

Capstone Project-2

Supervised ML – Regression

Bike Sharing Demand Prediction

Presented by:

Vishal Raul

Contents

1. Introduction
2. Defining Problem Statement
3. Data Explanation and Preparation
4. Methodology
5. Exploratory Data Analysis(EDA)
6. Data visualization
7. ML Model – Regression
8. Conclusion



Introduction



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.

Prediction of bike sharing demand can help bike sharing companies to allocate bikes better and ensure a more sufficient circulation of bikes for customers. This presentation proposes a real-time method for predicting bike renting based on historical data, weather data, and time data. This demand prediction model can provide a significant theoretical basis for management strategies and vehicle scheduling in public bike rental system. The design of the learning algorithm includes preprocess of feature explanation and data selection, modeling and validation.

Defining Problem Statement

- Our task is to predict the demand of bike rent based on the historical usage over different factors such as seasons, weather, temperature, humidity etc.



Data Summary:

Data Set Name :- SeoulBikeData.csv

Statistics –

- Rows - 8760
- Features – 14

Columns:-

'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'

Data fields:-

Date : year-month-day

Rented Bike count : Count of bikes rented at each hour

Hour : Hour of the day

Temperature : Temperature in Celsius

Humidity : %

Windspeed : m/s

Visibility : 10m

Dew point temperature : Celsius

Solar radiation : MJ/m²

Rainfall : mm

Snowfall : cm

Seasons : Winter, Spring, Summer, Autumn

Holiday : Holiday/No holiday

Functional Day : NoFunc(Non Functional Hours), Fun(Functional hours)

Loading the Dataset

```
#loading the dataset  
data = pd.read_csv('/content/drive/MyDrive/Capstone Project-02(Re)/SeoulBikeData.csv', encoding='latin1')
```

```
data.head()
```

	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

Attribute Information:- Dtypes and Null values

```
#data information
```

```
data.info()
```

```
<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
```

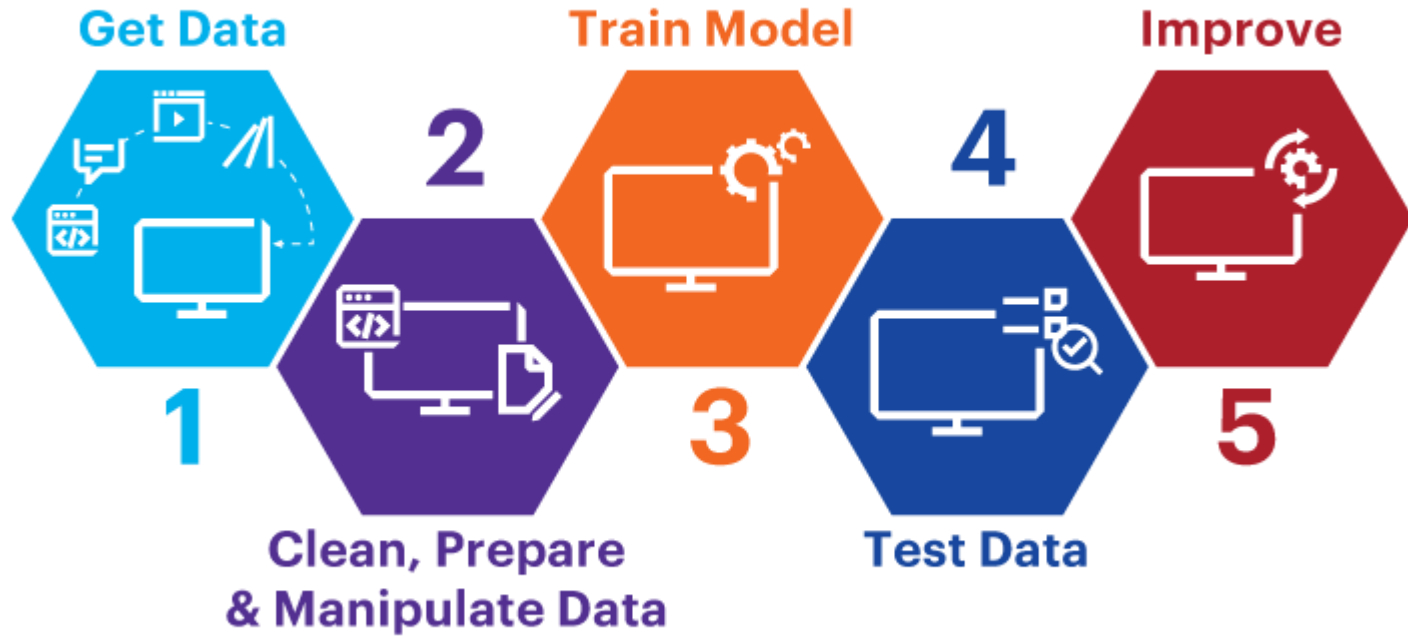
```
#data attribute null values
```

```
data.isna().sum()
```

Date	0
Rented Bike Count	0
Hour	0
Temperature(°C)	0
Humidity(%)	0
Wind speed (m/s)	0
Visibility (10m)	0
Dew point temperature(°C)	0
Solar Radiation (MJ/m2)	0
Rainfall(mm)	0
Snowfall (cm)	0
Seasons	0
Holiday	0
Functioning Day	0

```
dtype: int64
```


METHODOLOGY



Machine Learning Models

- Linear Regression
- Lasso Regression
- Ridge Regression
- Decision Tree Regression
- Random Forest Regression
- Gradient Boosting Regression
- XGBoost Regression



