Capstone Project - 2

Bike Sharing Demand Prediction

Supervised ML Regression Model

Shubham bhadouria

Concept of Bike Sharing

Bike-Sharing, as a new green public transportation mode, has been developed in several western cities and most of the developing countries are on the path of following the western model of bike sharing system.

It is a convenient, environmentally friendly way to get around town, but has flaws too. There are certain conditions that must be met for proper implementation of this program. Bike sharing Demand is affected by multiple factors including temperature, resources and many more.

Let’s predict Bike Sharing Demand in Seoul

1. 1. Defining problem statement
2. 2. EDA and feature Engineering
3. 3. Data Summary
4. 4. Feature selection
5. 5. Preparing Dataset for modelling
6. 6. Implementing Regression models
7. 7. Challenges
8. 8. Conclusions

Problem statements

- Prediction of demand of bike at each hour.
- What factors affect bike sharing count ?
- Reduce waiting time of public.

Need of machine learning to predict bike demand:

The idea of this project is to create a predictive model that identifies upcoming trends in bike sharing demand.

It is crucial to keep in mind that machine learning can only be used to memorize patterns that are present in the training data, so we can only recognize what we have seen before. When using Machine Learning we are making the assumption that the future will behave like the past, and this isn't always true.

Data pipeline

- *Preparing the data_1* : In this first part, we've done data inspection where we checked null values and duplicate observations and did multiple operations to make sure our dataset is up to the mark.

- *Preparing the data_2* : Checked all the features, extracted data feature to get more insight of the data. Now as dataset is ready, we moved to the next step.

- *EDA* : In this part, in order to see the trends we did some exploratory data analysis on the features and checked distribution of data points and correlation of variables.

- *Create model* : After Data preparation, We build the different supervised machine learning regression models.
- *Choose a Measure of Success* : After implementing each model, we measure it's accuracy by some evaluation matrices.

Data summary

- Date - year-month-day
- Rented Bike count - Count of bikes rented at each hour
- Hour - Hour of he day
- Temperature-Temperature in Celsius
- Humidity - %
- Wind speed - 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 - Yes(Functional Day), No(Non-Functional day)

