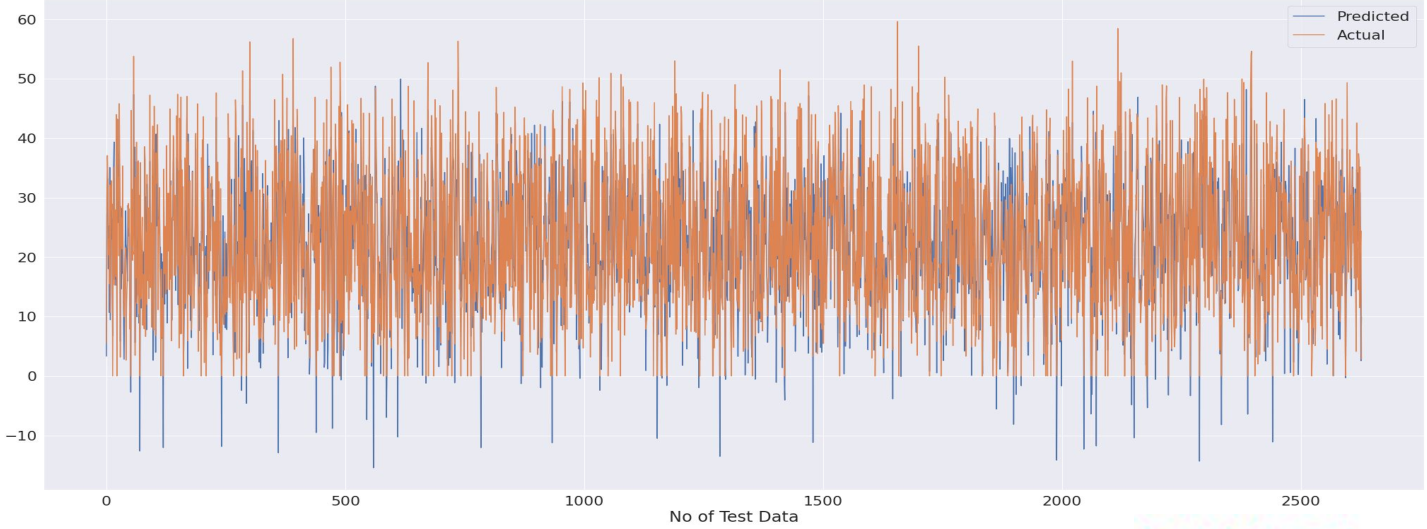
•MODELS AND MODEL SELECTION

- **Linear Regression:**
- Data frame shows the metrics from calculated on the test part of the datasets.
- Since R2 is represents the goodness of fit of a model. Here we have R2 as 0.768.
- Therefore, we can say that our model gave us good results.

METRCIS



	Model	MSE	RMSE	MAE	R2_score	Adjusted R2
1	Linear regression	35.843135	5.986914	4.642527	0.76826	0.764038

