Capstone Project-2
Bike Sharing Demand Prediction

<u>Member</u>

Sourabh Pramanik

Contents

Problem Statement

Here in this project we have one dataset-

1. SeoulBikeData

Currently the rental bikes are introduced in many urban cities for enhancement of mobility comfort. It is important to make the rental bikes available to public at right time as it lessen the waiting time.

In this project we will try to understand the Bike demand with respect to different timings of the day, months, seasons.

Data Summary

We will complete this project by using following steps-

- After reading the data we will perform Exploratory Data analysis.
- We will check the Null values and Outliers present in our Dataset.
- We will do some statistical analysis of the data.
- We will check the distribution of all the numerical columns and correlation between the variables.
- After then we will apply different Machine Learning models and will check the performance of the Models by using some performance metrics.

Data Summary

Outcomes of this Project -

- Factors which affect the Bike demand.
- Selection of appropriate model to predict the demand.
- Average count of bikes needed w.r.t every ‘Hour’, ‘Month’, ‘Seasons’.

Independent Variables

Some important Independent variables -

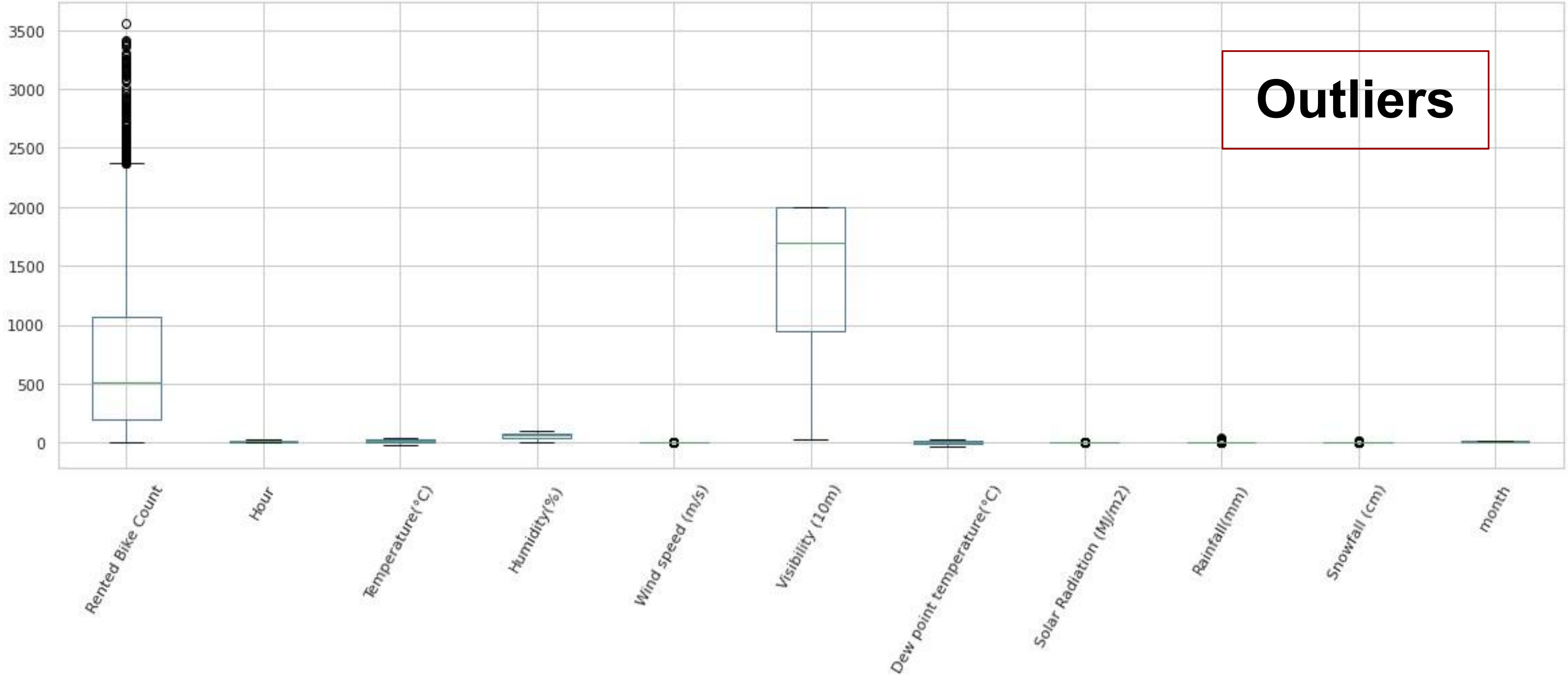
Hours of the day, Temperature, Humidity, Wind Speed, Rainfall, Snowfall, Holiday, Functional Day.