Basic Data Exploration

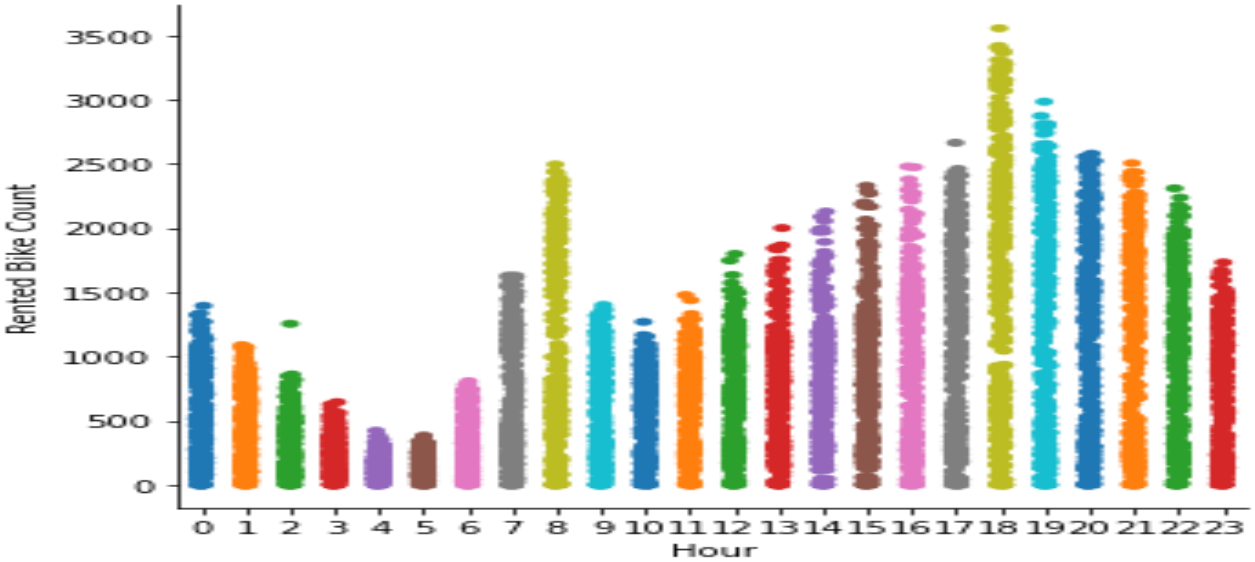
- This Dataset has 8760 Row and 14 Columns.
- The dataset has the information of 365 days(1/dec/2017 to 30/Nov/2018)
- Dataset contains no null values
- No duplicates values.
- There are 3 categorical features ‘Seasons’, ‘Holiday’, & ‘functioning Day’.
- From date column, we extract lots of features like year, month and weekdays.

```
bike_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

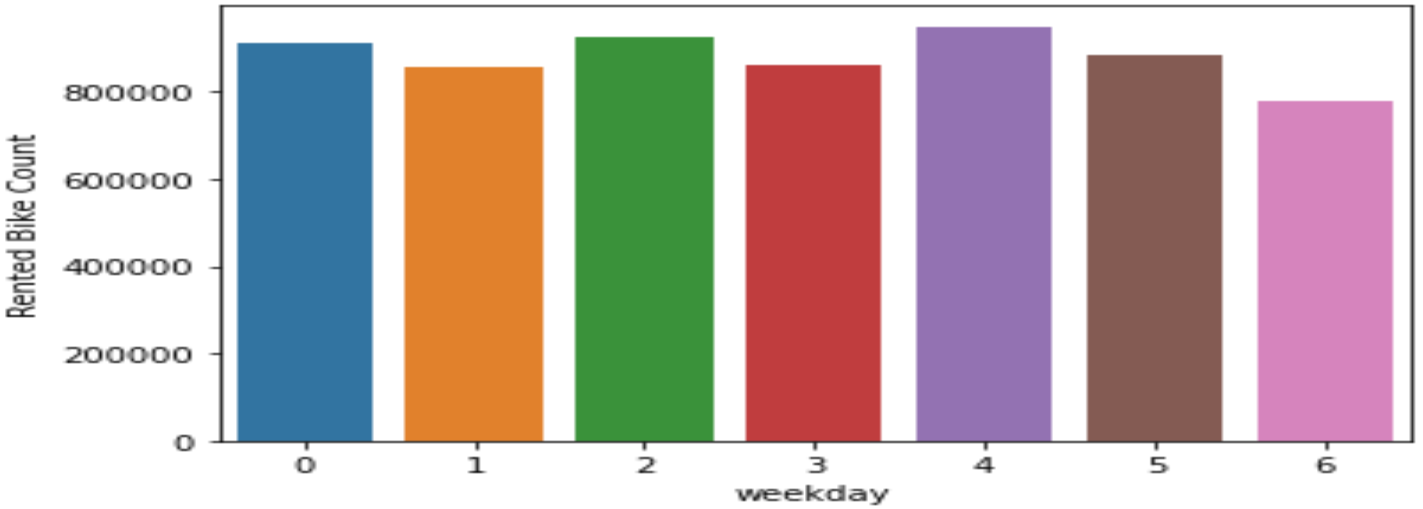
EDA

Rented bike count on various hours of the day



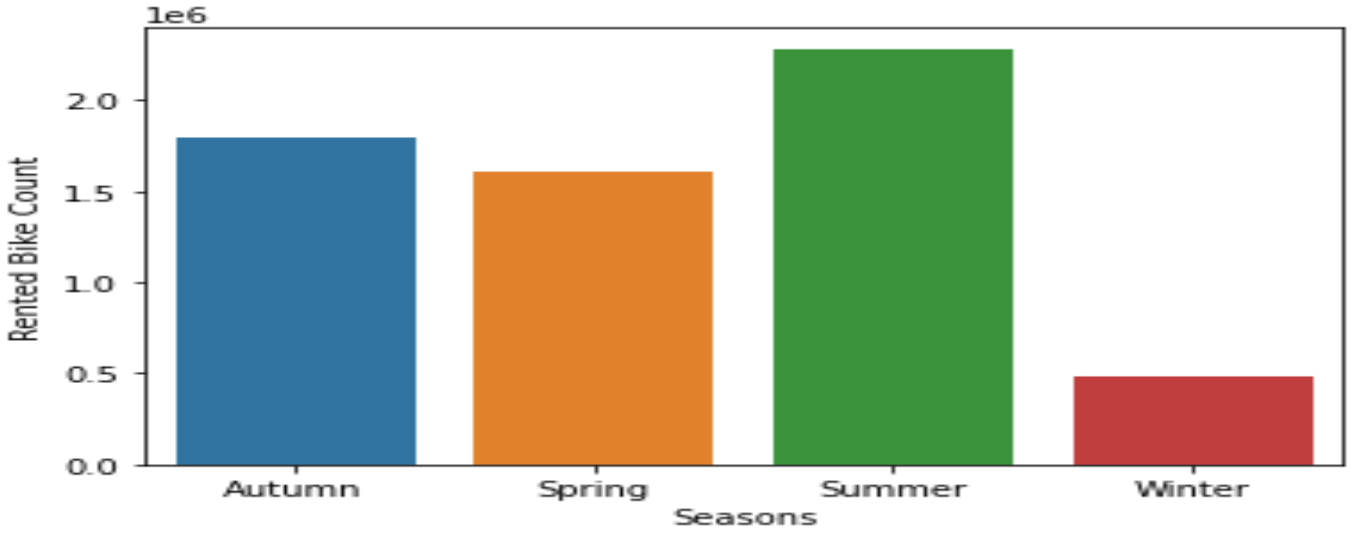
- Demand of bike is higher at 8am and max at 6pm. i.e. during opening and closing hours of office.

EDA - Bike count on various days of week



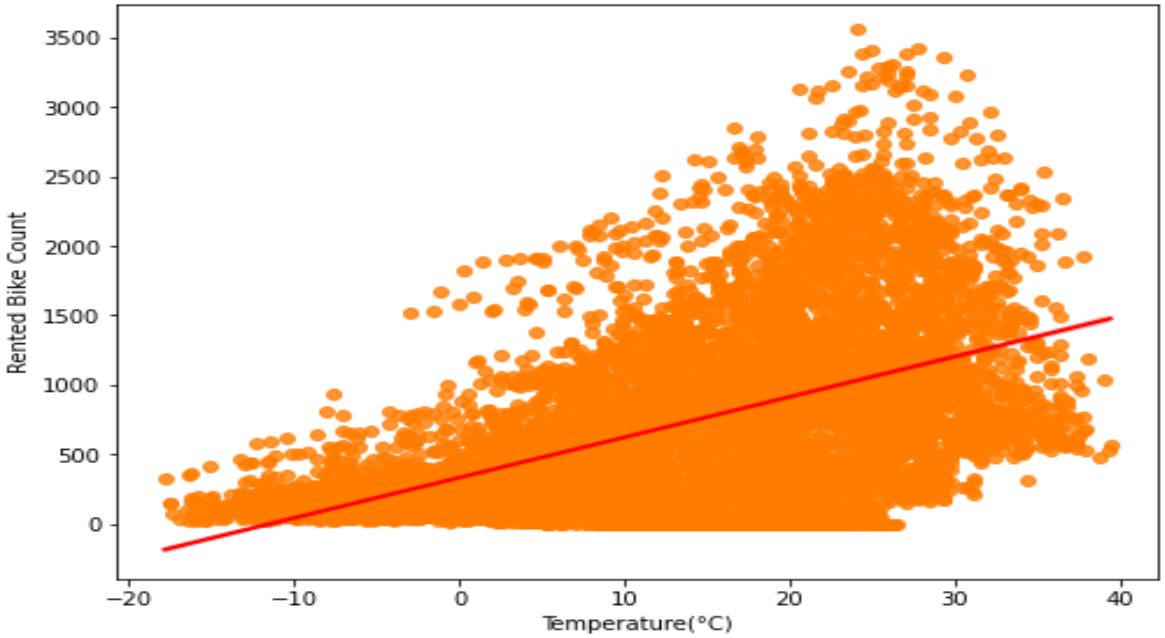
There is no significant difference in bike rented on various days of the weeks, but minimum bike rented on Sunday.

EDA – Total count of rented bike in various seasons of the year



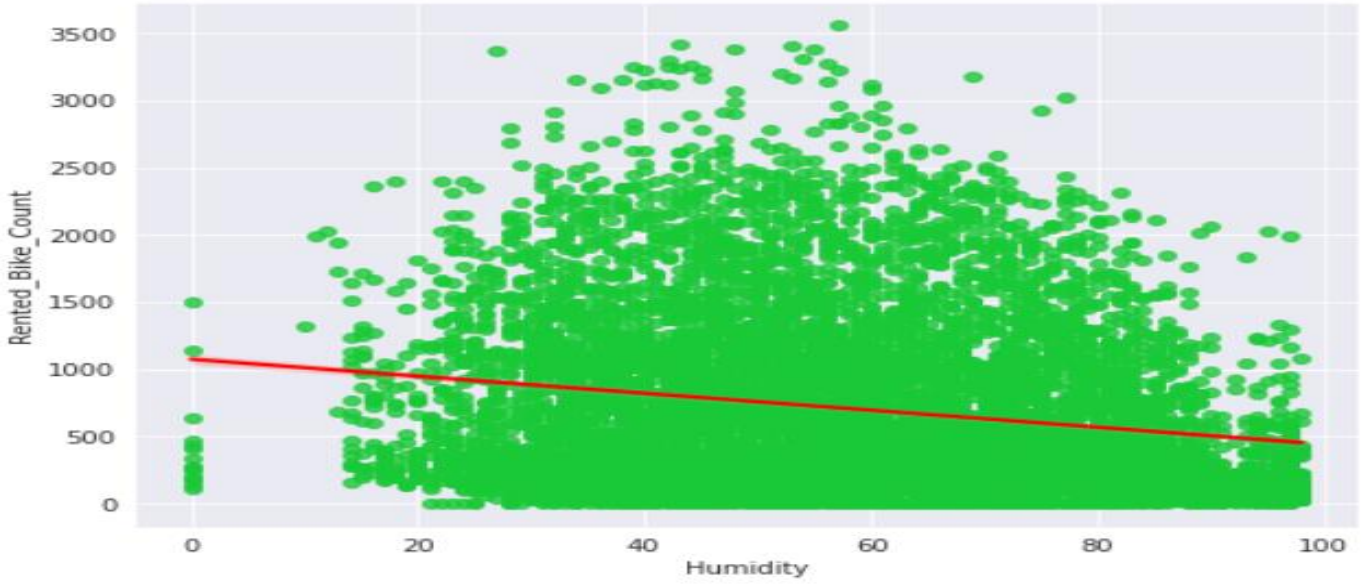
Rented bike count is less in winter and almost consistently higher in other months.