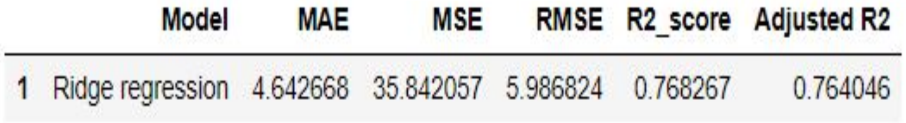


Continued…

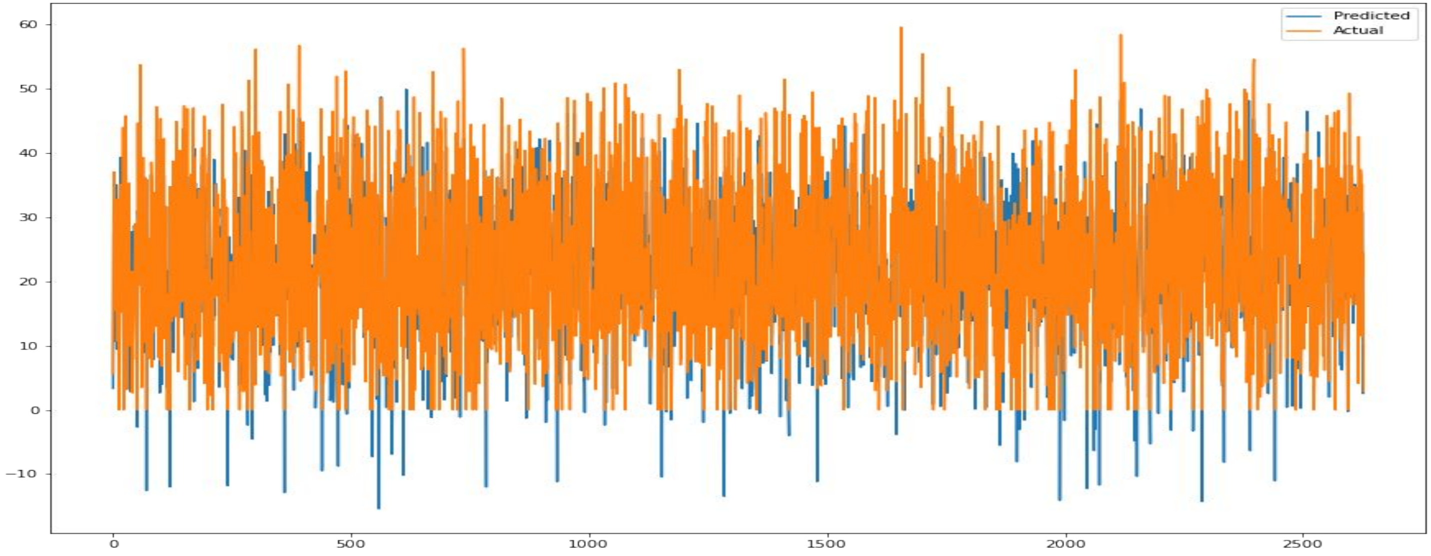
•MODELS AND MODEL SELECTION

- **Ridge Regression:**
- Data frame shows the metrics from calculated on the test part of the datasets.
- Since R2 is represents the goodness of fit of a model. Here we have R2 as 0.768.
- Therefore, we can say that our model gave us good results.

METRCIS



	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
1	Ridge regression	4.642668	35.842057	5.986824	0.768267	0.764046

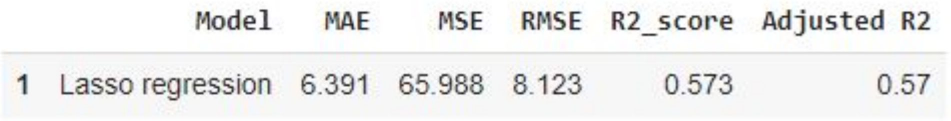


Continued…

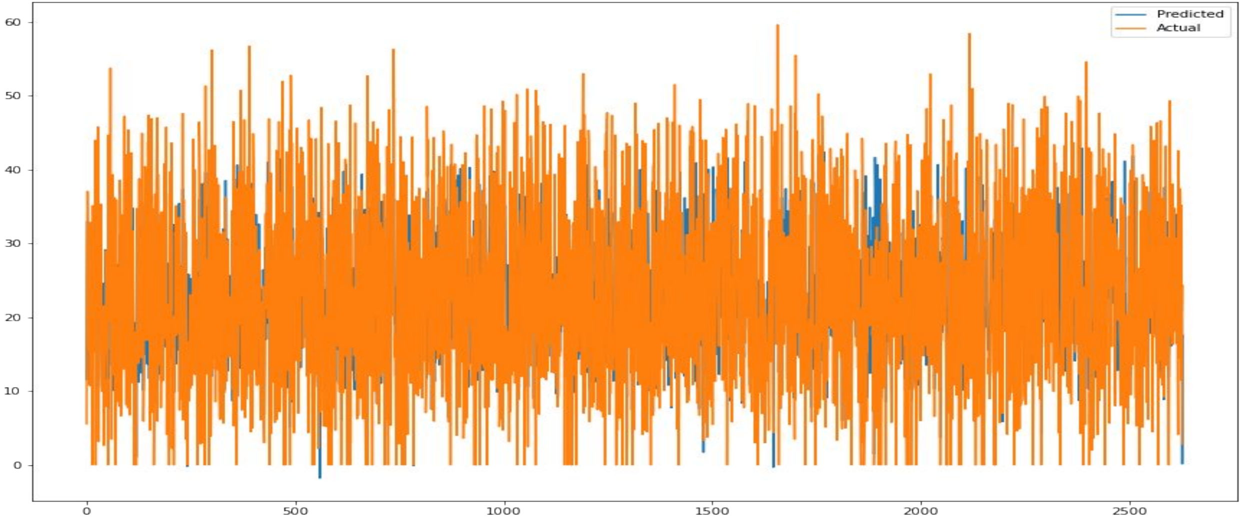
•MODELS AND MODEL SELECTION

- **Lasso Regression:**
- Data frame shows the metrics from calculated on the test part of the datasets.
- Since R2 is represents the goodness of fit of a model. Here we have R2 as 0.573.
- Therefore, we can say that this model does not give good results in comparison of the models deployed earlier.

