Capstone Project - 2

Bike Sharing Demand Predictions

Team Members

Devarshi Dwivedi
Jay Pardeshi
Priyadarshani Gaikwad
Samarjeet singh chhabra
Mohd. Anas Ansari

Table Of Contents

1. 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.

We have been provided with the bike rental demand data of 2017 December to November 2018 (1 year total), in order to perform supervised linear regression and identify the relationships between different variables. The data includes information on the number of bike rentals, the weather conditions, and the day of the week. The goal is to use this data to predict the demand for bike rentals in the future.

2. Understanding The Problem From A Business Perspective

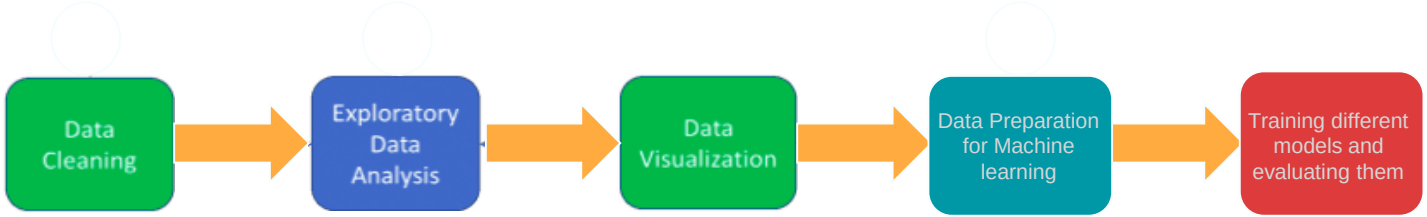
Why this business is progressing fast?

Revenue in the Bike-sharing segment is projected to reach US$7.96bn in 2022.
In the Bike-sharing segment, the number of users is expected to amount to 930.3m users by 2026.
User penetration is 10.0% in 2022 and is expected to hit 11.8% by 2026.
Bike sharing business has a profit margin of almost 60%.
Bike sharing business is eco-friendly way of commute and so it attracts more customers.
Bike sharing businesses are also purpose-driven businesses in a way, as they help the environment to reduce its carbon footprint and air pollution.

Understanding the problem.

As a business, our first priority is to benefit and serve our customers with the best service and no delay in service.
To avoid a bike shortage for our customers and to save their time, we need to be prepared beforehand. For that, we have to use our past data for predictions.
We can benefit our business by predicting how many bikes could be needed at a specified time, by just taking a look at the temperature, time, day, weather and many other things.
We can predict our revenue, profits and operation cost by predicting the total number of bikes that could be rented on a day.

3. Steps Followed



Data cleaning : We cleaned the data by dropping or replacing null values, deleting unwanted columns, Creating new features and we performed many other operations to get the required dataset.

EDA: We performed EDA on dataset to get better understanding.

Data Visualization: We visualized the trends and patterns using seaborn and matplotlib.

Data Preparation for Machine learning:- cleaned, transformed, removed outliers, fixed skewness, and scaled dataset/features for better model working.

Training models: Trained many different models to get the best model for prediction.

4. Understanding The Dataset Provided

The data has 8760 instances and 14 features.

All columns heading and data description:

Date - Date on which the bike was rented(year-month-day).

Rented Bike count - Number of bikes rented in that hour.

Hour - Hour of the day.

Temperature -Temperature at that time (Celsius).

Humidity - Percentage of humidity in air (%).

Wind Speed - Speed of the wind (m/s).

Visibility - How far is the visibility (*10m).

Dew point temperature - The dew point is the temperature the air needs to be cooled at constant pressure to in order to achieve a relative humidity of 100%(°C).

Solar radiation - Solar radiation is the energy received on an area on earth from the Sun (MJ/m²).

Rainfall - Measure of Rainfall (mm).

Snowfall - Measure of Snowfall (cm).

Seasons - which season is bike rented (Winter, Spring, Summer, Autumn).

Holiday - Was it a holiday or not (Holiday/No holiday).

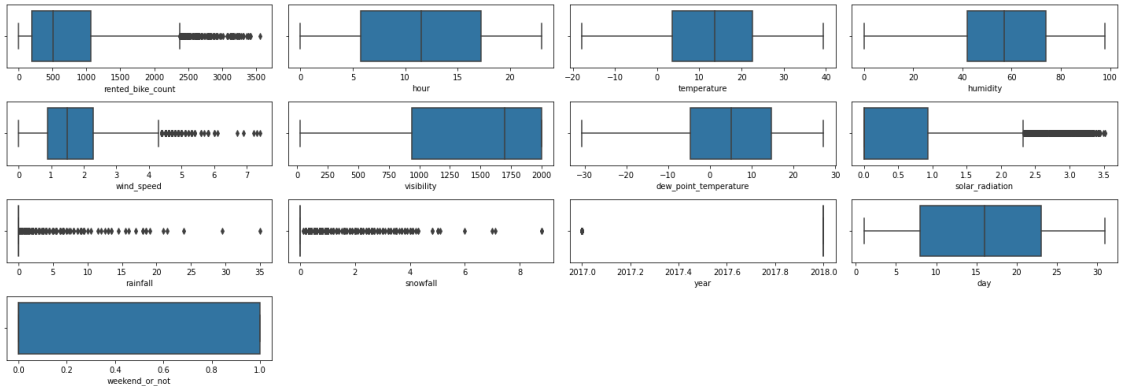
Functional Day - Was it a Functional day or not (Yes/No).