EDA - Relationship between bike count and Temperature



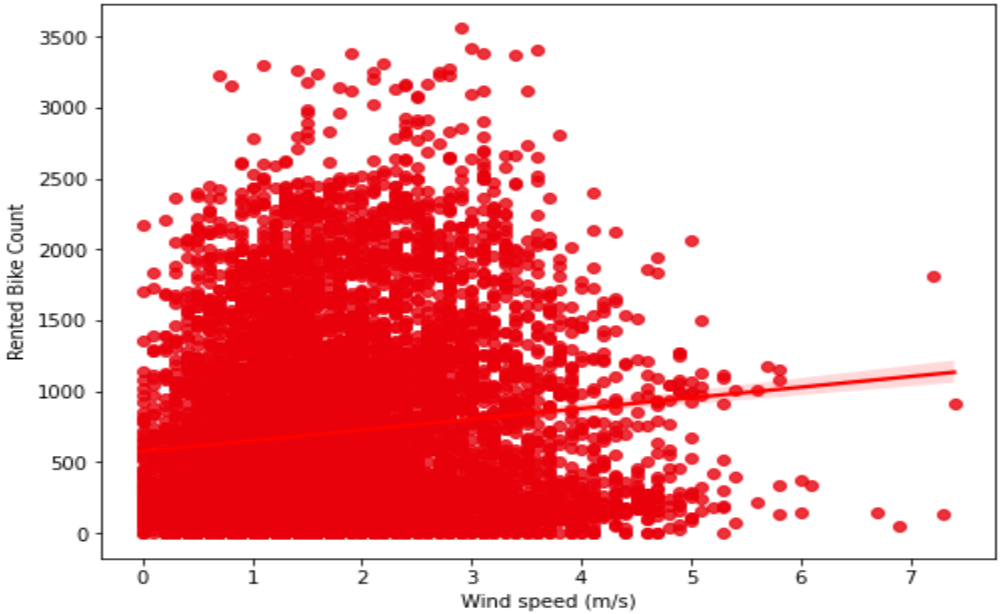
Temperature : Temperature is positively correlated. Rented bike count is highest between 20 °C and 30 °C.

EDA - Relationship between bike count and Humidity



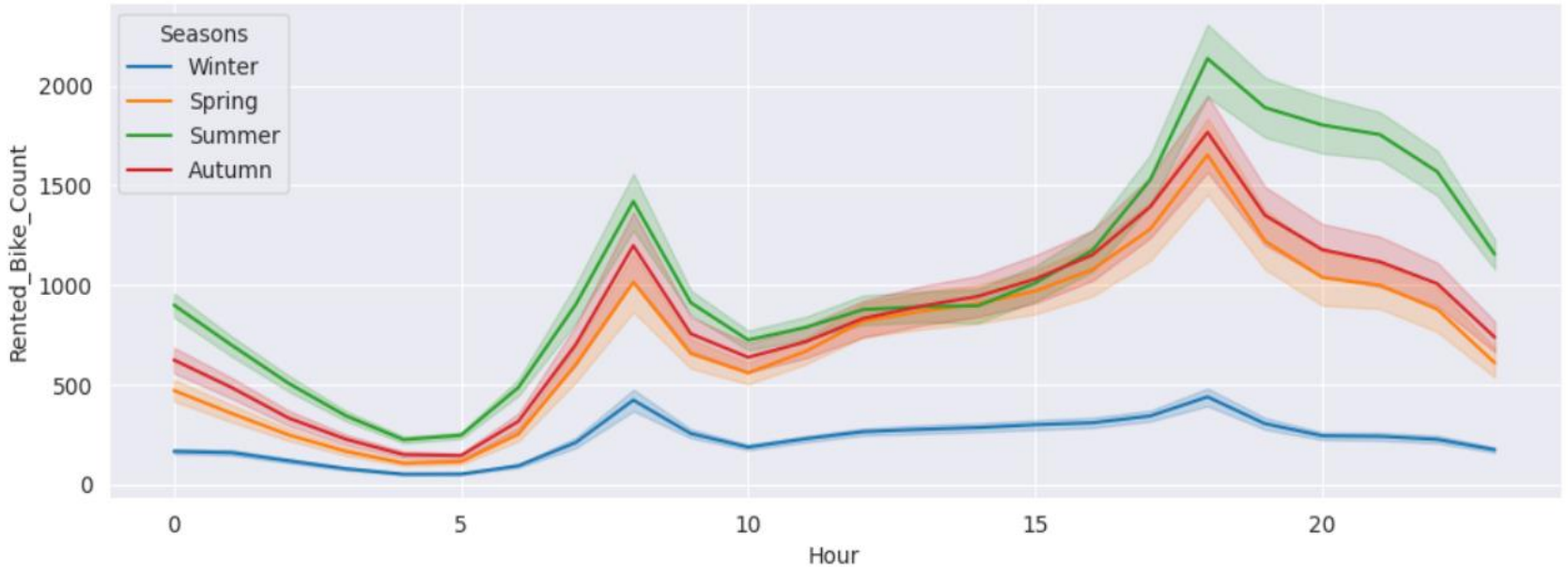
Humidity : Humidity is the amount of water vapor in the air. So, People preferring to borrow bike When there is optimum humidity.(neither too high nor too low.

EDA - Relationship between bike count and Windspeed



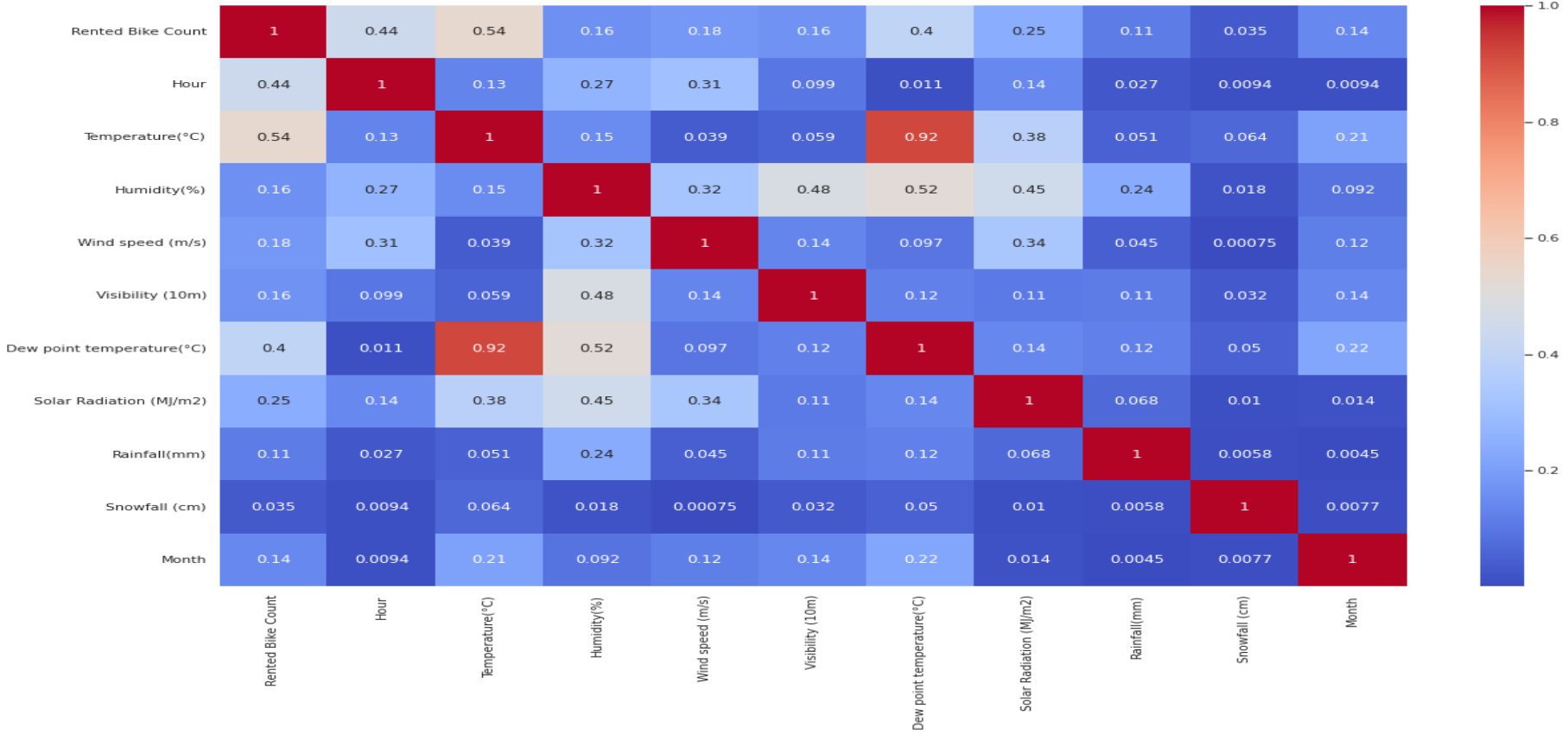
Windspeed : Consumers prefer bikes when wind speed is in particular range but looking at plot, we can conclude that wind speed doesn't affect our data much.

EDA - Rented bike count in different seasons



People prefer borrowing bikes more in a particular season. Rented bike count is highest in summer and least in winter. Peak demand of bike is at 8am and 6pm i.e. during office opening and closing times

Correlation between Different variables by using Heatmap



Correlation between Different variables by using Heatmap