METRCIS



	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
1	Lasso regression	6.391	65.988	8.123	0.573	0.57

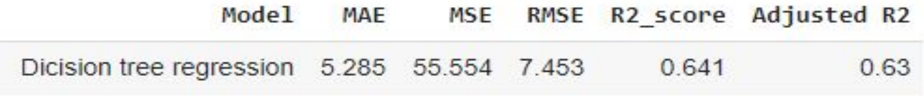


Continued...

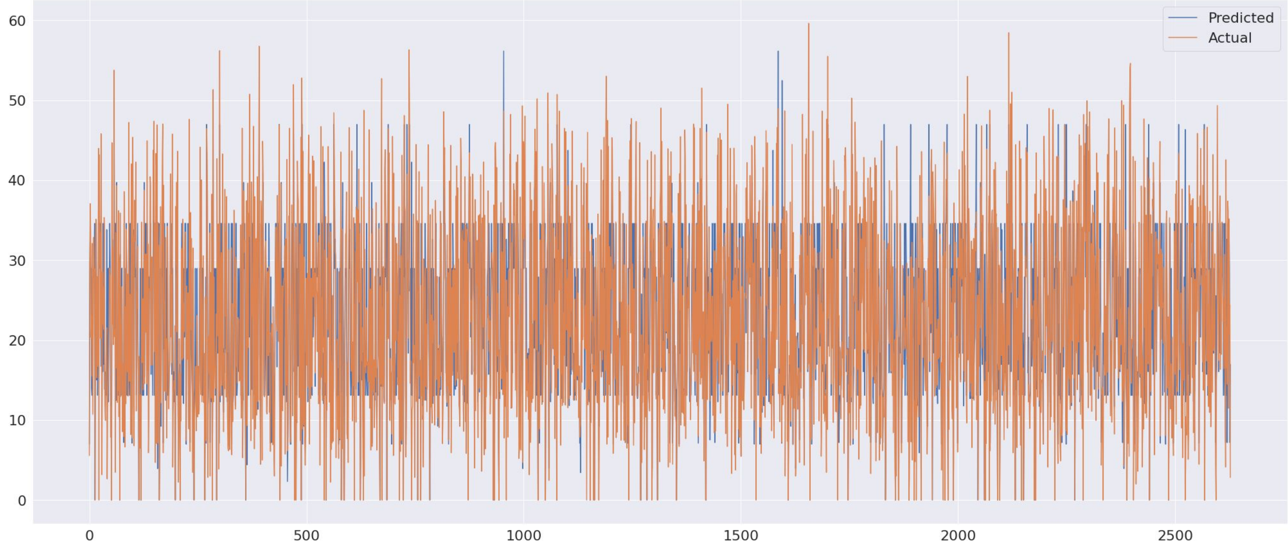
•MODELS AND MODEL SELECTION

- **DECISION TREE:**
- Data frame shows the metrics from calculated on the test part of the datasets.
- Since R2 is represents the goodness of fit of a model. Here we have R2 as 0.641.
- Therefore, we can say that our model gave us comparatively good results.