	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

5. Data Overview

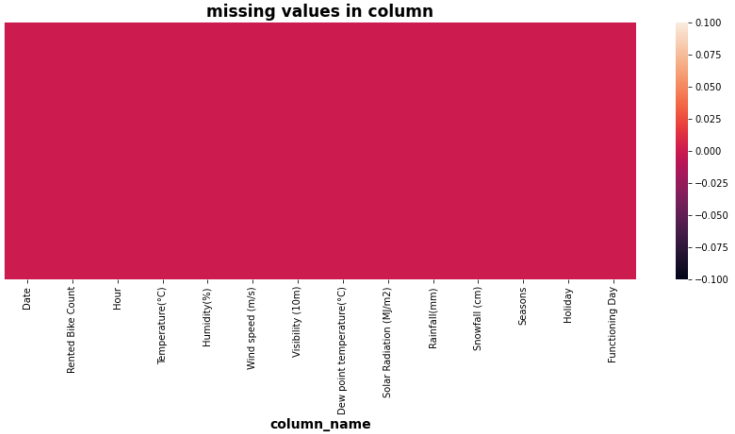
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8760 entries, 0 to 8759
Data columns (total 14 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Date                       8760 non-null   object
 1   Rented Bike Count          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)
memory usage: 958.2+ KB
```

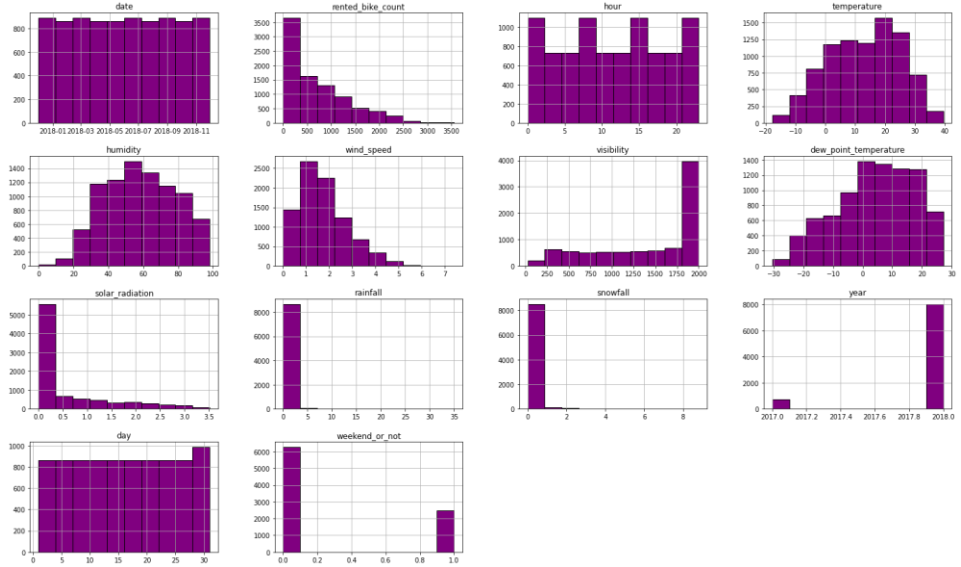
- We took a overview of the data together by using many methods such as **.head() , .tail() , .describe(), shape , .info()** and **etc.**



We checked for outliers in all numerical columns and we found **rented bike count** which is our dependent variable, has many outliers. Many other columns also have outliers which we will deal with later by using transformation.

6. Data Cleaning





We do not have any null values present, and also not any duplicate, whole dataset is very clean to be used for EDA.

Introduced some new columns such as month, day, year, day of week, weekday or weekend for EDA purposes.

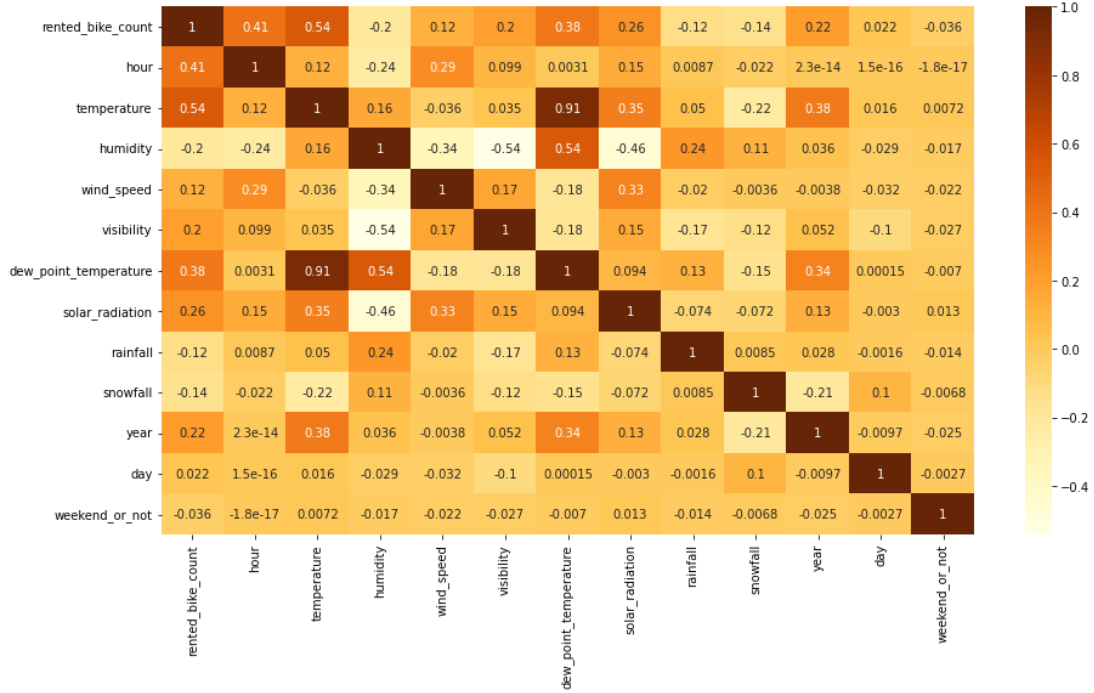
Brief of various column trends

Before we start getting insights from data here are histograms to have a brief picture of various column trends and data.

All columns with numerical data are represented in histograms.

7. EDA

1. Correlation between all features.

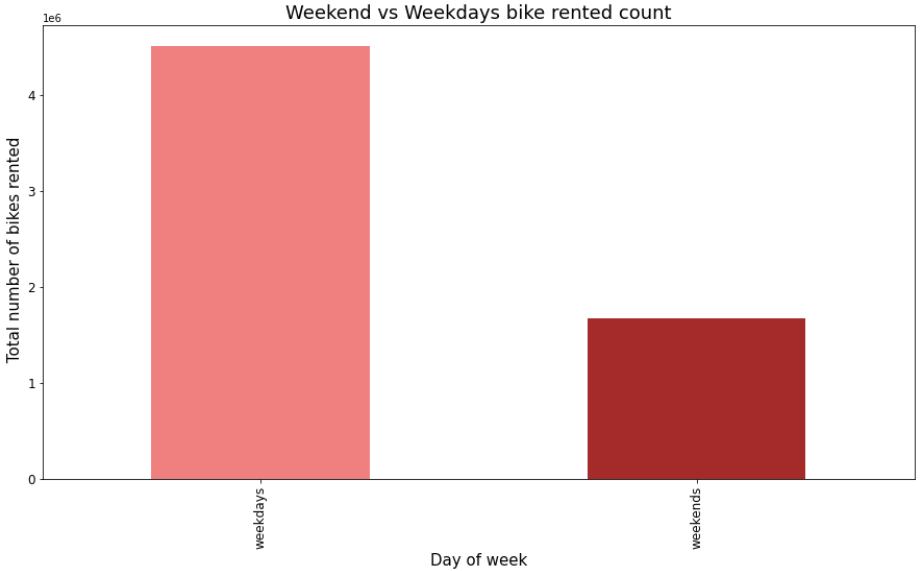


Some insights:-

- Day Hour is highly correlated to Number of bike rented, as the hour increase number of bikes rented also increase.
- Temperature, dew point and Bike rented count are highly correlated, as temperature and dew point increases , bike rented count increase
- Temperature and solar radiation are correlated.
- Wind speed and hour are correlated, as hours increase ina day wind speed also goes up.
- Temperature and dew point are extremely correlated.
- Visibility, solar radiation and Humidity are negatively correlated, as humidity increases, solar radiation and visibility decreases.

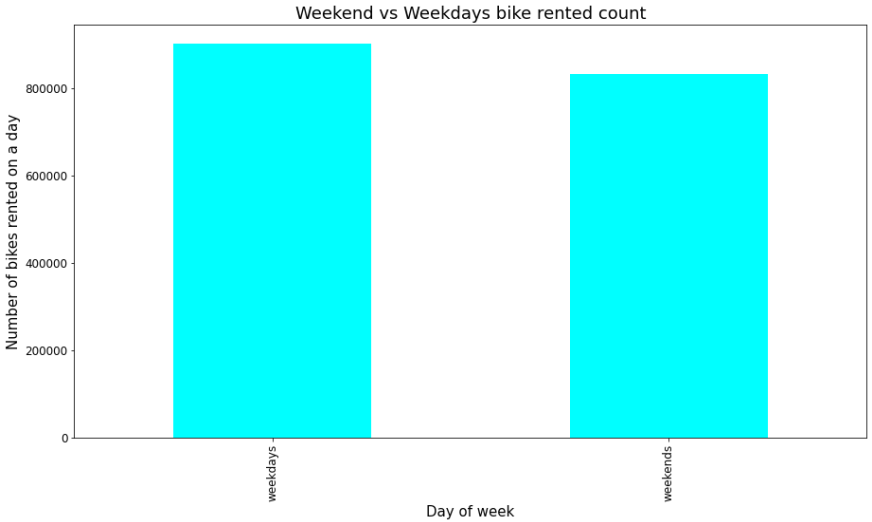
- Wind speed, solar radiation and visibility are positively correlated.
- Humidity and rainfall are positively correlated, more the rainfall more is the humidity and more the dew point