- Multicollinearity : Dew point temperature, Temperature and Humidity are variables which are highly correlated with each other. We find VIF value of each independent variable . After dropping dew point temperature column, VIF values of each independent variable are under 10.

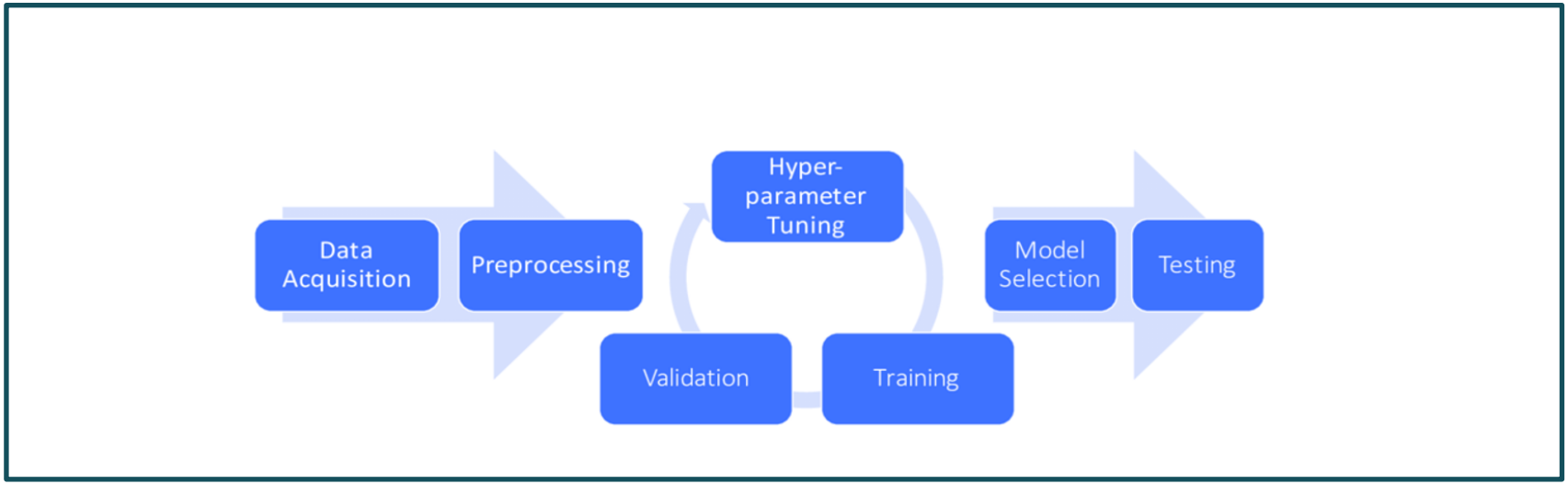
	variables	VIF
0	Hour	3.931253
1	Temperature(°C)	3.470934
2	Humidity(%)	6.284281
3	Wind speed (m/s)	5.119235
4	Visibility (10m)	5.809378
5	Solar Radiation (MJ/m2)	2.244557
6	Rainfall(mm)	1.070191
7	Snowfall (cm)	1.008483
8	Month	5.289116

Preparing Dataset for Modelling

- Dataset:

  Train set: (5385, 12)

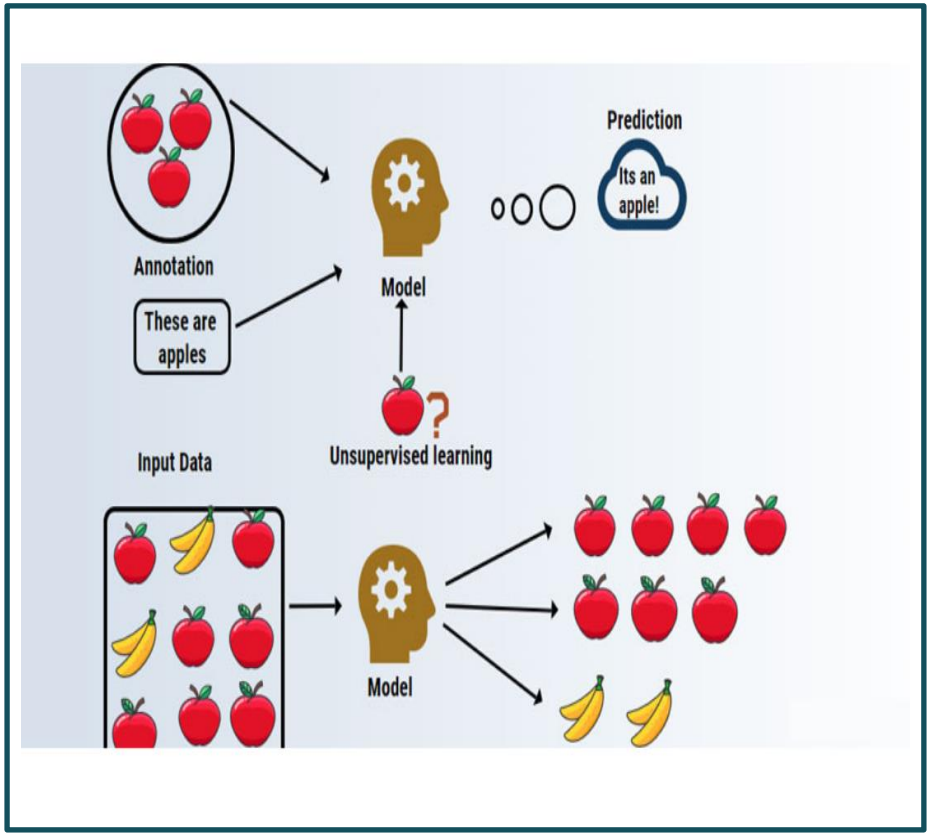
  Test set: (2308, 12)
- Carefully handled feature selection part as it affects R2 score.



Modeling Overview

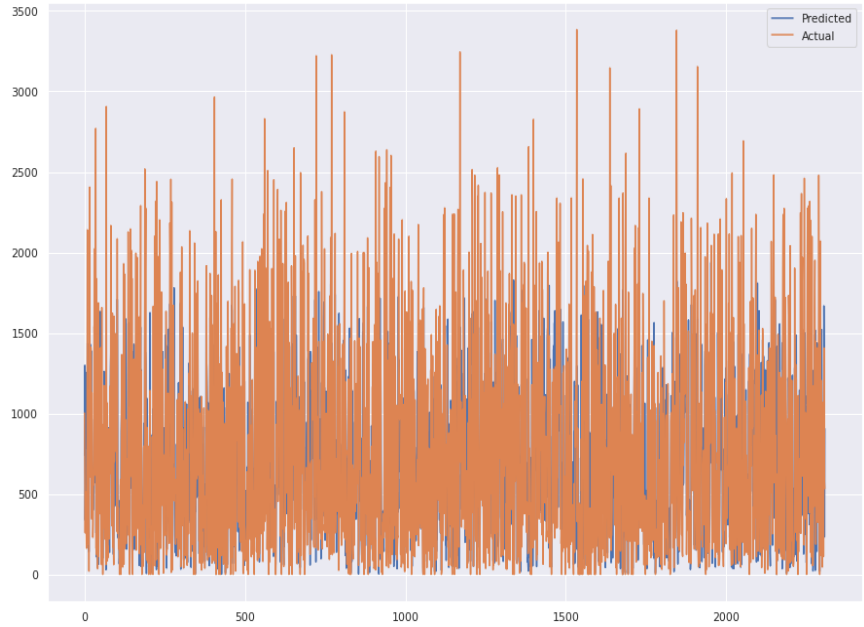
- Type – Supervised Learning
- In this project, we are using six model on our data set for getting best performance :

1. LINEAR REGRESSION
2. LASSO REGRESSION
3. RIDGE REGRESSION
4. DECISION TREE
5. RANDOM FOREST
6. XGBOOST



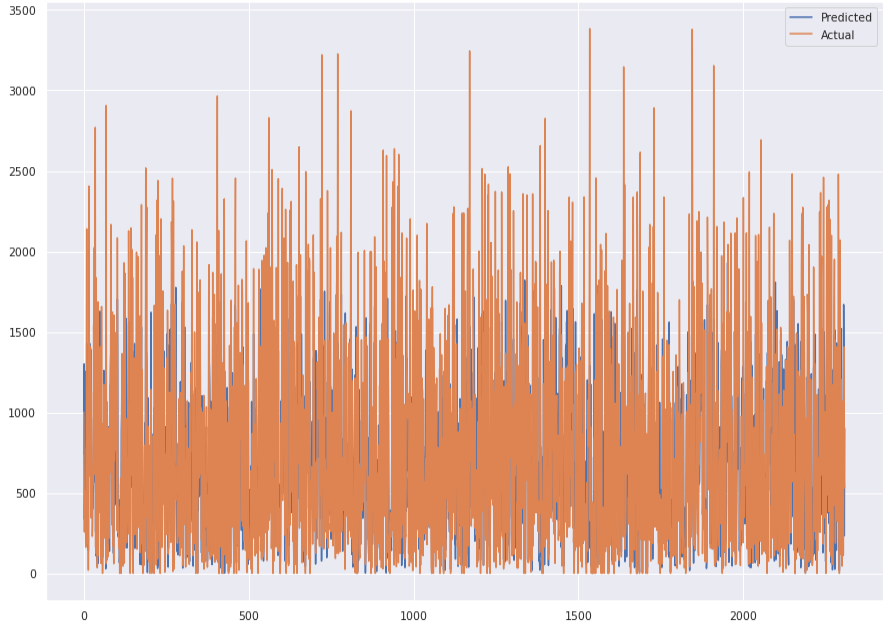
Linear regression