Descriptive Statistics

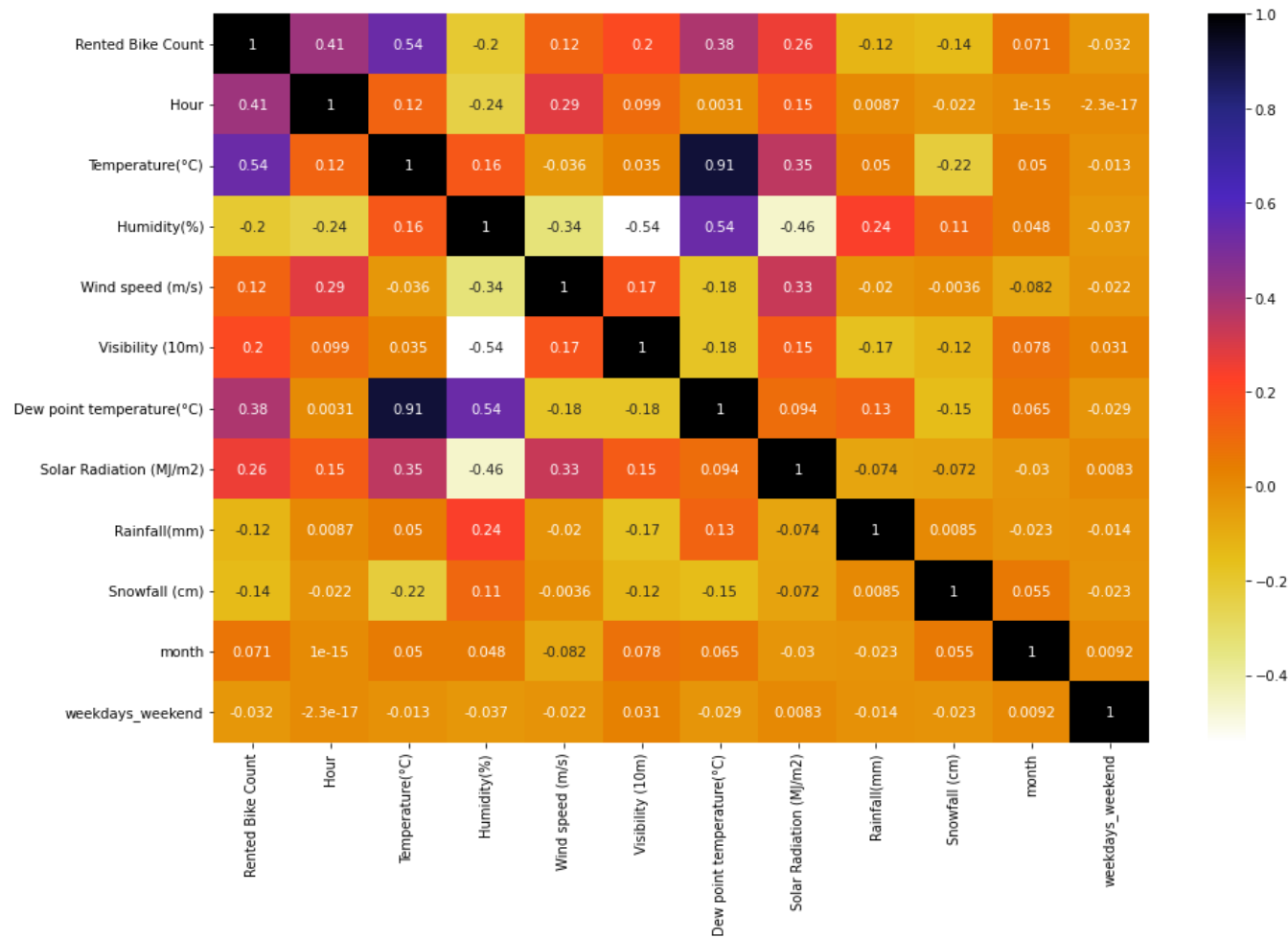


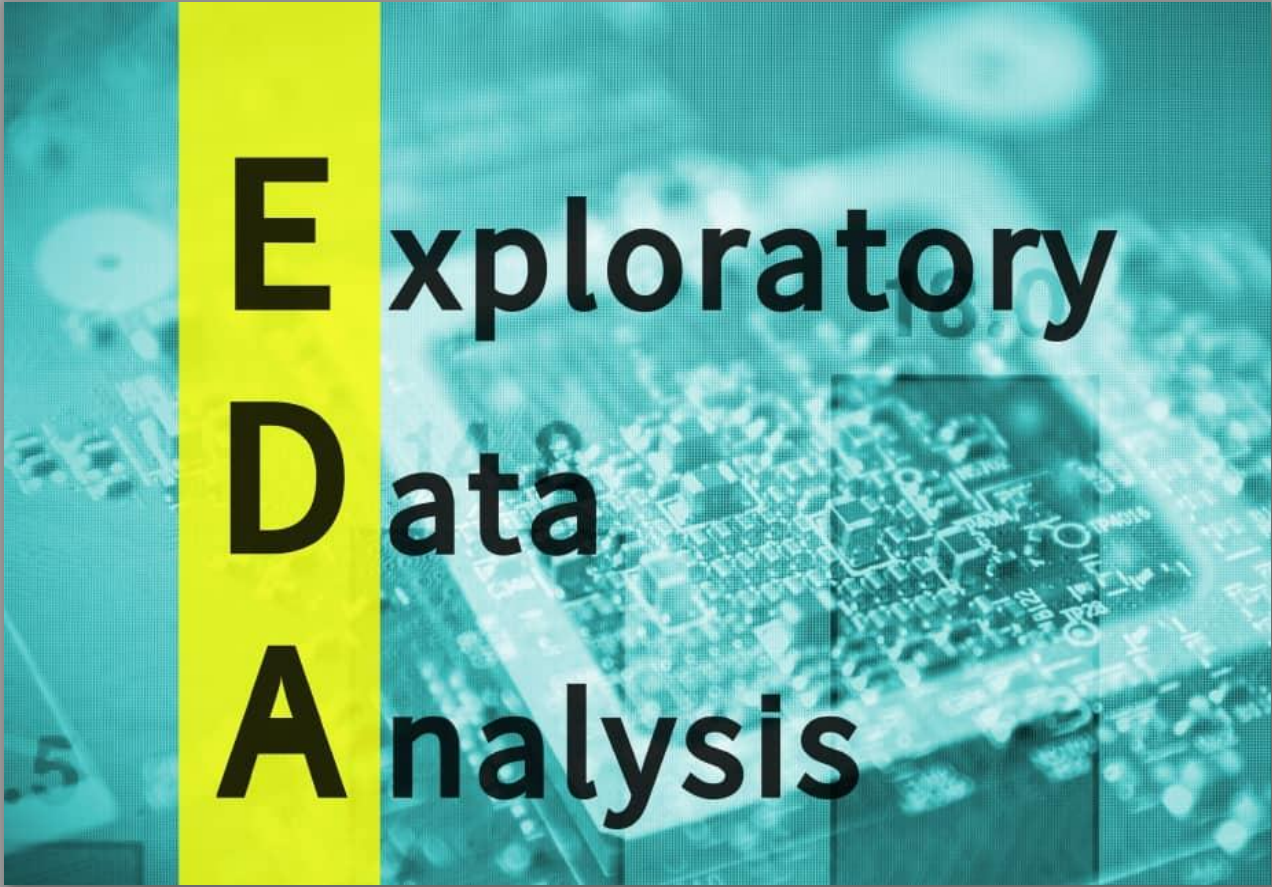
```
#descriptive statistics  
data.describe()
```