3. Does weekend has more rented bike count or weekdays?



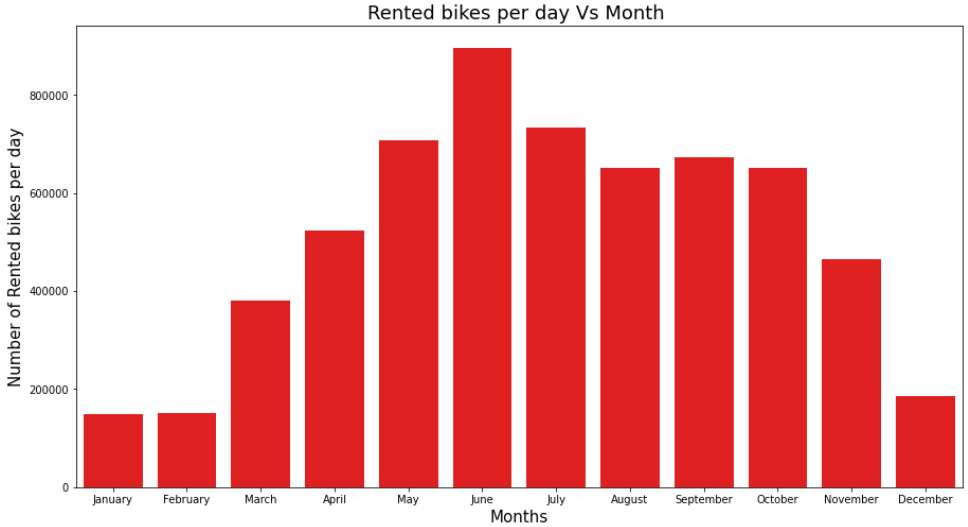
- We have more bookings on weekdays combined vs weekend

4. Which one has more bike rented counts: weekend or weekdays?



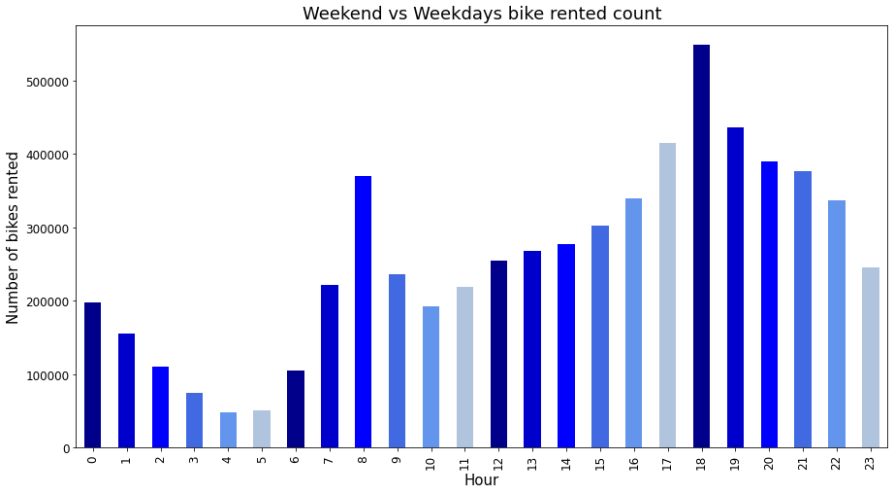
- On Weekdays on average 901325 bikes are rented
- On Weekends on average 832843 bikes are rented

5. Average Bike Rented Count every Month



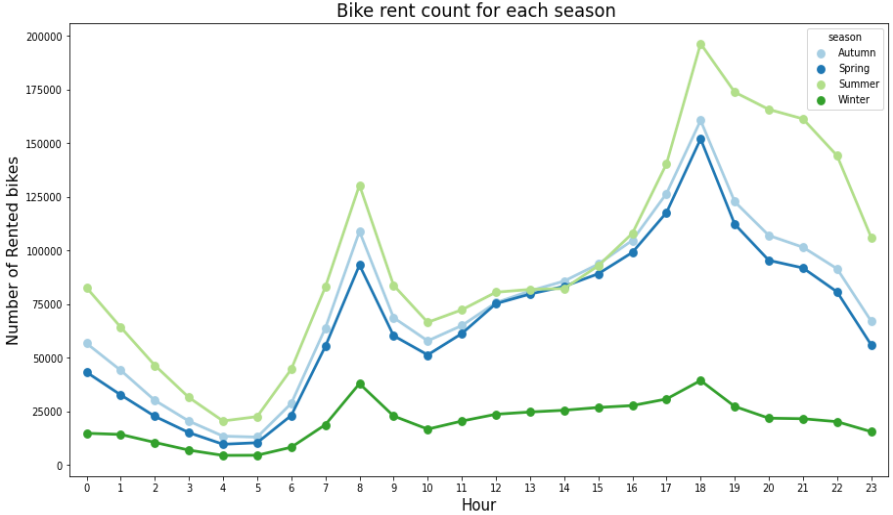
- Top 3 months where most bikes were rented are June, July and May.
- Peak periods when maximum bikes were rented is from May to October.
- The month in which least bikes were rented is January(150006) followed by February(151833) and December(185330).

6. Bike Rented Hour Wise.



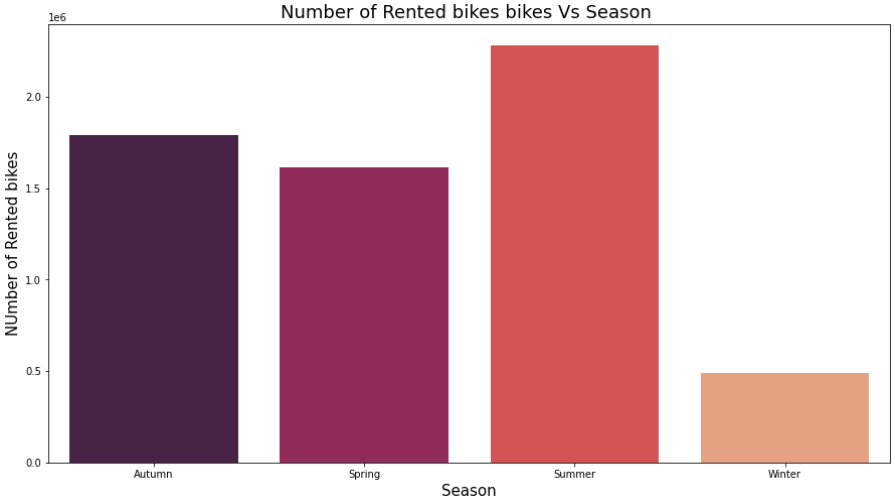
- Maximum number of bikes rented on average is in the 18th hour followed by 19th hour and 17th hour.
- Peak period for bike rented count is from 15th hour to 22nd hour, there is a slight increment in the 8th hour also.
- Minimum bikes were rented in 4th and 5th hour.

7. Hour wise bike rent count for each season.



- In Summer(green), Peak is at 18th hour and least is at 4th and 5th hour, and it has highest number of rented bike
- In Autumn(blue), Peak is at 18th hour and least is at 4th and 5th hour.
- In Spring(yellow), Peak is at 18th hour and least is at 4th and 5th hour.
- In winter(red), Peak is at 18th hour and least is at 4th and 5th hour, it has least number of bikes rented

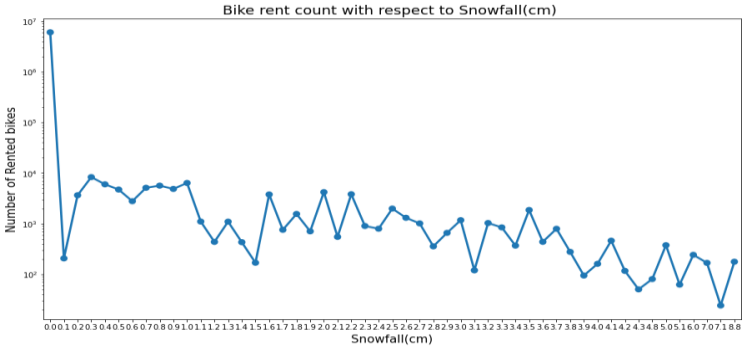
8. Season wise bike rent count



- We can see that there are very high demand for bike on rent in summer season , followed by autumn.
- Least numbers of bike were rented in winter season, that maybe because of cold and snow.

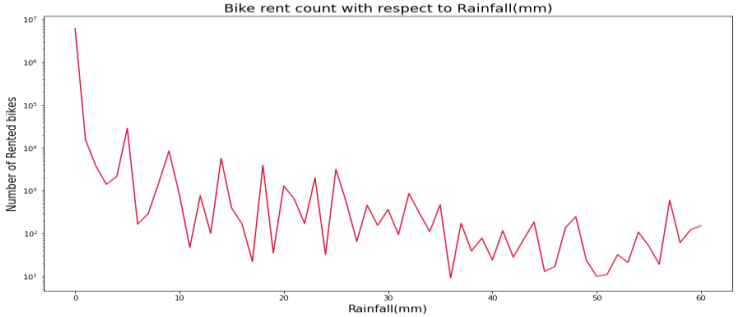
10. Snowfall effect on total bike rent count

- Above graph is in Logarithmic Y-axis Scale.
- As the Snowfall increases, the bike rent count decreases.



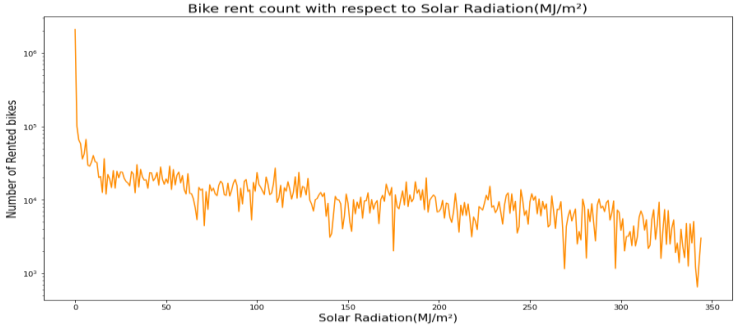
11. Effect of Rainfall on Bike rent count.

- Above graph is in Logarithmic Y-axis Scale.
- As the Snowfall increases, the bike rent count decreases.



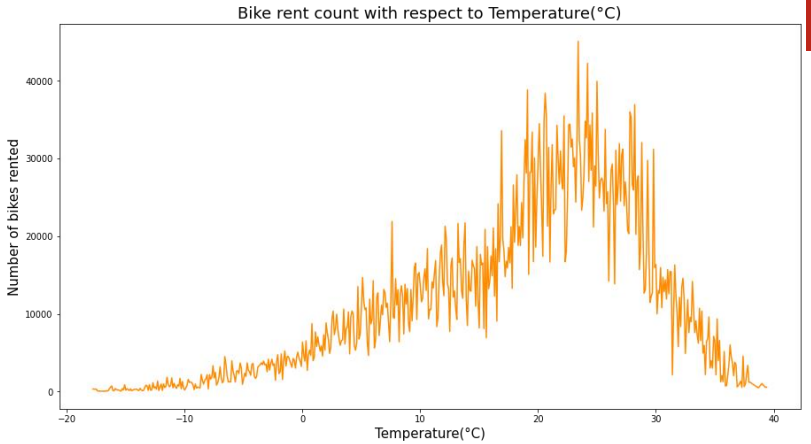