Dependent Variables

Dependent variable-

Rented Bike Count

Checking Missing values and Outliers



Outliers

EDA and Visualization

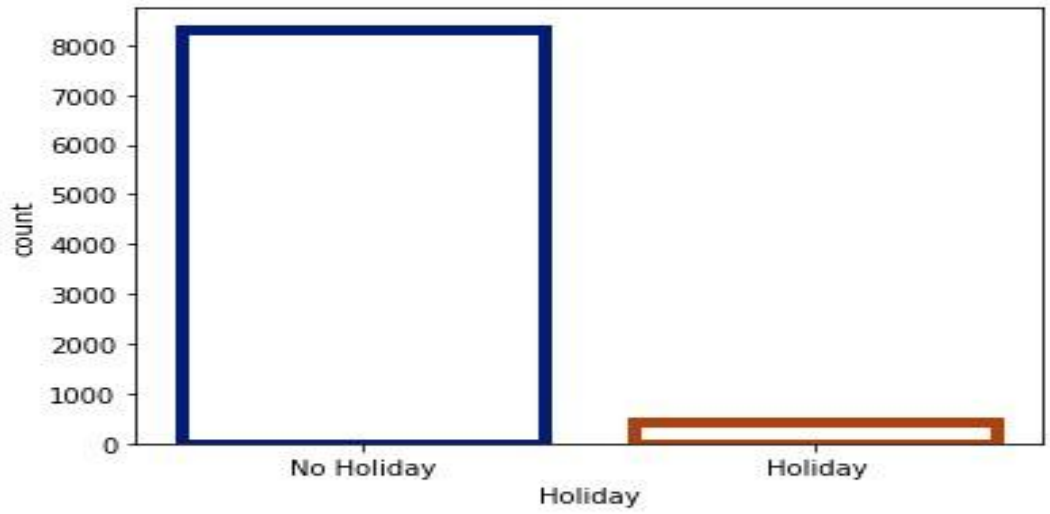
Statistical Distribution

```
df_copy.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	No Hol
count	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.00
mean	704.602055	11.500000	12.882922	58.226256	1.724909	1436.825799	4.073813	0.569111	0.148687	0.075068	6.526027	0.95
std	644.997468	6.922582	11.944825	20.362413	1.036300	608.298712	13.060369	0.868746	1.128193	0.436746	3.448048	0.21