	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	day	weekdays_weekend
count	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000
mean	704.602055	11.500000	12.882922	58.226256	1.724909	1436.825799	4.073813	0.569111	0.148687	0.075068	6.526027	15.720548	0.290411
std	644.997468	6.922582	11.944825	20.362413	1.036300	608.298712	13.060369	0.868746	1.128193	0.436746	3.448048	8.796749	0.453978
min	0.000000	0.000000	-17.800000	0.000000	0.000000	27.000000	-30.600000	0.000000	0.000000	0.000000	1.000000	1.000000	0.000000
25%	191.000000	5.750000	3.500000	42.000000	0.900000	940.000000	-4.700000	0.000000	0.000000	0.000000	4.000000	8.000000	0.000000
50%	504.500000	11.500000	13.700000	57.000000	1.500000	1698.000000	5.100000	0.010000	0.000000	0.000000	7.000000	16.000000	0.000000
75%	1065.250000	17.250000	22.500000	74.000000	2.300000	2000.000000	14.800000	0.930000	0.000000	0.000000	10.000000	23.000000	1.000000
max	3556.000000	23.000000	39.400000	98.000000	7.400000	2000.000000	27.200000	3.520000	35.000000	8.800000	12.000000	31.000000	1.000000

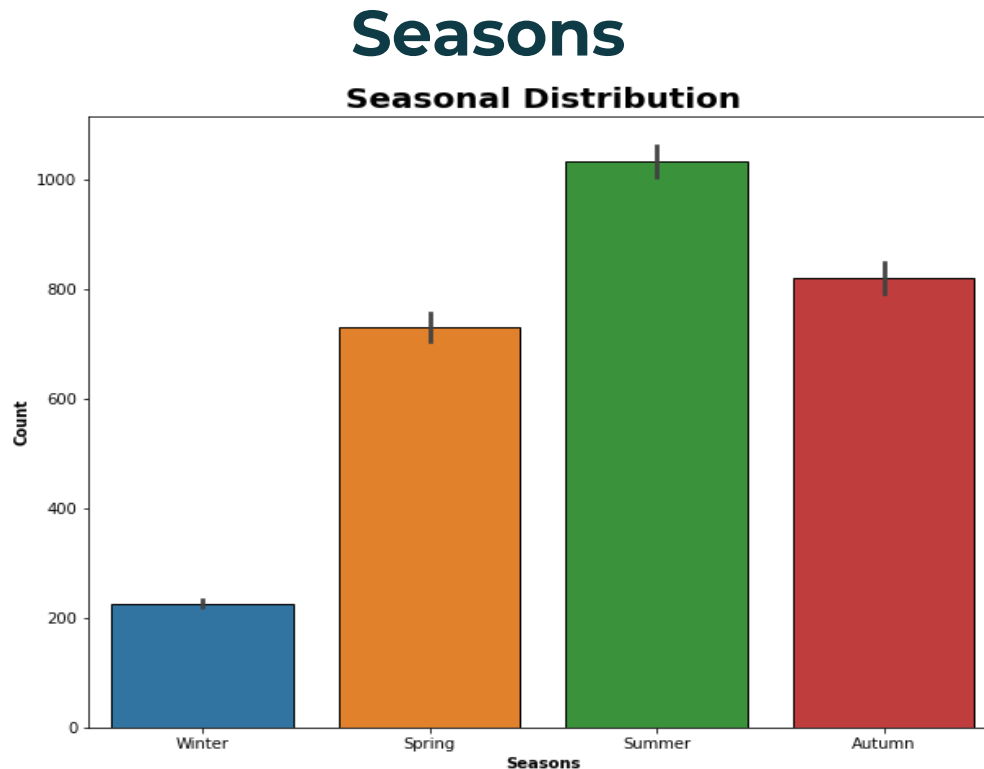
- We can observe that there were Rented bike count having 0 bike which means no bike demand.