METRCIS



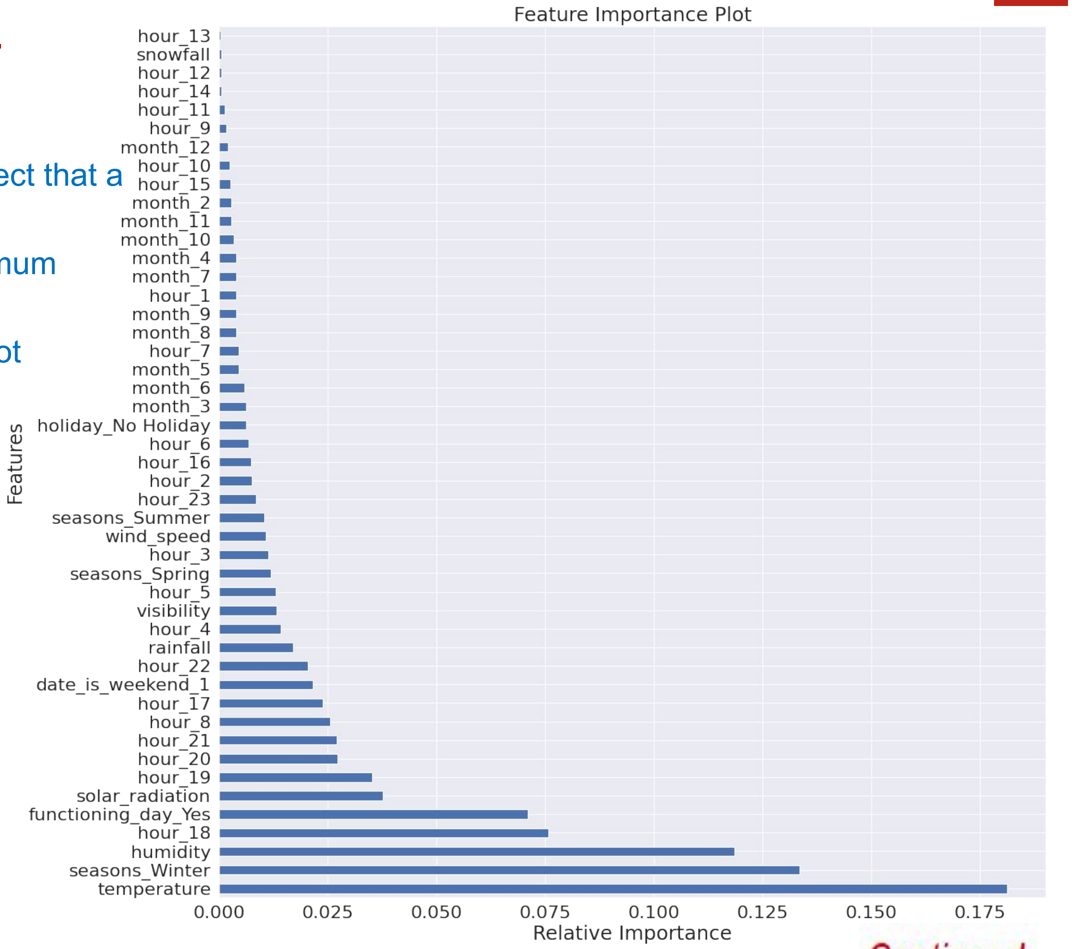
	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
	Dicision tree regression	5.285	55.554	7.453	0.641	0.63



Continued…

•MODELS AND MODEL SELECTION

- **DECISION TREE:**
- Importance of the feature simply depicts the effect that a feature affects the target variable.
- Temperature tops the list as it posses the maximum relative importance score.
- Feature that is least apart from those are one hot encoded is snowfall.

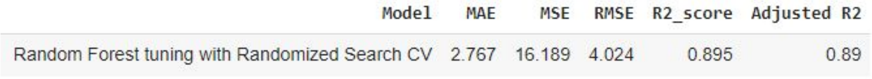


Continued...

•MODELS AND MODEL SELECTION

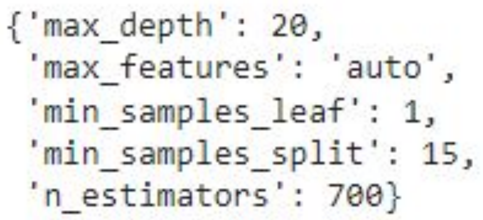
- **Random Forest tuning with Randomized Search CV:**
- Data frame shows the metrics from calculated on the test part of the datasets.
- Using Randomized Search CV we got our best parameters.
- Since R2 is represents the goodness of fit of a model. Here we have R2 as 0.895.
- Therefore, we can say that our model gave us good results.

METRCIS



Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Random Forest tuning with Randomized Search CV	2.767	16.189	4.024	0.895	0.89

BEST HYPERPARAMETERS



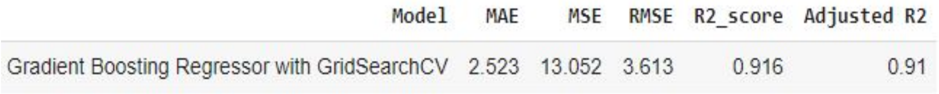
```
{'max_depth': 20,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 15,
 'n_estimators': 700}
```

Continued…

•MODELS AND MODEL SELECTION

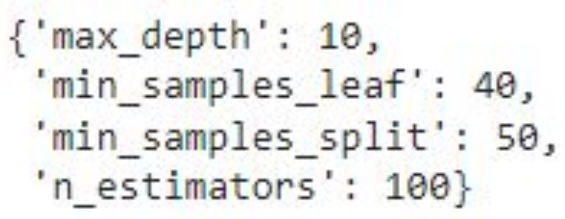
- **Gradient Boosting Regressor with Grid Search CV:**
- Data frame shows the metrics from calculated on the test part of the datasets.
- Using Grid Search CV we got our best hyper parameters.
- Since R2 is represents the goodness of fit of a model. Here we have R2 as 0.916.
- Therefore, we can say that our model gave us great results.