min	0.000000	0.000000	-17.800000	0.000000	0.000000	27.000000	-30.600000	0.000000	0.000000	0.000000	1.000000	0.00
25%	191.000000	5.750000	3.500000	42.000000	0.900000	940.000000	-4.700000	0.000000	0.000000	0.000000	4.000000	1.00
50%	504.500000	11.500000	13.700000	57.000000	1.500000	1698.000000	5.100000	0.010000	0.000000	0.000000	7.000000	1.00
75%	1065.250000	17.250000	22.500000	74.000000	2.300000	2000.000000	14.800000	0.930000	0.000000	0.000000	10.000000	1.00
max	3556.000000	23.000000	39.400000	98.000000	7.400000	2000.000000	27.200000	3.520000	35.000000	8.800000	12.000000	1.00

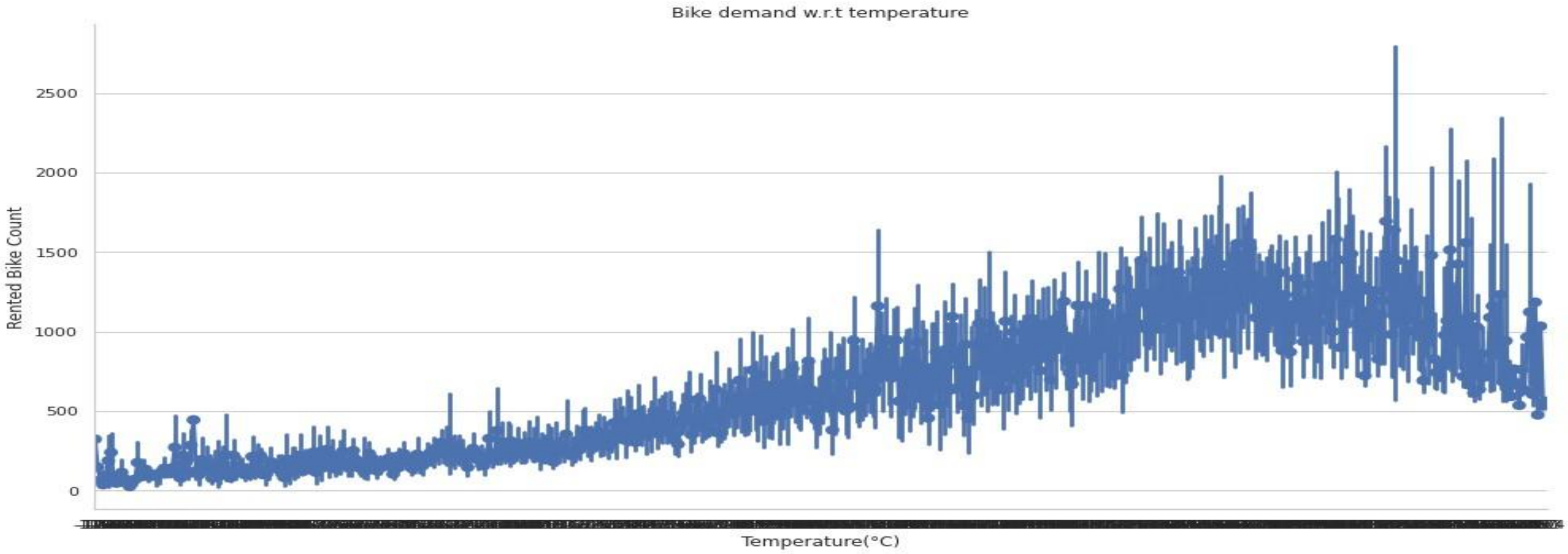
EDA and Visualization

Let’s see the total count of bikes w.r.t ‘Season’, ‘Days of week’, ‘Holidays’ -

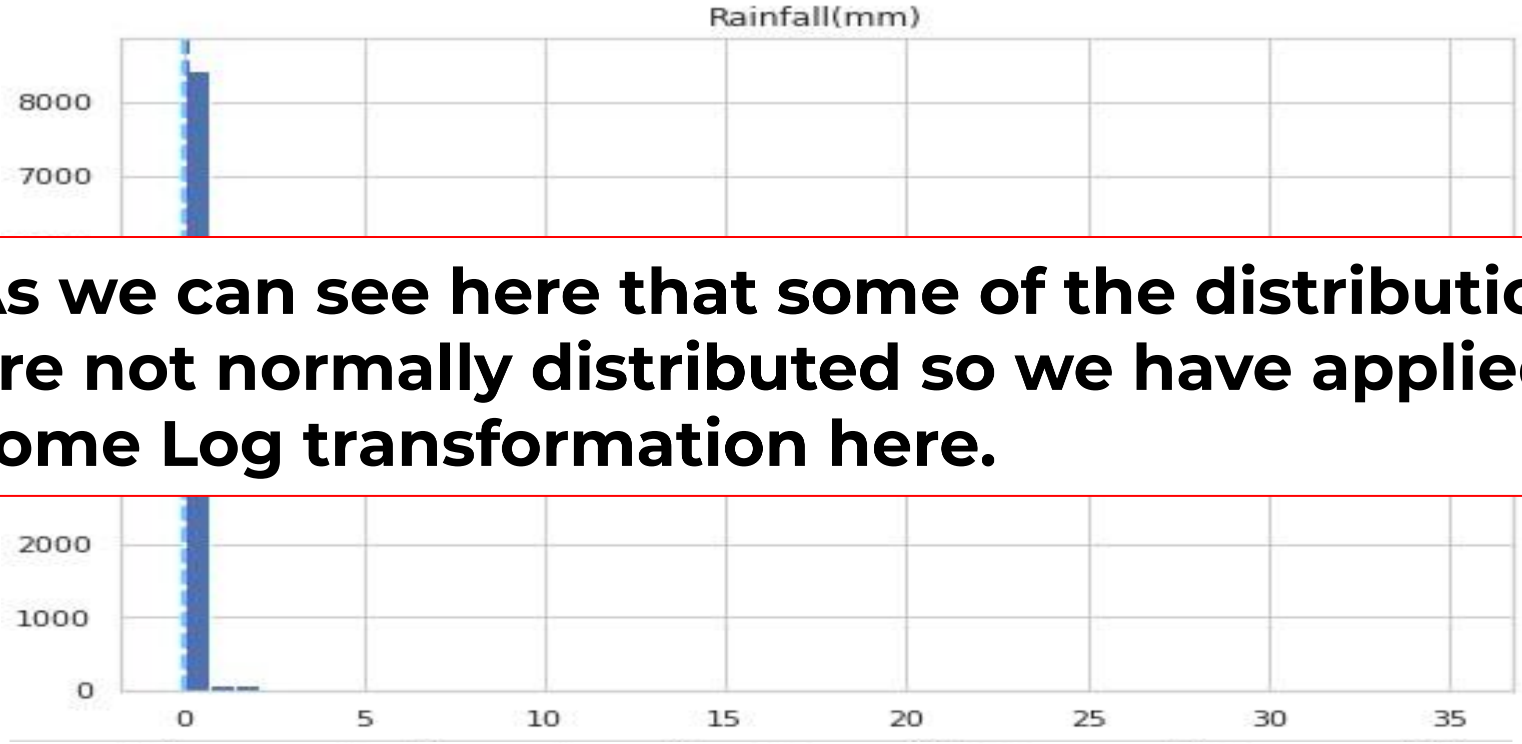


EDA and Visualization

Now we will see the Bike demand w.r.t ‘Every hour in a day’, ‘Month’, ‘Season’,’ Temperature’-

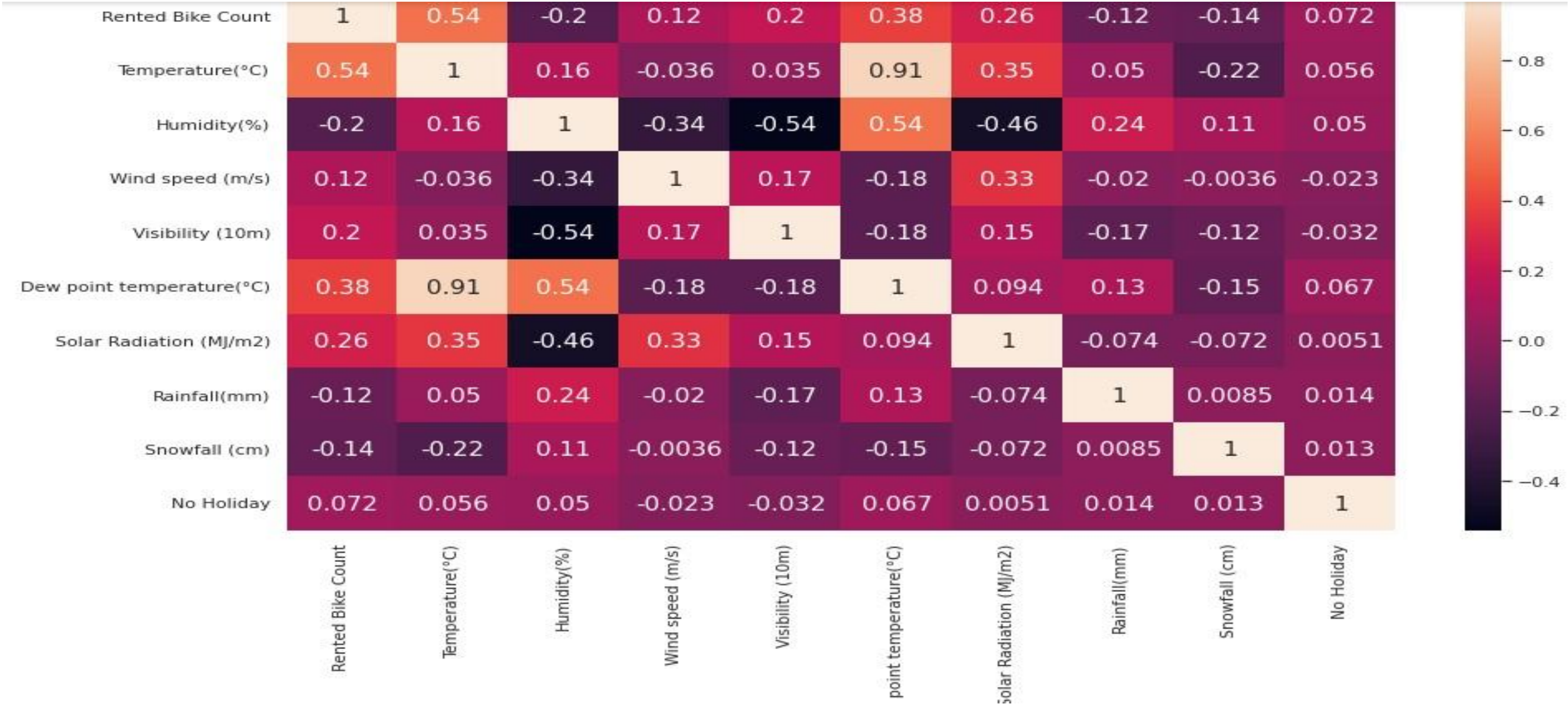


EDA and Visualization



As we can see here that some of the distribution are not normally distributed so we have applied some Log transformation here.

EDA and Visualization



Test and Train Split