Extracting the correlation of the dataset using Heatmap

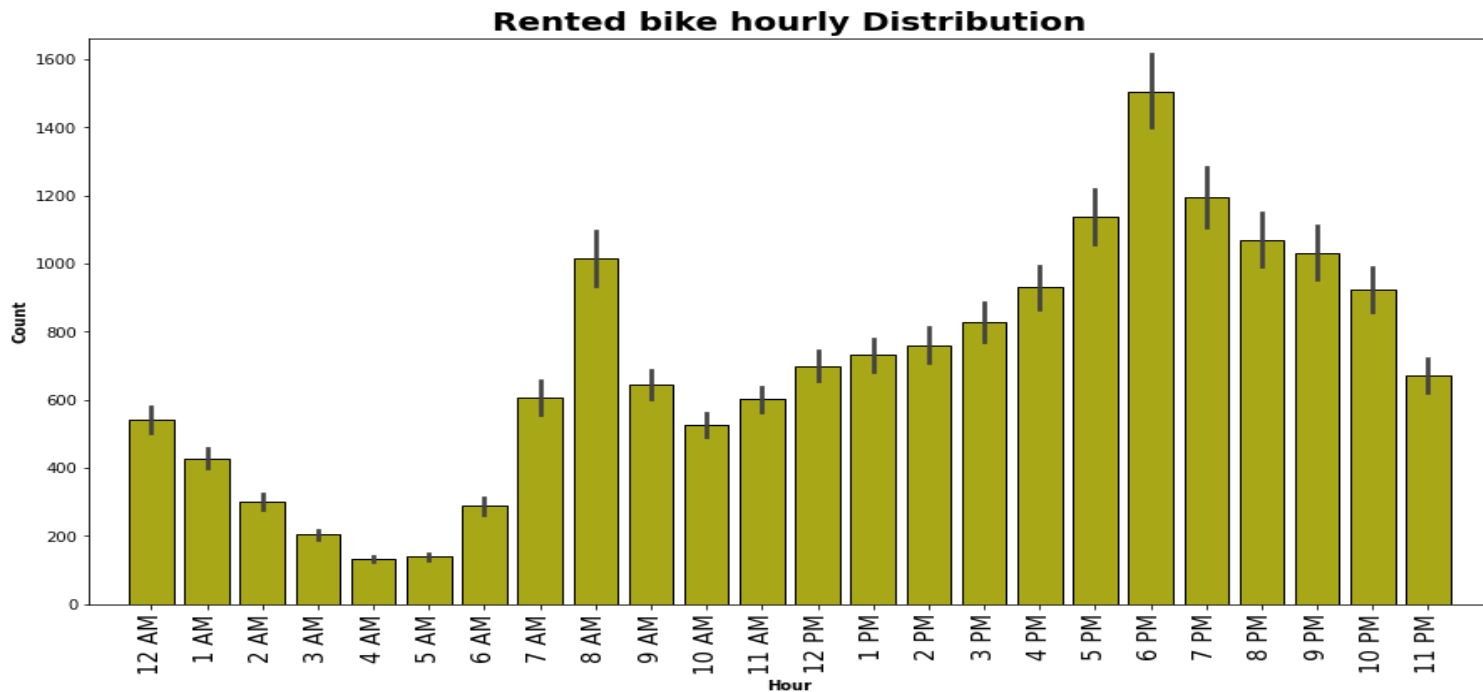




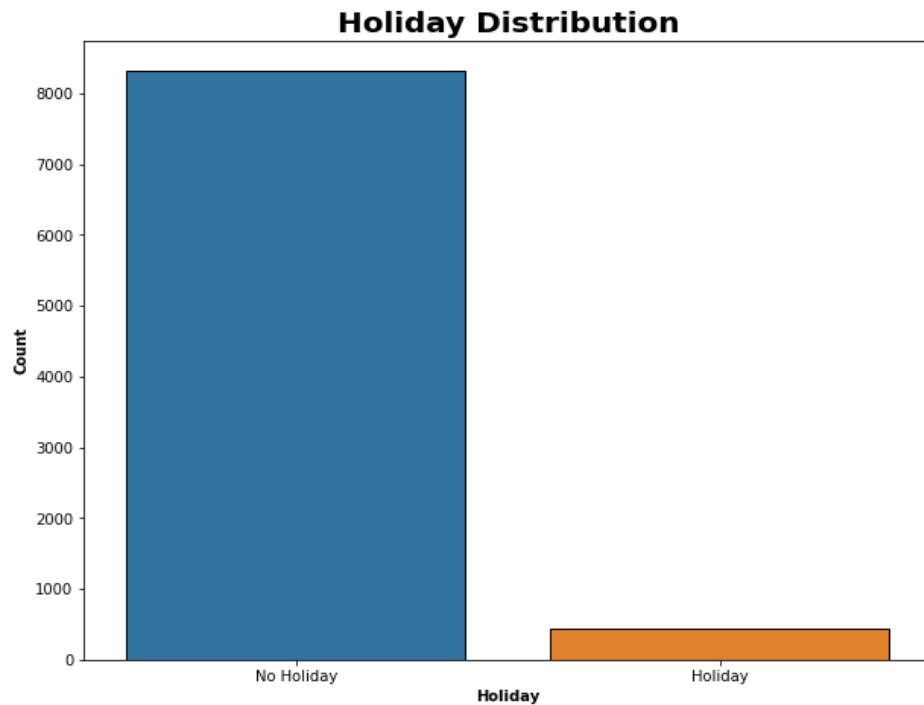
Exploratory **D**ata **A**nalysis



- From above graph in Summer season bike has more demand less demand in the winter season.