```
Test MSE : 52.40539224141344
Test R2 : 0.5763625324127122
Test Adjusted R2 :  0.5741474345429747
```



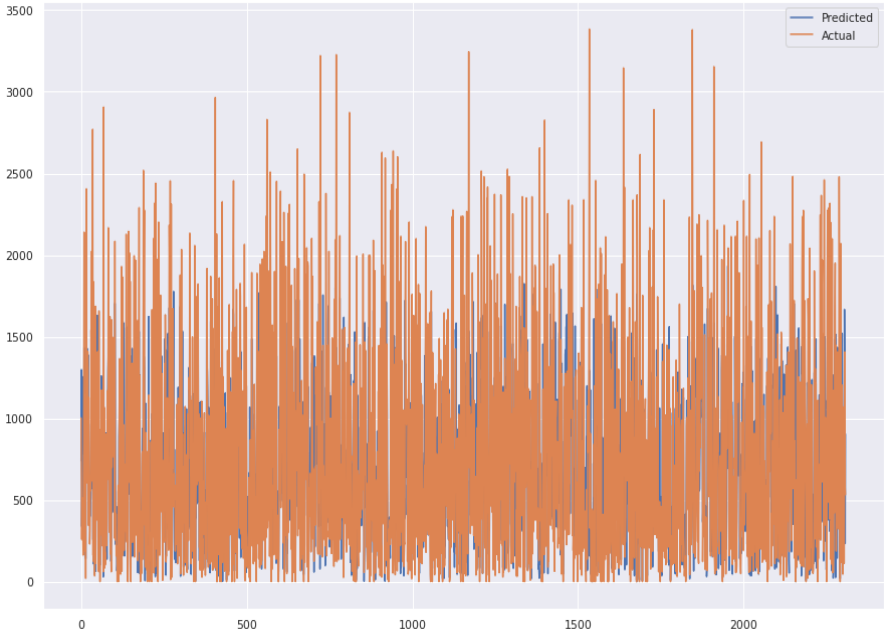
Lasso regression