METRCIS



Model	MAE	MSE	RMSE	R2_score	Adjusted R2
Gradient Boosting Regressor with GridSearchCV	2.523	13.052	3.613	0.916	0.91

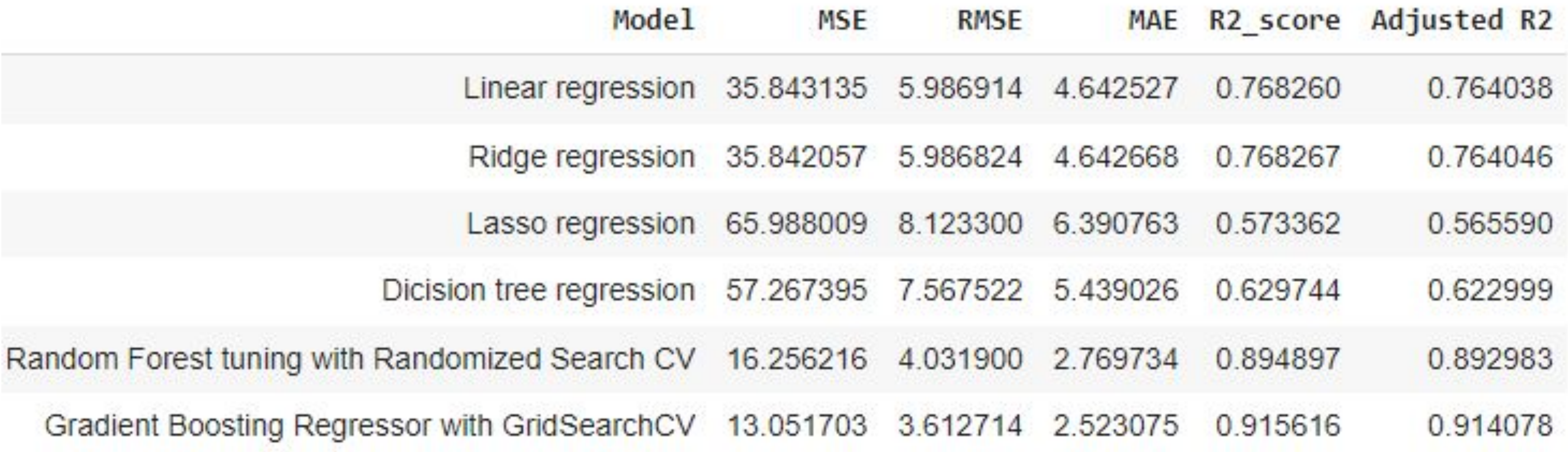
BEST HYPERPARAMETERS



```
{'max_depth': 10,
 'min_samples_leaf': 40,
 'min_samples_split': 50,
 'n_estimators': 100}
```

Continued...

COMPARISON AMONG ALL THE MODELS ON METRIC SCORE:



Model	MSE	RMSE	MAE	R2_score	Adjusted R2
Linear regression	35.843135	5.986914	4.642527	0.768260	0.764038
Ridge regression	35.842057	5.986824	4.642668	0.768267	0.764046
Lasso regression	65.988009	8.123300	6.390763	0.573362	0.565590
Dicision tree regression	57.267395	7.567522	5.439026	0.629744	0.622999
Random Forest tuning with Randomized Search CV	16.256216	4.031900	2.769734	0.894897	0.892983
Gradient Boosting Regressor with GridSearchCV	13.051703	3.612714	2.523075	0.915616	0.914078

- **OBSERVATIONS**

  - There was a high correlation between the dependent variables specifically dew point temperature and temperature due to which the VIF was going very high.
  - Temperature, Wind Speed, Dew Point Temperature, Solar Radiation, Visibility are positively correlated with the target variable.
  - Humidity, Rainfall, Snowfall are negatively correlated with the target variable.
  - In general people used rented bikes during their commuting hours i.e. from 7am to 9am in morning and 5pm to 7pm in the evening.
  - Weekdays are the ones where the demand of the bikes is comparatively high as compared with the weekends.
  - Summer season was the most preferred season throughout the year where the count was very high
  - As seen in the comparison table Linear Regression as well as Ridge Regression gave us good comparatively results.
  - But the Lasso Regression provided the worst results.
  - Decision Tree gave the results that were fairly good.
  - Random Forest tuning with Randomized Search CV, gave us better results but the exceptionally best result was obtained from Gradient Boosting Regressor with Grid Search CV.
  - Temperature came out to be the most important feature out of all from the list.

- **Conclusion:**

  - Since Gradient Boosting Regressor with Grid Search CV gave us the best results i.e. R2 score to be .916.
  - Therefore, we can deploy it for our predictions. Also, As this data is time dependent, the values for variables will not always be consistent. Therefore, we need constantly keep checking for the models.

Thank You