12. Effect of Solar radiation on Bike rented count

- Above graph is in Logarithmic Y-axis Scale.
- As the Solar Radiation(MJ/m²) increases, the bike rent count decreases rapidly(log scale).



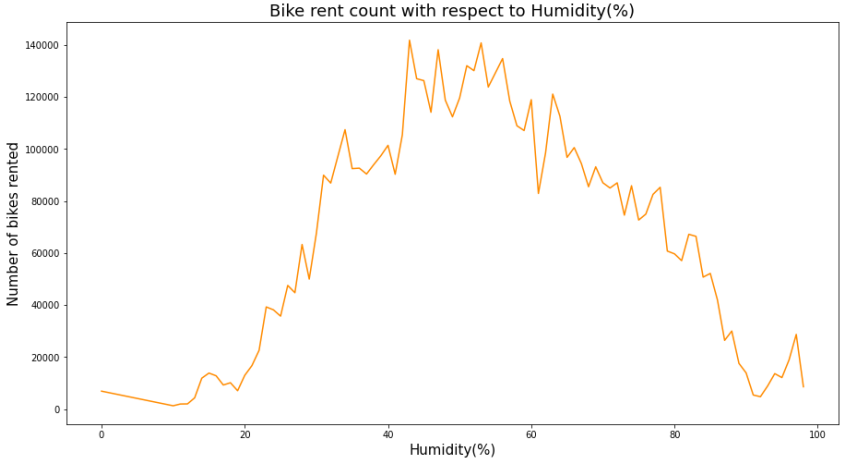
13. Bike Rent count with respect to temperature.

- We can see only at an optimal temperature maximum number of bikes were rented.
- The peak period is between temperature 15°C to 28°C. Temperature at which maximum bikes were rented is 23.4°C. Temperature at which minimum bikes were rented is -16.9°C.

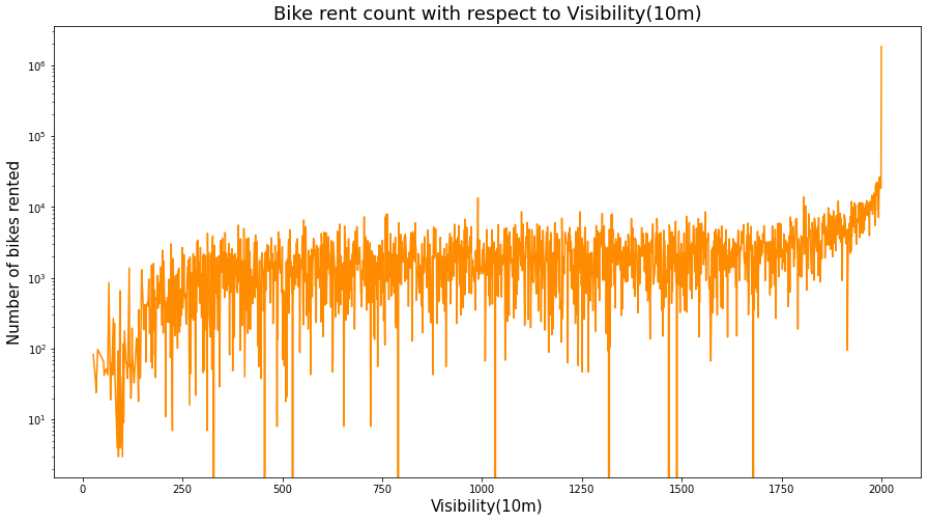


14. Bike Rent count with respect to Humidity.

- We can see only at an optimal Humidity percentage, maximum number of bikes were rented.
- Most appropriate humidity percentage is between 30% to 80%. Humidity at which maximum bikes were rented is 43%. Humidity at which minimum bikes were rented is 10%.

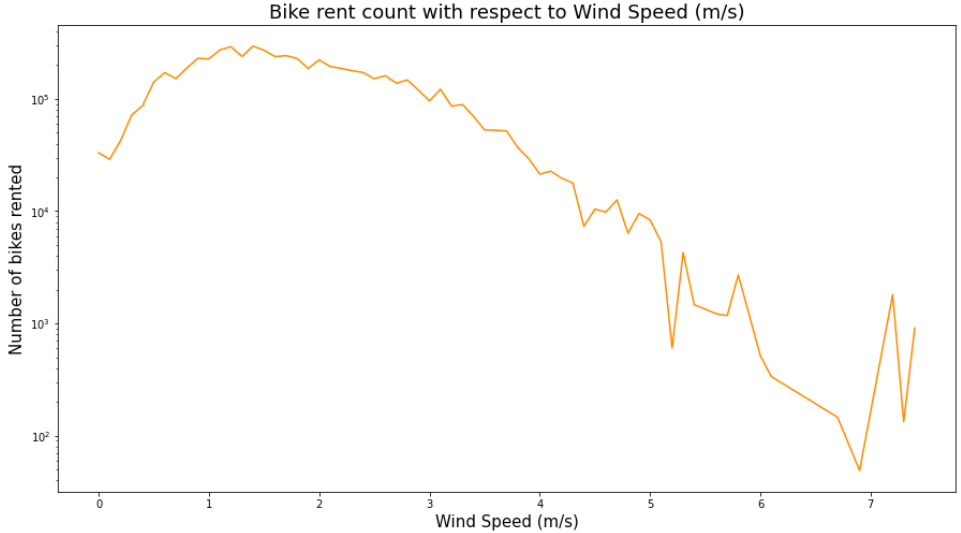


15. Effect of Visibility on no. of Bikes rented.



- We can see as the visibility increases so the number of bike which are rented.
- Visibility at which maximum bikes were rented is 20000 Meters

16. Effect of Wind speed on No. of bike rented.

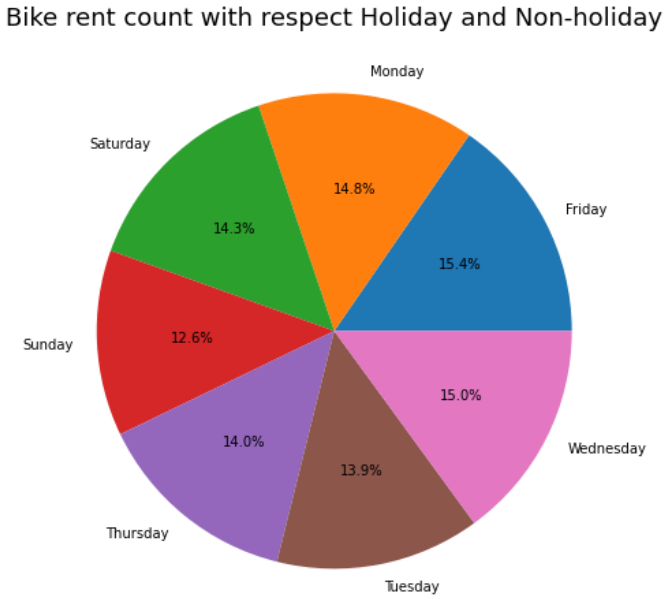


- We can see as the Wind Speed(m/s) increases the number of bike which are rented decreases.
- Wind Speed(m/s) at which maximum bikes were rented is 1.4 m/s.
- Wind Speed(m/s) at which minimum bikes were rented is 6.9 m/s

17. Bike rent count on holidays vs Non-holidays

- We have only 3.5% of total rented bikes count on Holiday, majority of bikes were rented on working days(not-holiday).

18. Rented bike count on each day of a week



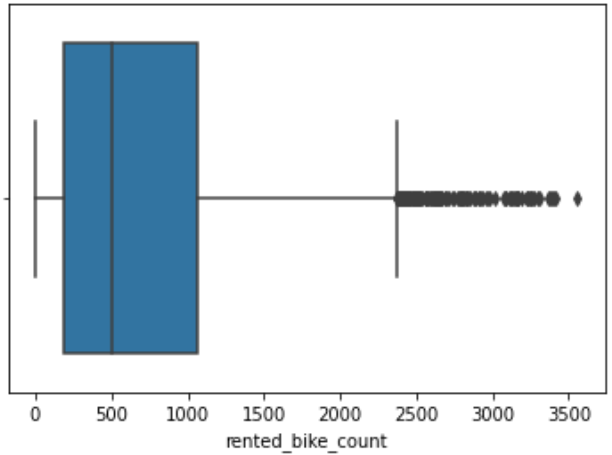
- Maximum bikes were rented on Friday(15.4%), On second place we have Wednesday(15.0%) and than Monday(14.8%).

8. Feature Engineering And Data Preparation for Model Training

1) Deleting Unuseful data :- deleted some columns which were unuseful to model.

2) Encoding all Categorical Features :- One-hot encoded all categorical features.

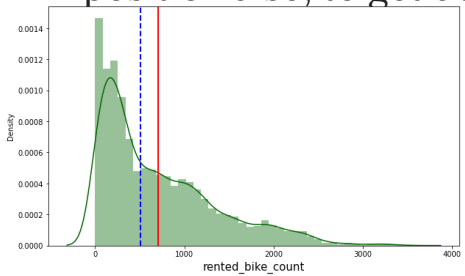
3) Dealing with outliers :- Used square root transformation later to deal with skewness and outliers in our dependent variable and some independent variable.