```
# Independent variables

X = final_df.drop(['Rented Bike Count','Bike Count','Dew point temperature(°C)','Wind speed (m/s)','Solar Radiation (MJ/m2)','Visibility (10m)'], axis=1)
```

```
# Dependent Variable
y = final_df['Bike Count']
```

```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=30)
```

Scaling

```
from sklearn.preprocessing import MinMaxScaler
scale = MinMaxScaler()
X_train = scale.fit_transform(X_train)
X_test = scale.fit_transform(X_test)
```

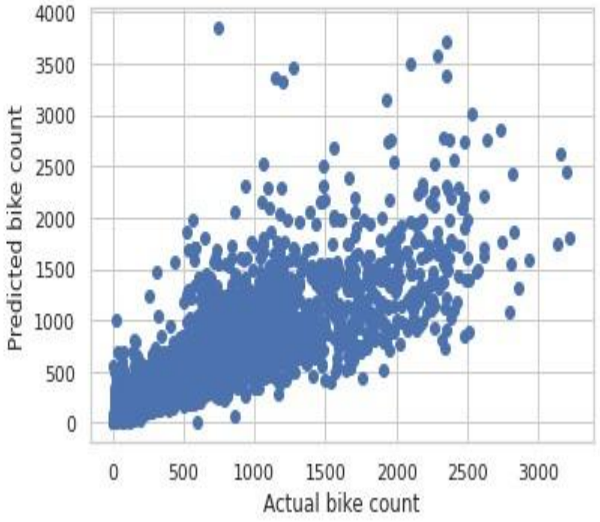
Applying Machine Learning Models and Hypertuning

1. Linear Regression
2. Lasso regression
3. Decision Tree
4. Random Forest
5. XGBoost Algorithm

We will Hypertune our Models by using GridSearchCV.

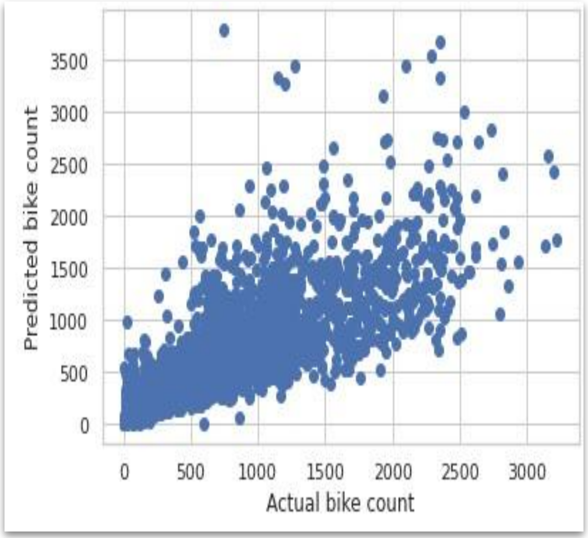
Performance of the Models after hypertuning

```
R2 : 0.6326705123041109
Adjusted R2 :  0.6253980729126162
```



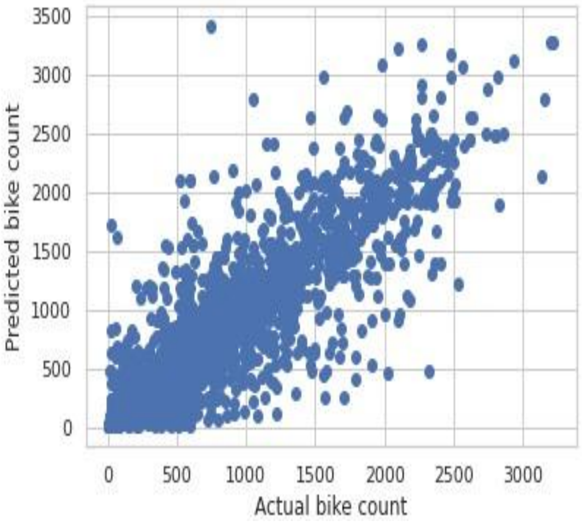
Linear Regression