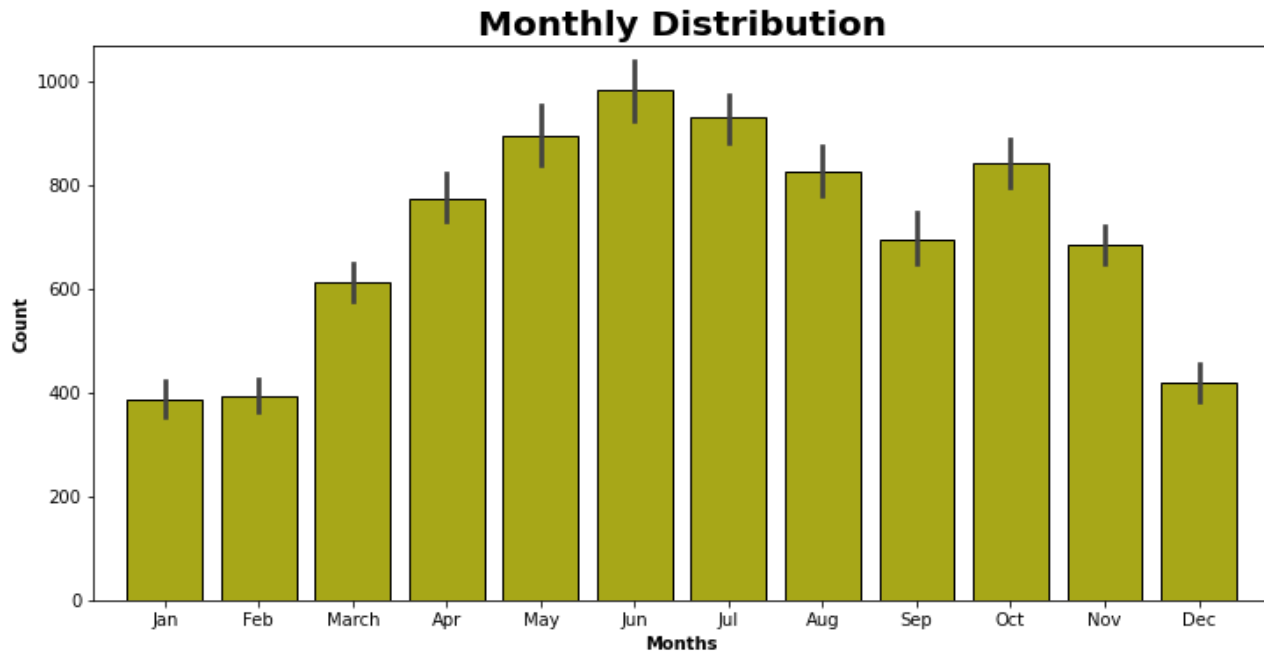


- High rise of Rented Bikes from 8:00 a.m. to 9:00 p.m. means people prefer rented bike during rush hour. we can clearly see that demand rises most at 8 a.m. and 6:00 p.m. so we can say that during office opening and closing time there is much high demand.

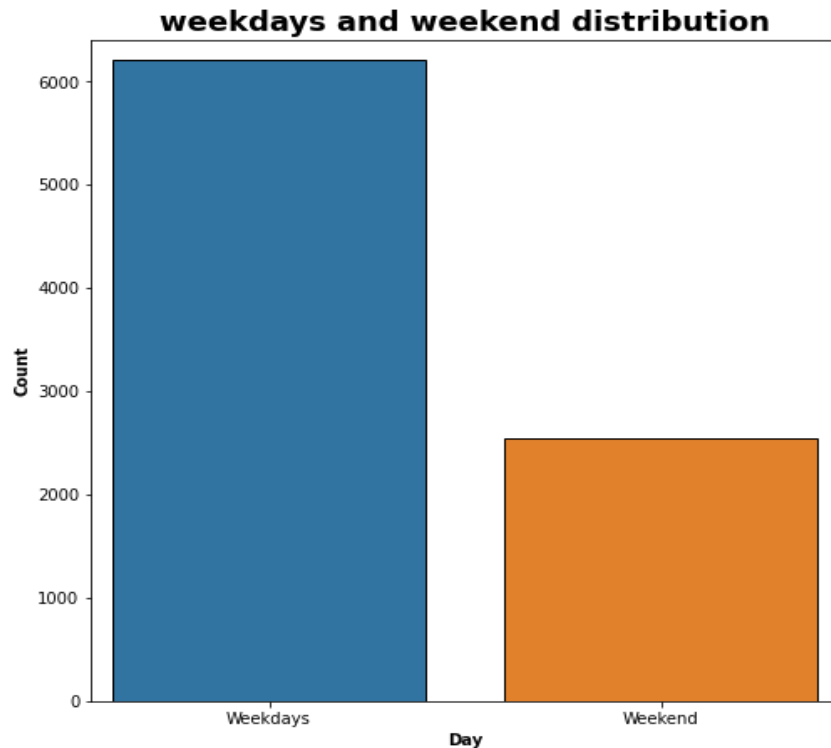


- **Bike demand is more on No-holiday as compared to holiday.**

Month

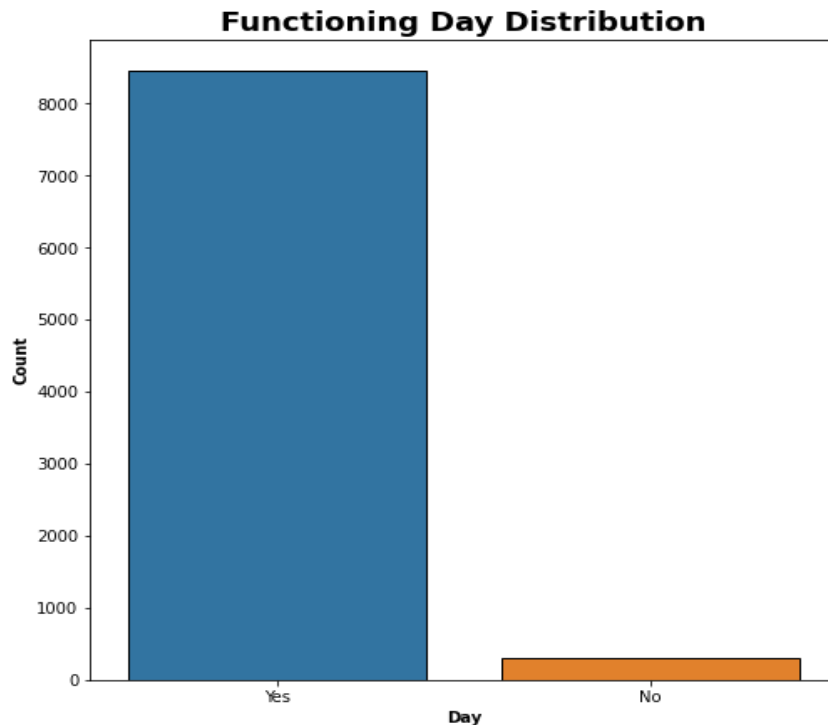


- we can see that there less demand of Rented bike in the month of December January, February i.e. during winter seasons Also demand of bikes is maximum during May, June, July i.e. Summer.