```
Test MSE : 928867.9691859926
Test R2 : 0.5757889541238373
Test Adjusted R2 :  0.5735708571519358
```



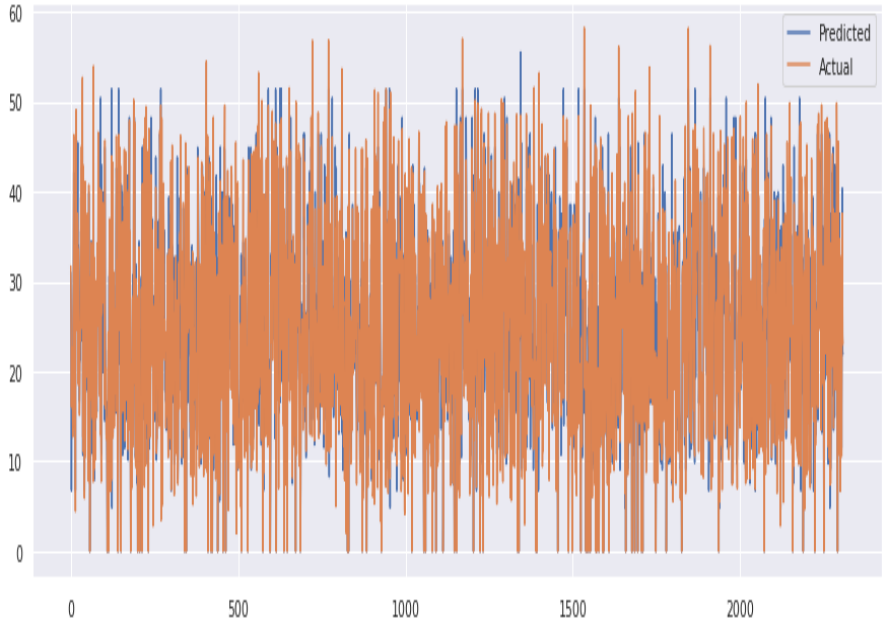
Ridge regression