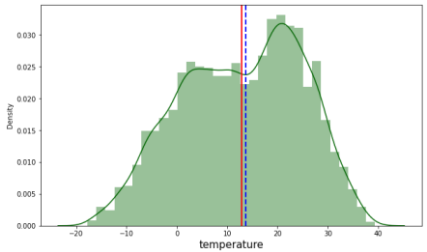


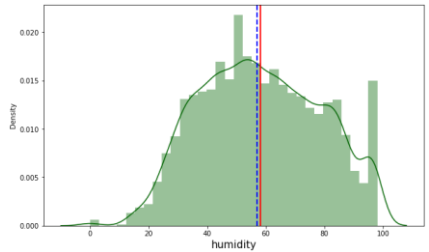
- we will be using square root transformation on it, and it will help in removing outliers, so we will not remove outliers in this way!

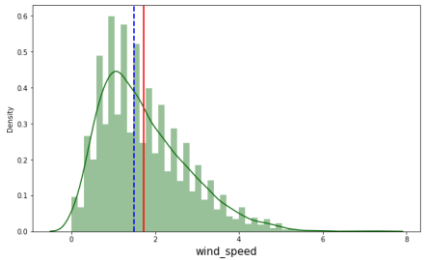
4) Analyzing all numerical columns:-

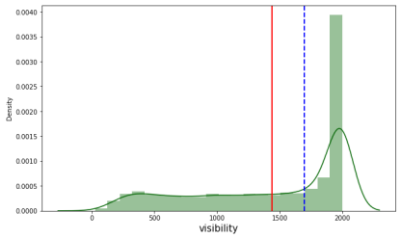
Used distribution plot on all numerical columns and visualized there mean and median position also, to get details about skewness.

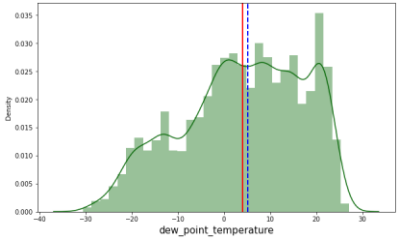


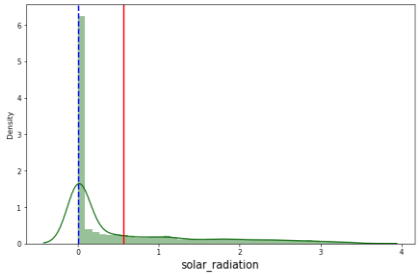


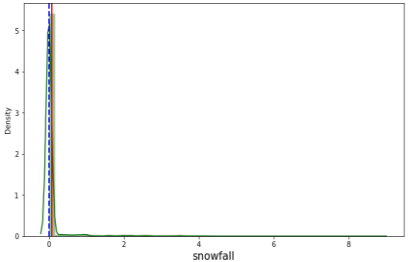


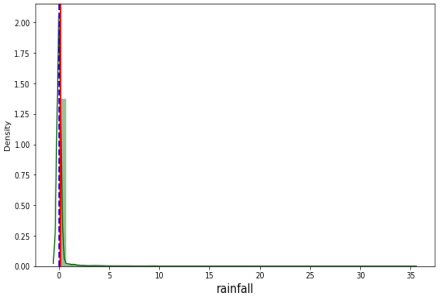










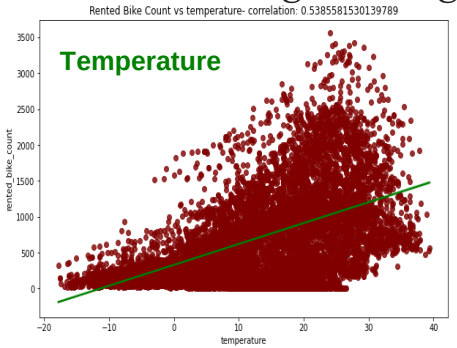


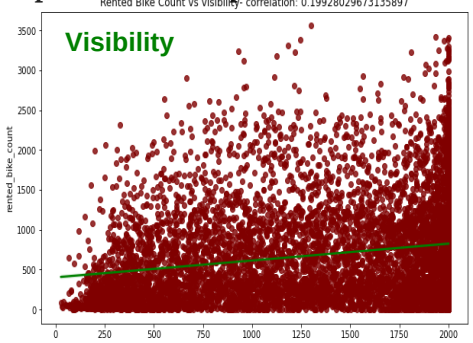
Some Insights:-

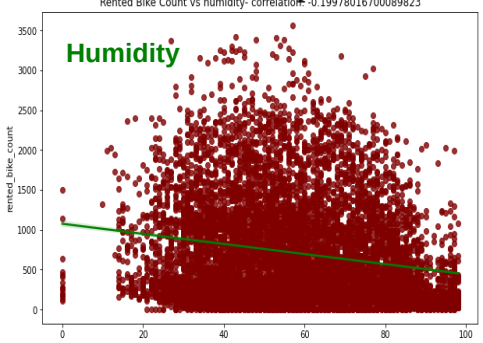
- If mean is greater than median it means it is positively skewed
- We can see many columns are skewed and for better model working we will transform some columns which are useful.

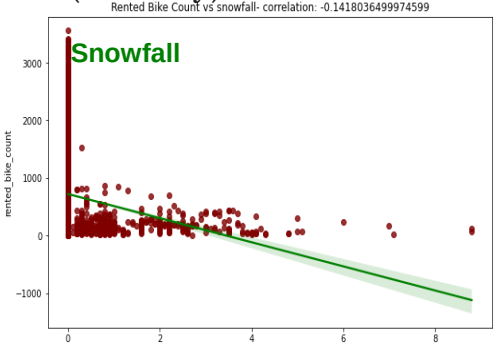
5) Regression plot for all columns:-

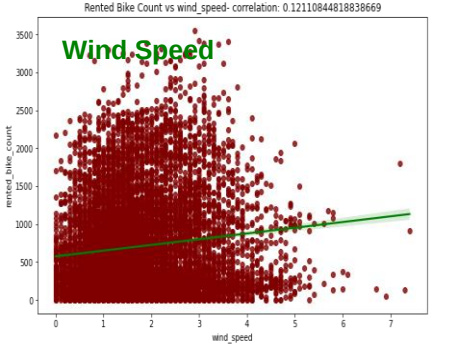
Plotted regression graph with respect to Rented bike count, except for encoded(dummy) variables.

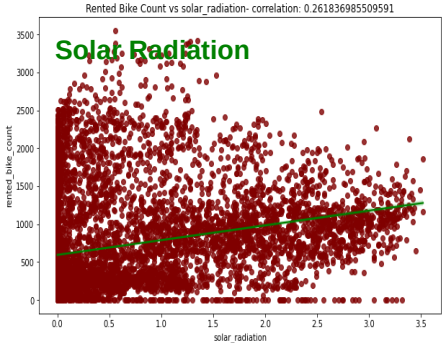


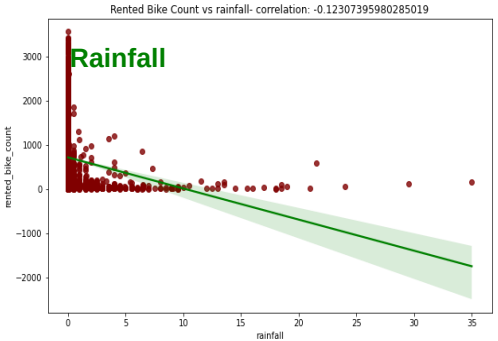


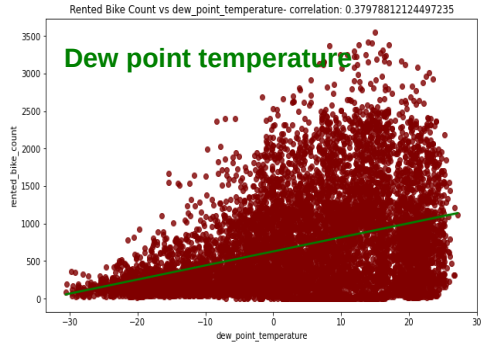












- All numeric Columns are correlated to Rented bike count columns and most of them are positively and few are negatively.
- Only snowfall, humidity and rainfall are negatively correlated.
- Temperature and Dewpoint temperature are more correlated to rented_bike_count column.

6) Multicollinearity check (VIF):-

	variables	VIF
0	temperature	29.075866
1	humidity	5.069743
2	wind_speed	4.517664
3	visibility	9.051931
4	dew_point_temperature	15.201989
5	solar_radiation	2.821604
6	rainfall	1.079919
7	snowfall	1.118903

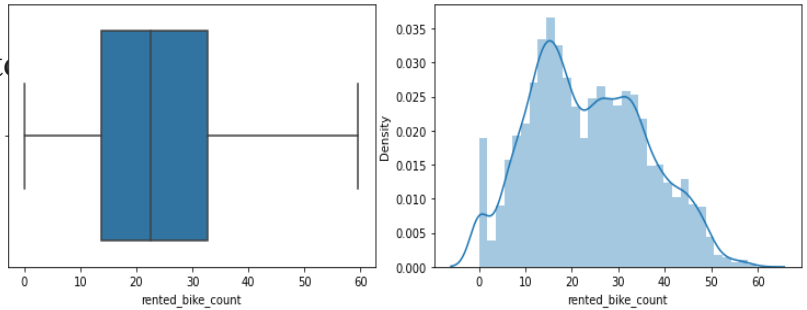
- Checked VIF for all independent variables(except Encoded ones) and found 2 very high VIF values for Temperature and Dew point temperature. And we deleted Dew point temperature.
- VIF below 1 is good, between 1-5 is moderate and after 5 its bad. So we focus mostly on those features which have VIF more than 5.

After:-

	variables	VIF