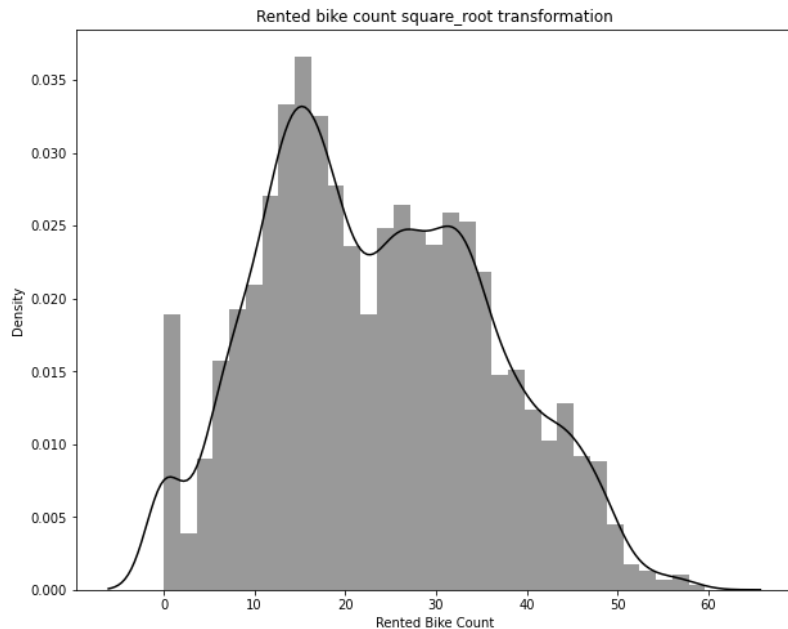
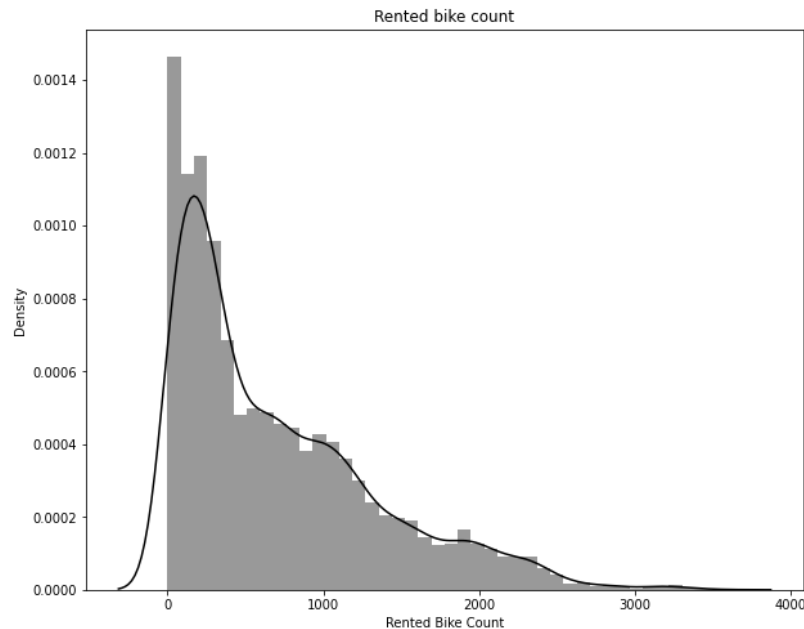
- Users demand more bike on weekdays as compared to weekend.

Functioning Day



- On Functioning Day Bike demand is more as compared to other days.

Square transformation



- Taking the square root and the logarithm of the observation in order to make the distribution normal belongs to a class of transforms called power transforms.

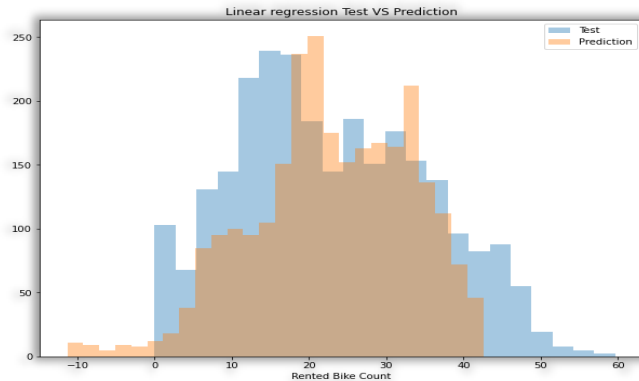
Machine Learning Model

AI

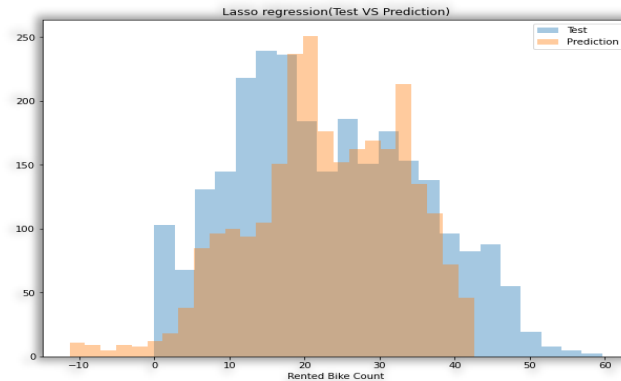


ML Model Prediction

1.Linear Regression

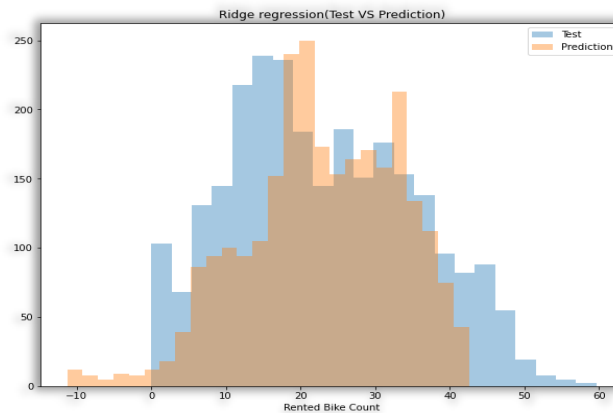


2.Lasso Regression



Lasso ($\alpha=0.0014$)