```
Test MSE : 184484.89482933382
Test R2 : 0.576023102475957
Test Adjusted R2 :  0.5738062298091646
```



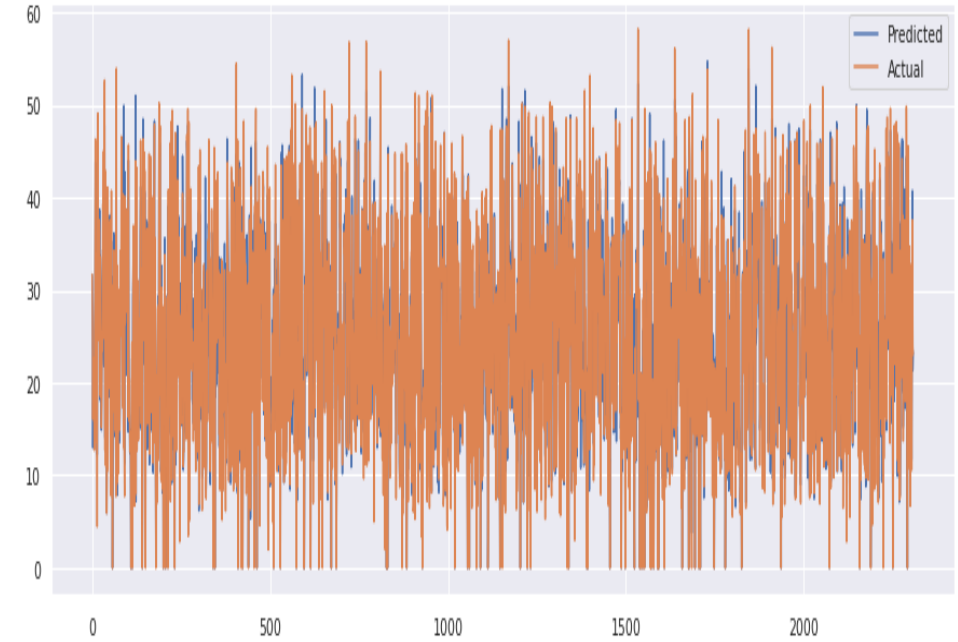
Decision Tree