0	temperature	3.166007
1	humidity	4.758651
2	wind_speed	4.079926
3	visibility	4.409448
4	solar_radiation	2.246238
5	rainfall	1.078501
6	snowfall	1.118901

7) Transforming Data to get rid of Skewness:-

- We transformed many features to get rid of skewness and t deal with outliers. All columns that were transformed were Rented bike count, Wind speed, Solar radiation, rainfall, snowfall
- We used Square root transformation as it was better in removing skewness.
- Transformation also removed outliers in the important features(mainly in rented bike count).



After transforming, Skewness was fixed and outliers were also removed:-

9. Model Implementation and Explainability

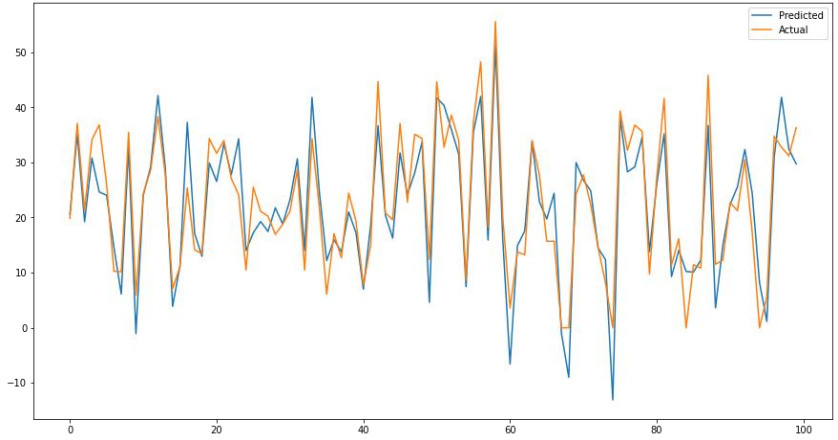
We will be Using these algorithms to train models.

1. Linear
2. Lasso
3. Ridge
4. Elastic Net
5. Polynomial
6. Decision_Tree
7. Random_Forest
8. Gradient_Boosting
9. Xtreme_GB

Linear Regression:-

Model Score : 0.8004990242543752
MSE : 29.988633654750085
RMSE : 5.47618787613702
R2 : 0.8064999529326514
Adjusted R2 : 0.8004601988133526

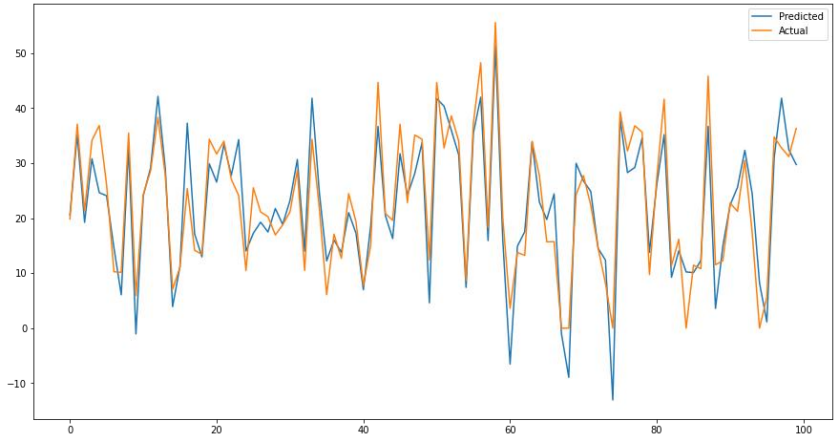
- We do not have desirable Adj R2 value.
- GridSearchcv is used for best parameters value.
- Model does not have very good accuracy.



Lasso Regression (L1)

Model Score : -30.925751009427138
MSE : 29.997003919111656
RMSE : 5.476952064708222
R2 : 0.8064459442516766
Adjusted R2 : 0.8004045043490493

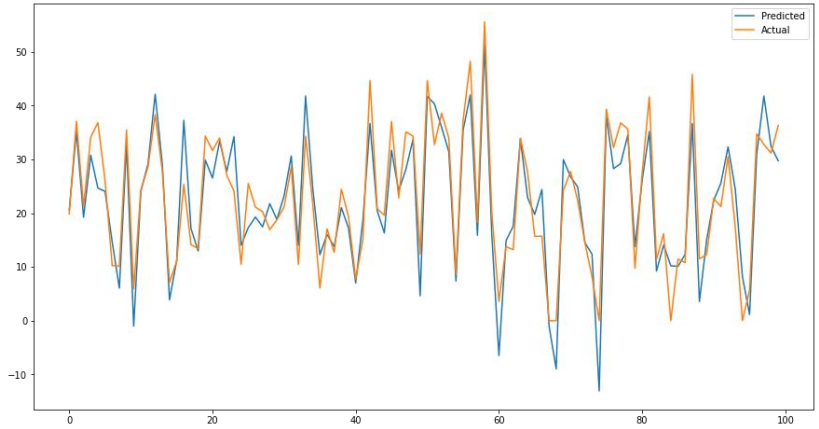
- We do not have desirable Adj R2 value.
- GridSearchcv is used for best parameters value.



Ridge Regression(L2)

Model Score : -30.926265211754078
MSE : 29.998153263431927
RMSE : 5.477056989244491
R2 : 0.8064385281692159
Adjusted R2 : 0.8003968567869829

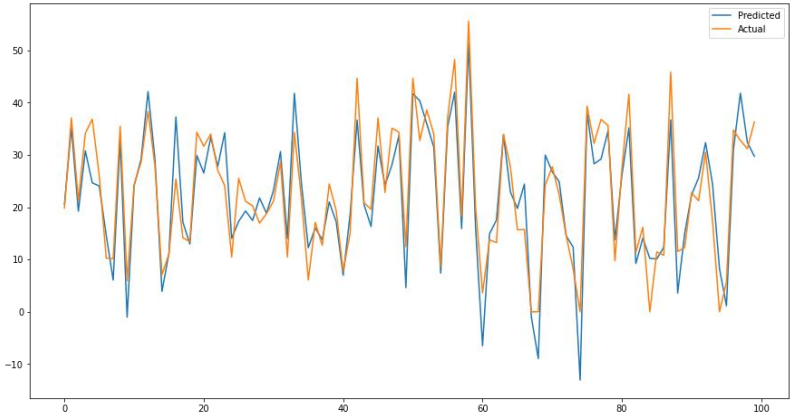
1)We do not have desirable Adj R2 value.
2)GridSearchcv is used for best parameters value.



Elastic Regression

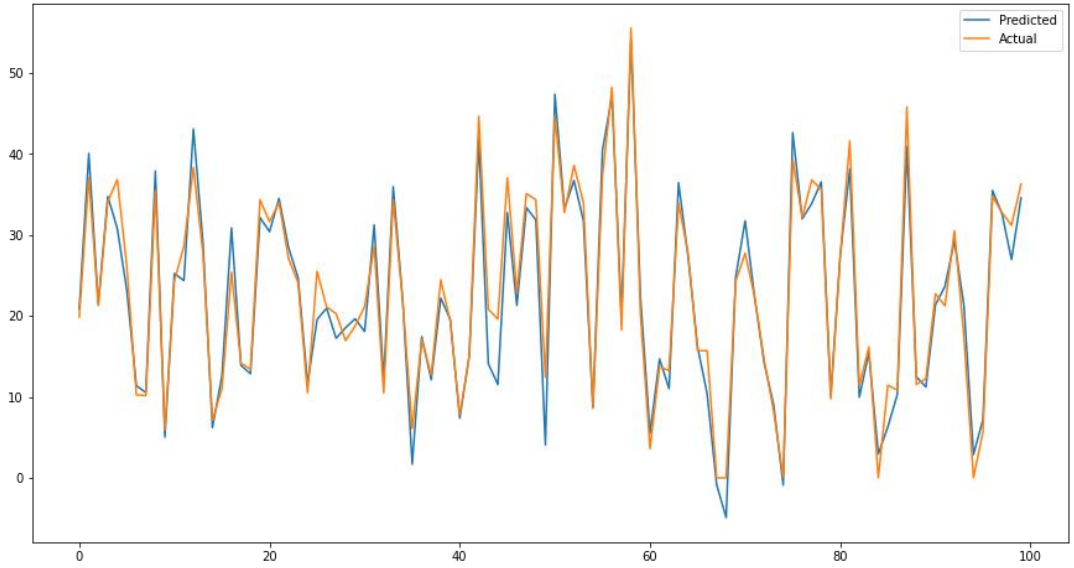
Model Score : -30.925938714201557
MSE : 29.997410014096054
RMSE : 5.476989137664603
R2 : 0.8064433239456134
Adjusted R2 : 0.8004018022548698

1)We do not have desirable Adj R2 value.
2)GridSearchcv is used for best parameters value.



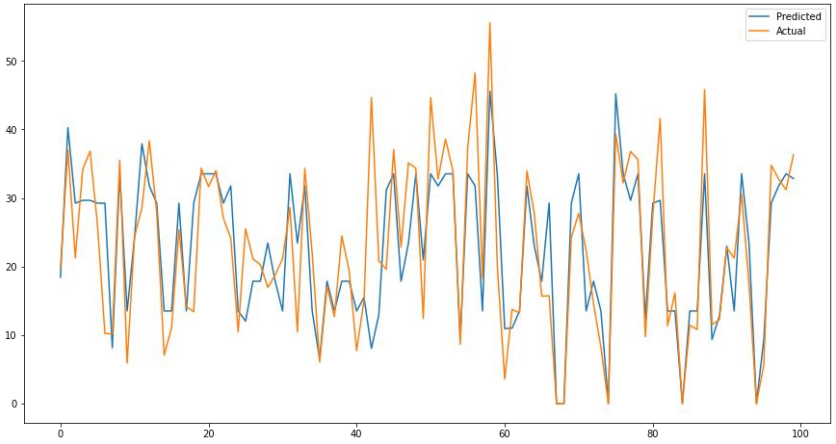
Polynomial Regression

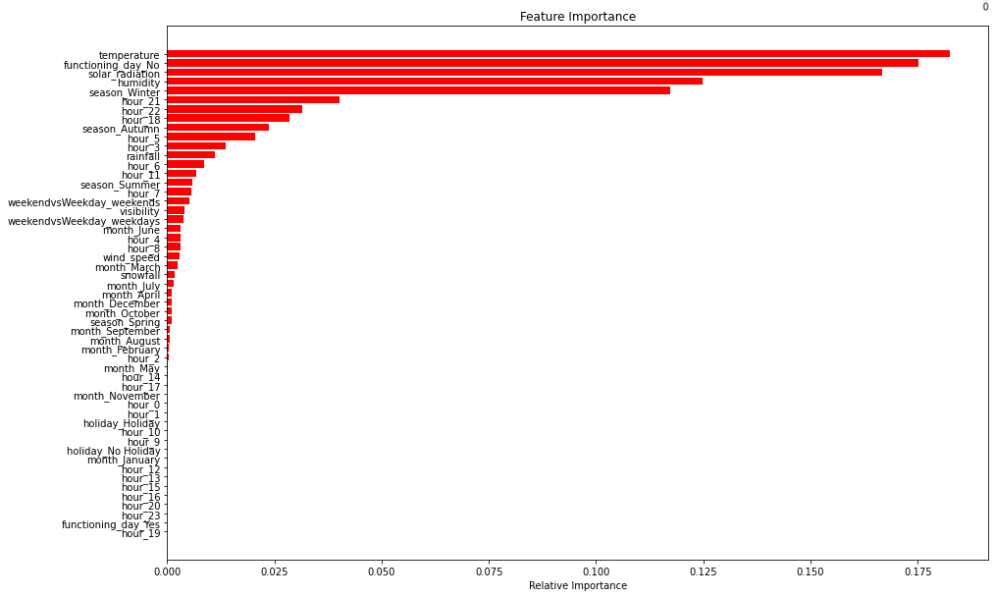