```
R2 : 0.6319563058761604
Adjusted R2 :  0.6246697265282117
```



Lasso Regression

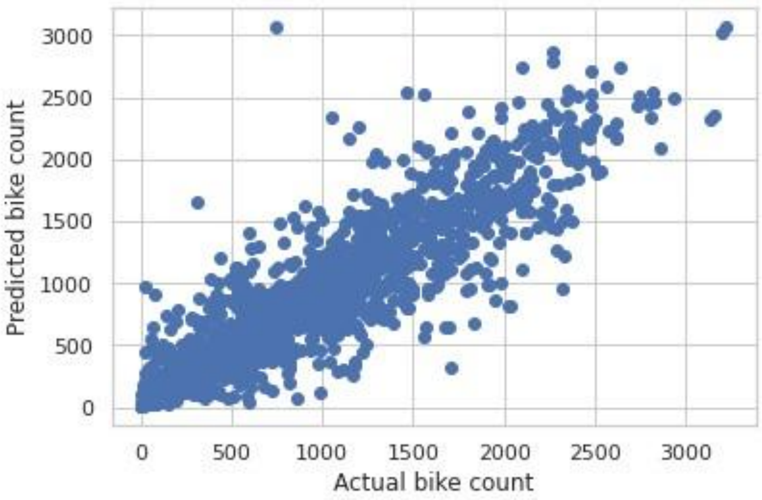
```
R2 : 0.7705060184265151
Adjusted R2 :  0.7659624652198972
```



Decision Tree

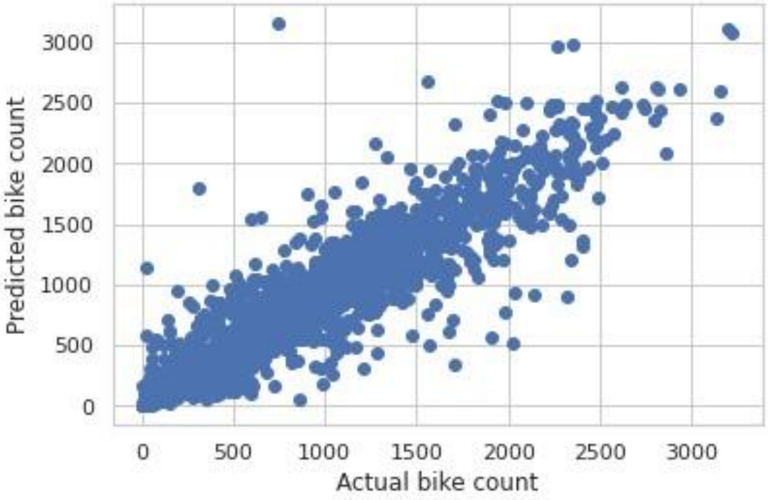
Performance of the Models after hypertuning

```
R2 : 0.8584216329755355
Adjusted R2 :  0.8556186451190729
```



Random Forest

```
R2 : 0.8776896212473072
Adjusted R2 :  0.87526810365542
```



XGBoost

Comparison

1. **Linear Regression-**

- R2 Score - 0.6326705123041109
- Adjusted R2 Score - 0.6253980729126162

2. **Lasso Regression-**

- R2 Score - 0.6319563058761604
- Adjusted R2 Score - 0.6246697265282117

3. **Dicision Tree-**

- R2 Score - 0.7705060184265151
- Adjusted R2 Score - 0.7659624652198972

4. **Random Forest-**

- R2 Score - 0.8590723534971378
- Adjusted R2 Score - 0.8562822486944802

5. **XGBoost Algorithm-**

- R2 Score - 0.8776896212473072
- Adjusted R2 Score - 0.875268103655542

Exploration Conclusion

1. There are 2 rental patterns for Working days and Non-working days.
2. People generally prefer bikes at moderate to high temperature.
3. Demand of rental bikes is high between February to October.
4. Bike demand is high on clear day and lowest on Rainy or Snowy day.

After all the machine learning operation we are coming to the conclusion that Random Forest , XGBoost model performing well than the other models. So in future if we want to do the prediction we can use these models.

Challenges

Elimination of features.

Finding best parameters for the model.

References

https://www.analyticsvidhya.com/

https://towardsdatascience.com/

https://stackoverflow.com/

Q & A