```
Test MSE : 76743.02248045518
Test R2 : 0.823631801357063
Test Adjusted R2 :  0.8227096146974922
```



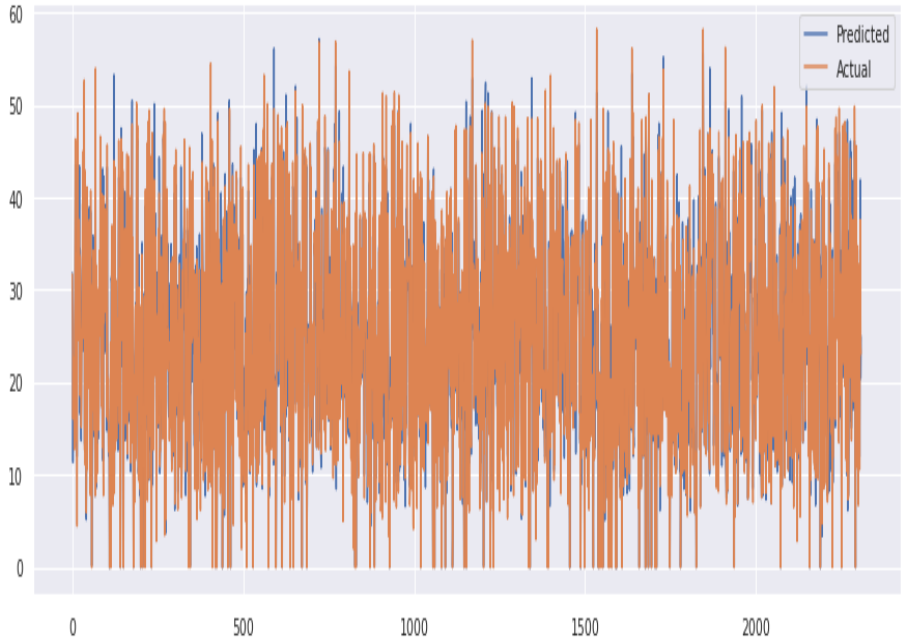
Random forest

```
Test MSE : 58033.498941468875
Test R2 : 0.8666293906803012
Test Adjusted R2 :  0.8659320280171916
```



XGboost

```
Test MSE : 16.277046359443275
Test R2 : 0.8946367091154187
Test Adjusted R2 :  0.8940857899473947
```



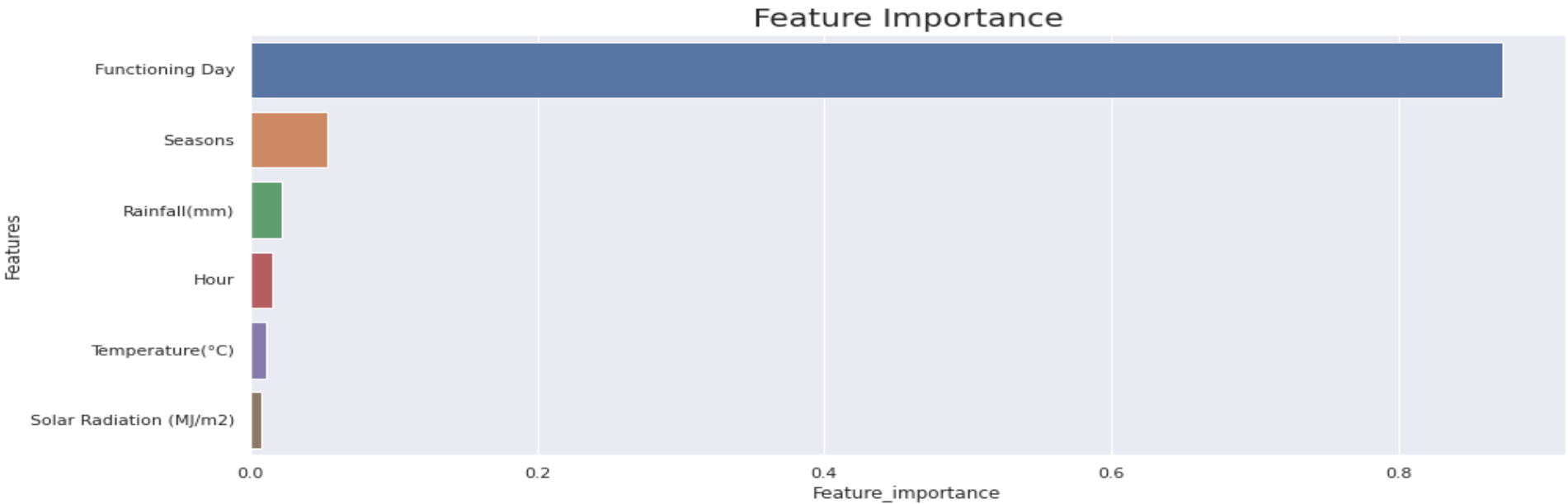
Comparison of different models on the basis of various evaluation matrices

	Model Name	Train MSE	Test MSE	Train R^2	Test R^2	Train Adjusted R^2	Test Adjusted R^2
0	Linear Regression	176191.029574	52.405392	0.583305	0.576363	0.582375	0.574147
1	Lasso Regression	176488.594182	928867.969186	0.582602	0.575789	0.581669	0.573571
2	Ridge Regression	176363.451769	184484.894829	0.582898	0.576023	0.581966	0.573806
3	DecisionTree Regressor	52168.403781	76743.022480	0.876621	0.823632	0.876345	0.822710
4	Random Forest	24142.557298	58033.498941	0.942902	0.866629	0.942775	0.865932
5	XGBoost Regressor	2.641893	16.277046	0.953255	0.894637	0.953151	0.894086

XGBoost outperforms all the other models with r square score of 0.894637.

Feature importance with XGBoost model Grid Search-cv

Functional day is the most relevant feature here.



Challenges

- Large dataset to handle
- Feature engineering
- Feature selection - Making sure we don’t miss any important feature.
- Careful tuning of hyperparameters as it affects R2 score.
- Computation time

Conclusion

- We implemented eight M.L. models. After comparing r square and adjusted r square values of all the models, we found that XGBoost has the highest r square values. Also it has least MSE among all models. So, we can conclude that XGBoost is the best model to predict rented bike count.
- People prefer borrowing bikes more summer season. Rented bike count is highest in summer and least in winter. Peak demand of bike is at 8am and 6pm i.e. during office opening and closing time.
- People prefers to borrow a bike during non- holidays and functional days of the week.
- Our model is ready for deployment.


THANK
YOU
24Slides