Model Score : 0.9389366493423017
MSE : 11.05553028489746
RMSE : 3.324985757096932
R2 : 0.9286647849611745
Adjusted R2 : 0.9264381851984785



Decision Tree Regression

Model Score : 0.7002252272794816
MSE : 54.54673655044349
RMSE : 7.385576250398034
R2 : 0.6480401137512566
Adjusted R2 : 0.6370543222487929



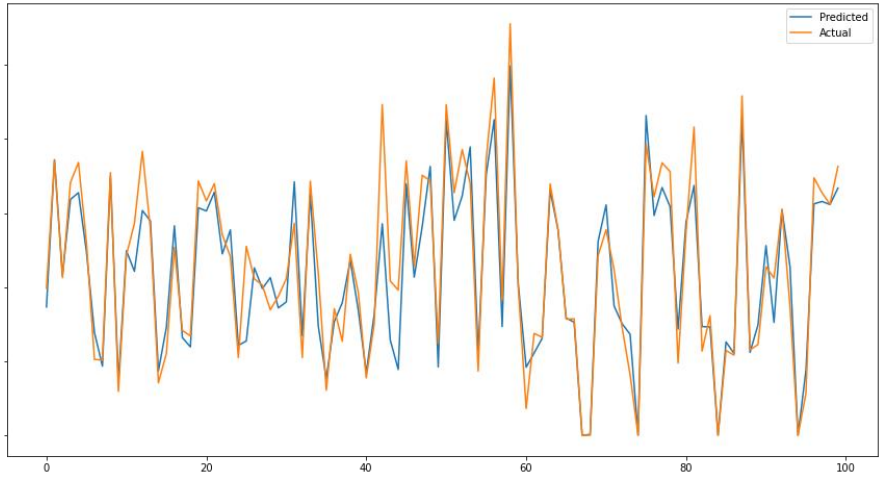


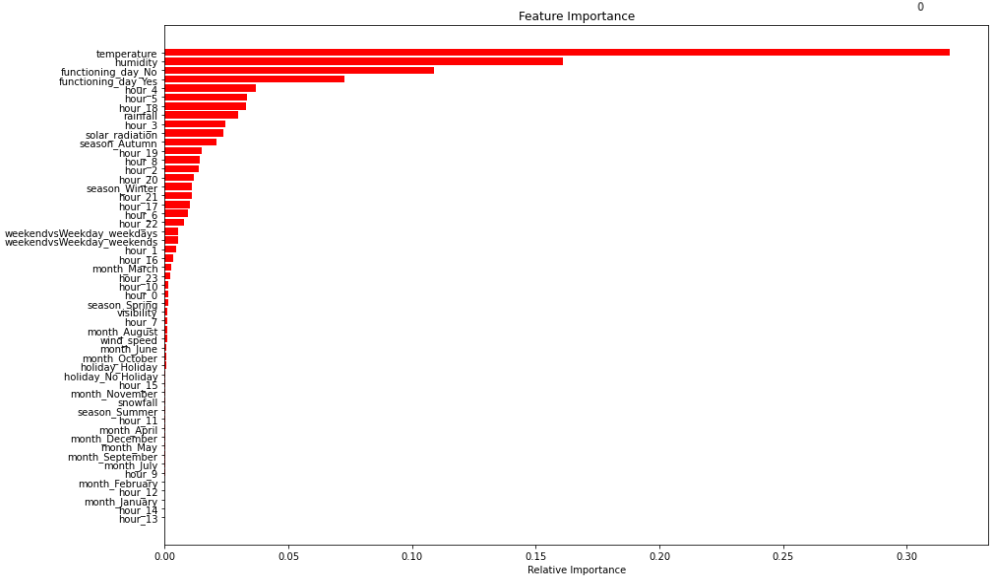
Feature Importance insights of Decision Tree:-

- Most important features for decision tree regressor are temperature, functioning-day-No and humidity.

Random forest

Model Score : 0.9017108894740667
MSE : 19.517834504546066
RMSE : 4.417899331644629
R2 : 0.8740622217483334
Adjusted R2 : 0.8701313016968973



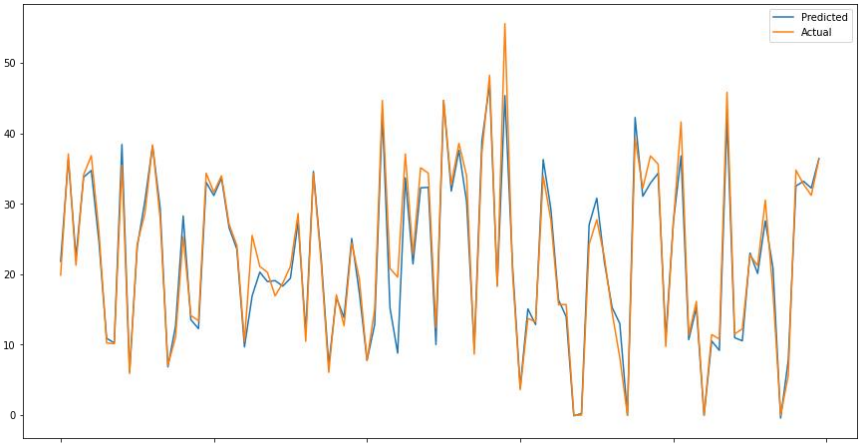


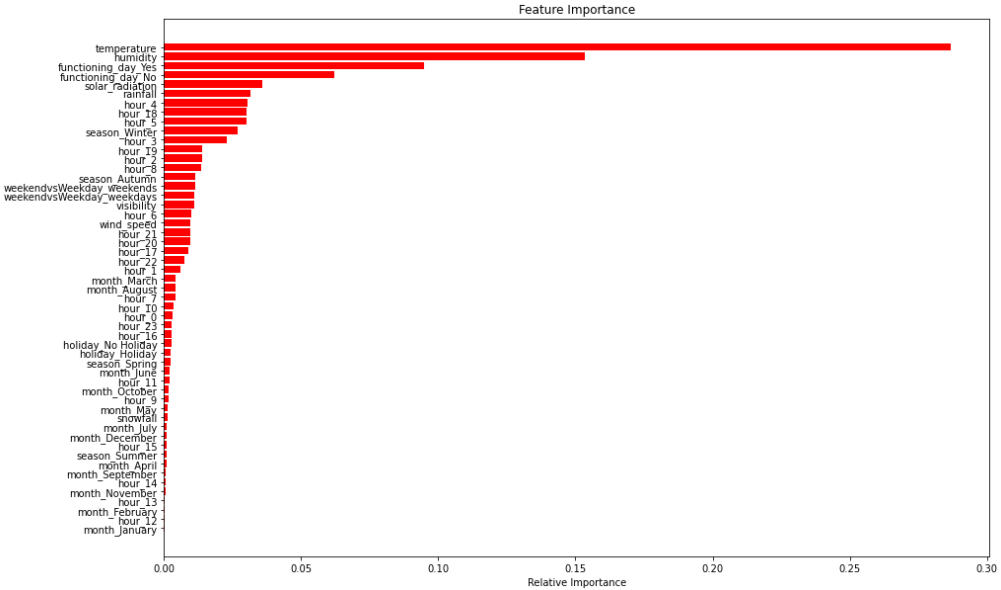
Feature Importance insights of Random Forest:-

- Most important features for Random forest are temperature, humidity, functioning-day-Yes and functioning-day-No

Gradient boosting

Model Score : 0.9999844256861317
MSE : 11.720656684593171
RMSE : 3.423544462190198
R2 : 0.9243730926110474
Adjusted R2 : 0.9220125354310624



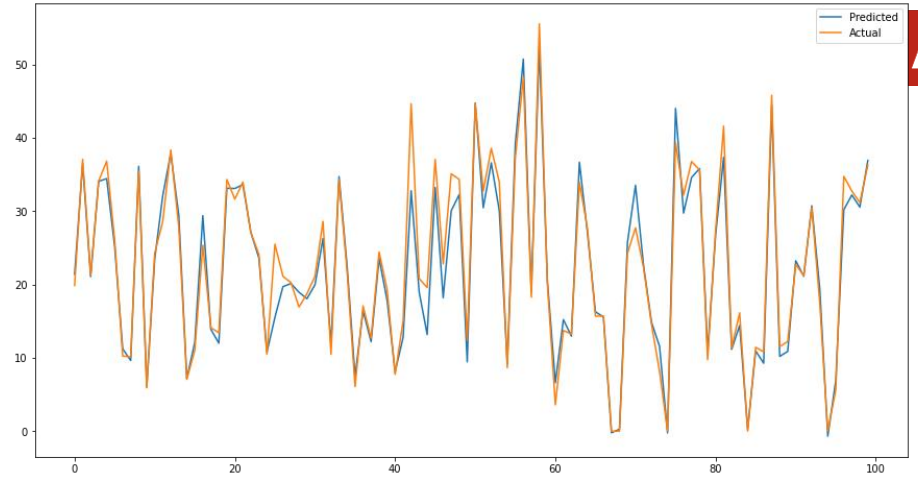


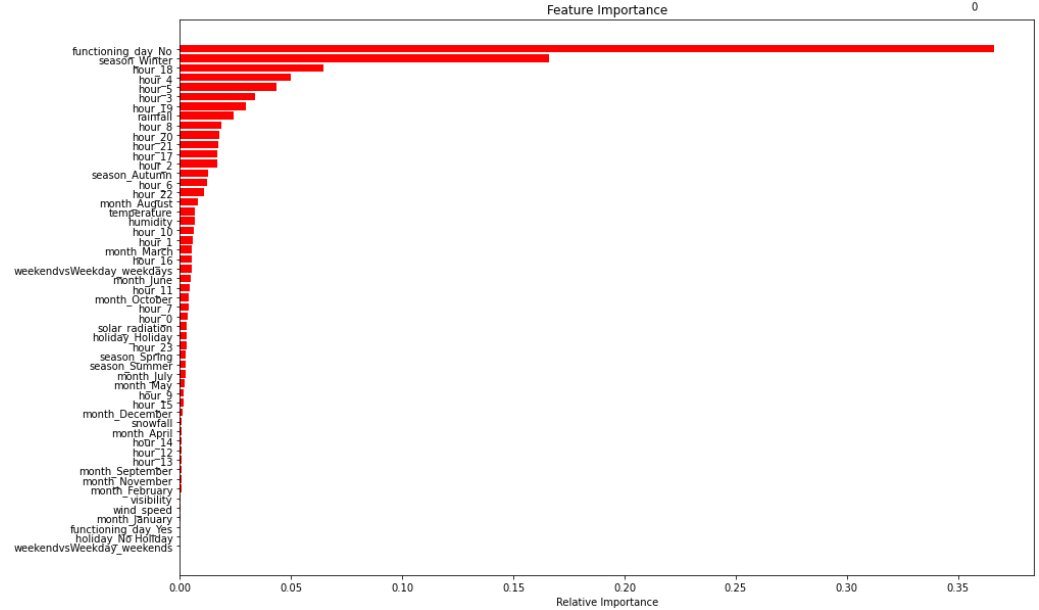
Feature Importance insights of Gradient Boosting:-

- Most important features for Gradient Boosting are temperature, humidity, functioning-day-Yes and functioning-day-No.

eXtreme Gradient boosting

Model Score : 0.999829426576048
MSE : 10.712847758067157
RMSE : 3.273048694728992
R2 : 0.9308759255497794
Adjusted R2 : 0.9287183425427937



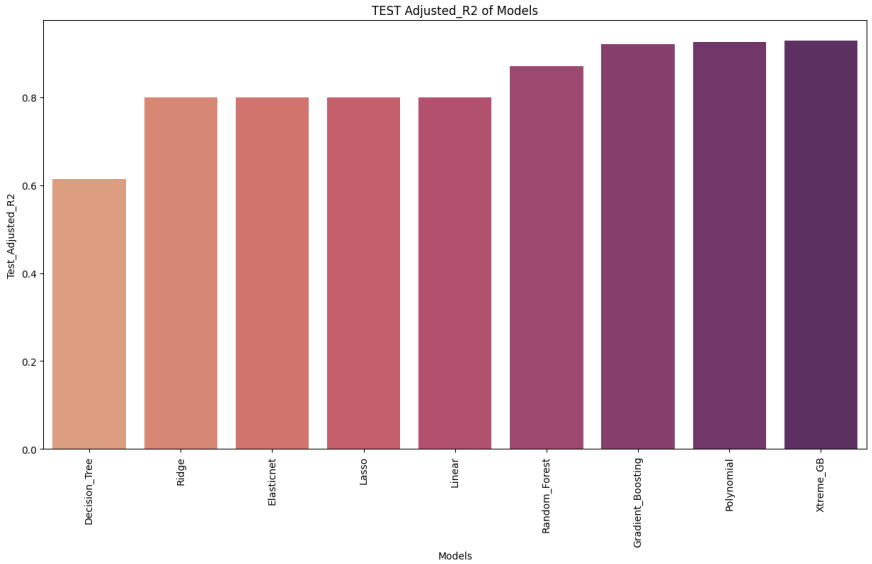


Feature Importance insights of eXtreme Gradient Boosting:-

- Most important features for eXtreme Gradient boosting are functioning-day-No, season_winter and hour_18

10. Evaluation Of All Models

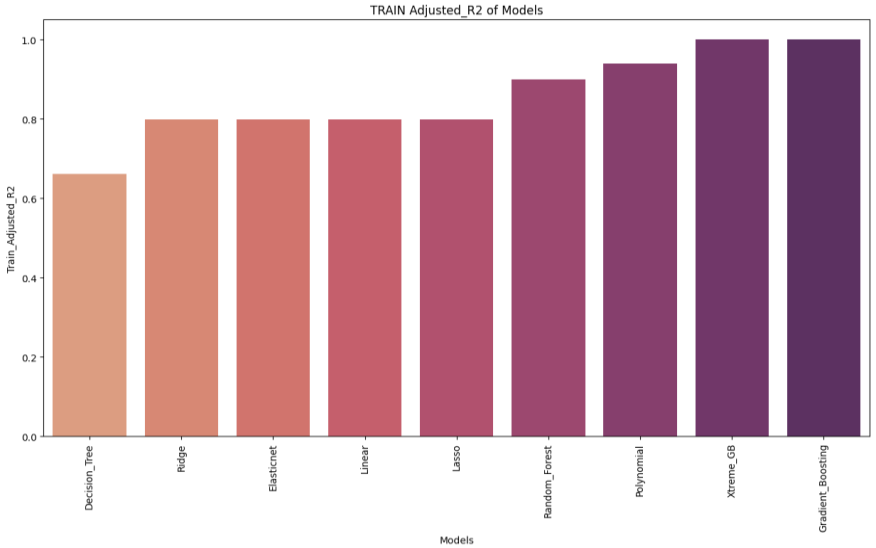
For Test Dataset



	Models	Test_Score	Test_Mean_square_error	Test_Root_Mean_square_error	Test_R2	Test_Adjusted_R2
0	Linear	0.800499	29.988634	5.476188	0.806500	0.800460
1	Lasso	-30.925751	29.997004	5.476952	0.806446	0.800405
2	Ridge	-30.926265	29.998153	5.477057	0.806439	0.800397
3	Elasticnet	-30.925939	29.997410	5.476989	0.806443	0.800402
4	Polynomial	0.938937	11.055530	3.324986	0.928665	0.926438