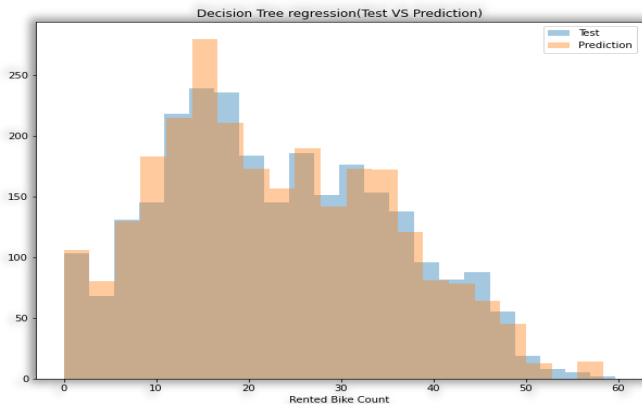
3.Ridge Regression



Ridge ($\alpha=5$)

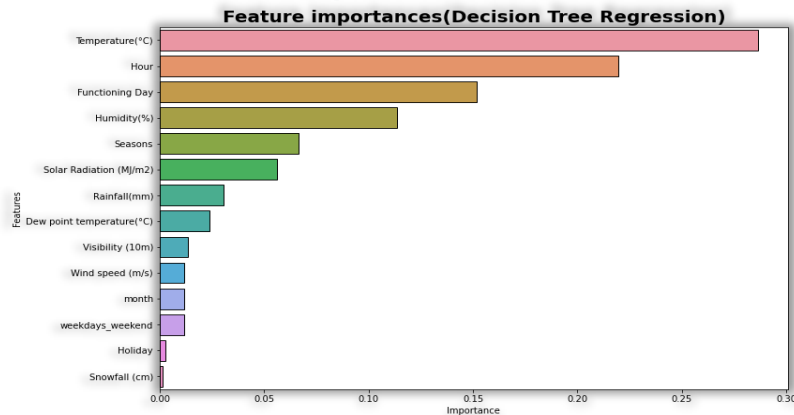
ML Model Prediction

4. Decision Tree Regression

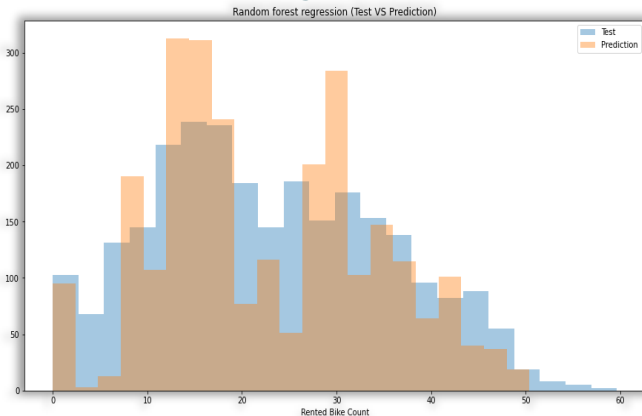


`DecisionTreeRegressor(max_depth=20,min_samples_split=0.1)`

➤ Feature importance's

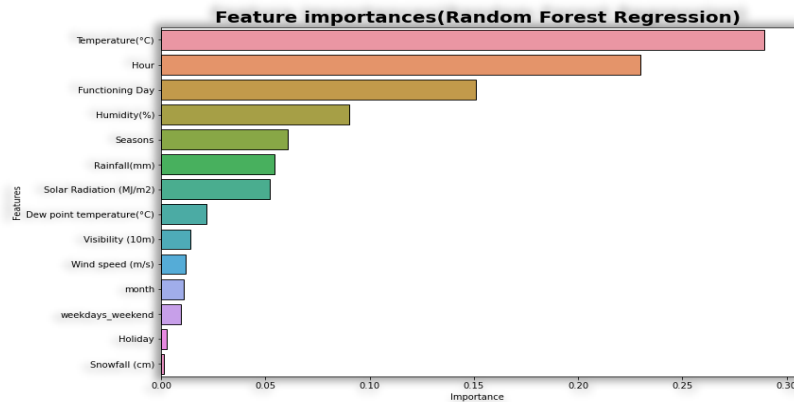


5. Random Forest Regression



`RandomForestRegressor(max_depth=20,max_leaf_nodes=8, n_estimators=60)`

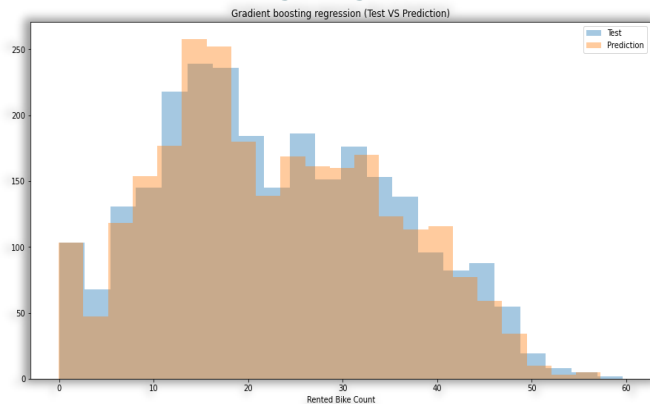
➤ Feature importance's



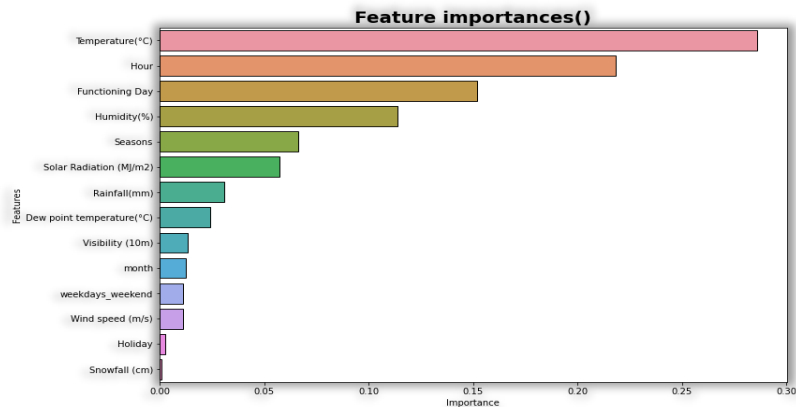
ML Model Prediction



6. Gradient boosting Regression

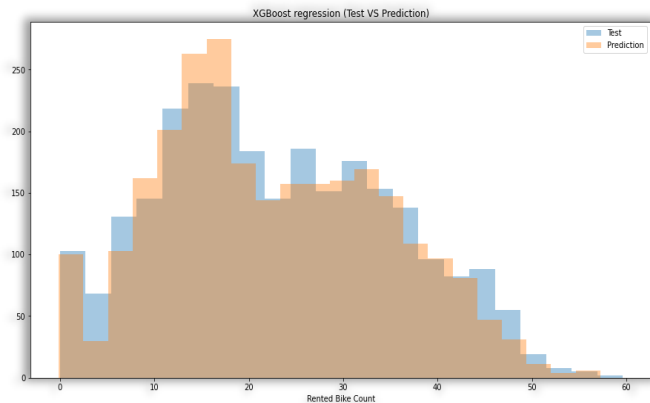


➤ Feature importance's

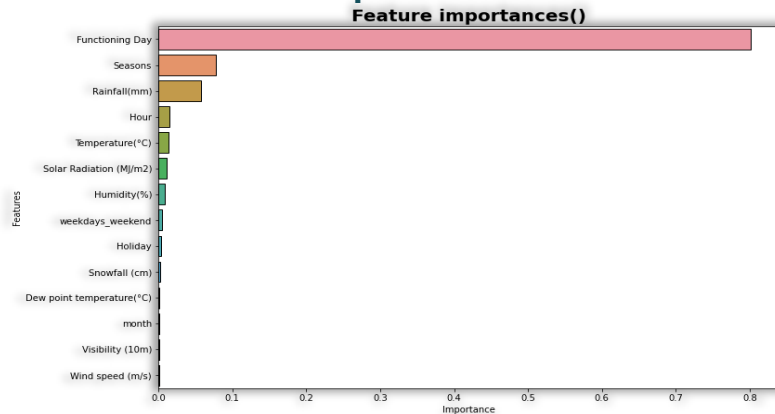


`GradientBoostingRegressor(max_depth=15, 'n_estimators':100)`

7. XGBoost Regression



➤ Feature importance's

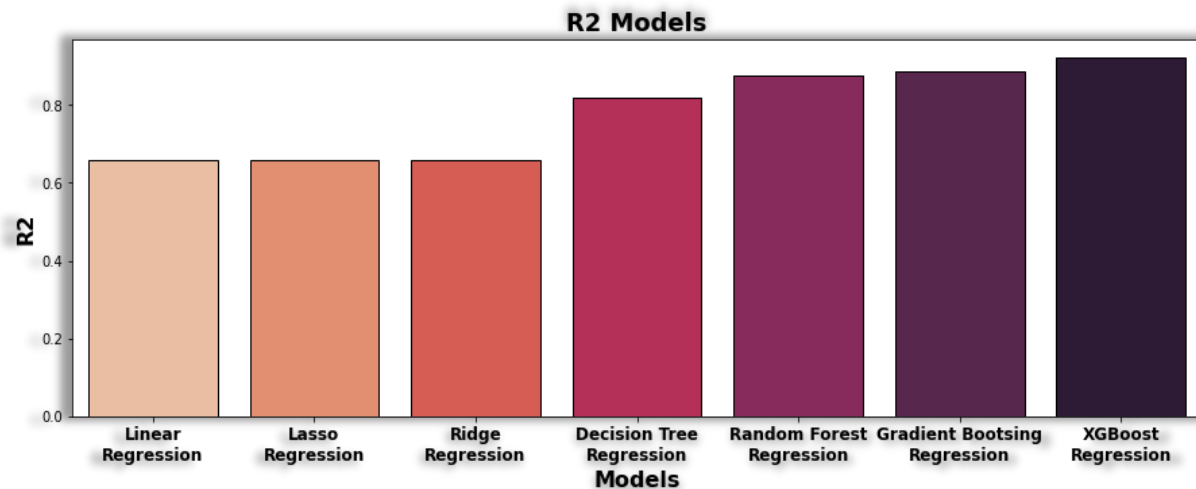


`XGBRegressor(max_depth=15, n_estimators=150)`

ML Model Evaluation Result



	Models	Mean_square_error	Root_Mean_square_error	R2	Adjusted_R2
0	Linear Regression	52.822839	7.267932	0.658480	0.656650
1	Lasso Regression	52.822824	7.267931	0.658480	0.656650
2	Ridge Regression	52.839987	7.269112	0.658369	0.656538
3	Decision Tree Regression	27.798881	5.272464	0.820269	0.819306
4	Random Forest Regression	17.788651	4.217659	0.884989	0.884373
5	Gradient Boosting Regression	19.166850	4.377996	0.876079	0.875415
6	XGBoost Regression	11.983092	3.461660	0.922525	0.922110



Conclusion



- As we can see the total amount of bike rentals increases with the temperature per month. Whereas it seems that the rentals are independent of the wind speed and the humidity, because they are almost constant over the months. This also confirms on the one hand the high correlation between rentals and temperature and on the other hand that nice weather could be a good predictor. So people mainly rent bikes on nice days and nice temperature. This could be important of planning new bike rental stations.
- It is quite evident from the results that XGBoost is the best model that can be used for the Bike Sharing Demand Prediction since the performance metrics (mse, rmse) shows lower and (r^2 , adjusted r^2 = 92 %) show a higher value for the XGBoost model. So, finally this model is best for predicting the bike rental count on daily basis.

References

1. <https://www.kaggle.com/>
2. <https://www.analyticsvidhya.com/>
3. <https://www.geeksforgeeks.org/>
4. <https://learn.almabetter.com/>