5	Decision_Tree	0.700225	54.546737	7.385576	0.648040	0.637054
6	Random_Forest	0.901711	19.517835	4.417899	0.874062	0.870131
7	Gradient_Boosting	0.999984	11.720657	3.423544	0.924373	0.922013
8	Xtreme_GB	0.999829	10.712848	3.273049	0.930876	0.928718

- The best accuracy algorithm for our test dataset is eXtreme Gradient boosting algorithm with Adj_R2 score of 0.928718.
- We have other algorithm such as polynomial, Gradient boosting, Random forest who also have optimal Adj_R2 score.

For Train Dataset



	Models	Train_Score	Train_Mean_square_error	Train_Root_Mean_square_error	Train_R2	Train_Adjusted_R2
0	Linear	0.800499	30.925854	5.561102	0.800499	0.798979
1	Lasso	-30.925751	30.925751	5.561093	0.800500	0.798979
2	Ridge	-30.926265	30.926265	5.561139	0.800496	0.798976
3	Elasticnet	-30.925939	30.925939	5.561109	0.800498	0.798978
4	Polynomial	0.938937	9.465800	3.076654	0.938937	0.938471
5	Decision_Tree	0.663369	52.183249	7.223797	0.663369	0.660803
6	Random_Forest	0.900540	15.417927	3.926567	0.900540	0.899782
7	Gradient_Boosting	0.999913	0.013439	0.115928	0.999913	0.999913
8	Xtreme_GB	0.999829	0.026442	0.162609	0.999829	0.999828

- The best accuracy algorithm for our train dataset is Decision tree algorithm with Adj_R2 score of 1.
- We have other algorithm such as Gradient boosting, Xtreme_GB and polynomial who also have optimal Adj_R2 score.

11. Conclusions

1. The best accuracy model for our test dataset is eXtreme Gradient boosting with Adj_R2 score of '0.928718'. We have other algorithm such as polynomial, Gradient boosting, Random forest who also have optimal Adj_R2 score.
2. The best Model to be used for this business is eXtreme Gradient boosting and polynomial for better accuracy.
3. The best accuracy model for our train dataset is Decision tree with Adj_R2 score of '1'. We have other models such as Gradient boosting, Xtreme_GB and polynomial who also have optimal Adj_R2 score.

12. Challenges

1. Pre-processing the data was one of the challenges we faced including removing highly correlated data so as to not hinder the performance of our regression model.
2. Selecting the appropriate models to maximize the accuracy of our predictions was one of the challenges faced.

Thank You