Capstone Project - 2

Seoul Bike Sharing Demand Prediction

ML Supervised Regression

Team Members:
Milan Naik
K S Apoorva

Contents

- Problem Statement
- Data Summary
- Data Analysis
- Analysis Details
- Challenges
- Conclusions

Problem Statements

- Prediction of bike count required at each hour.
- Reduce waiting time of public.

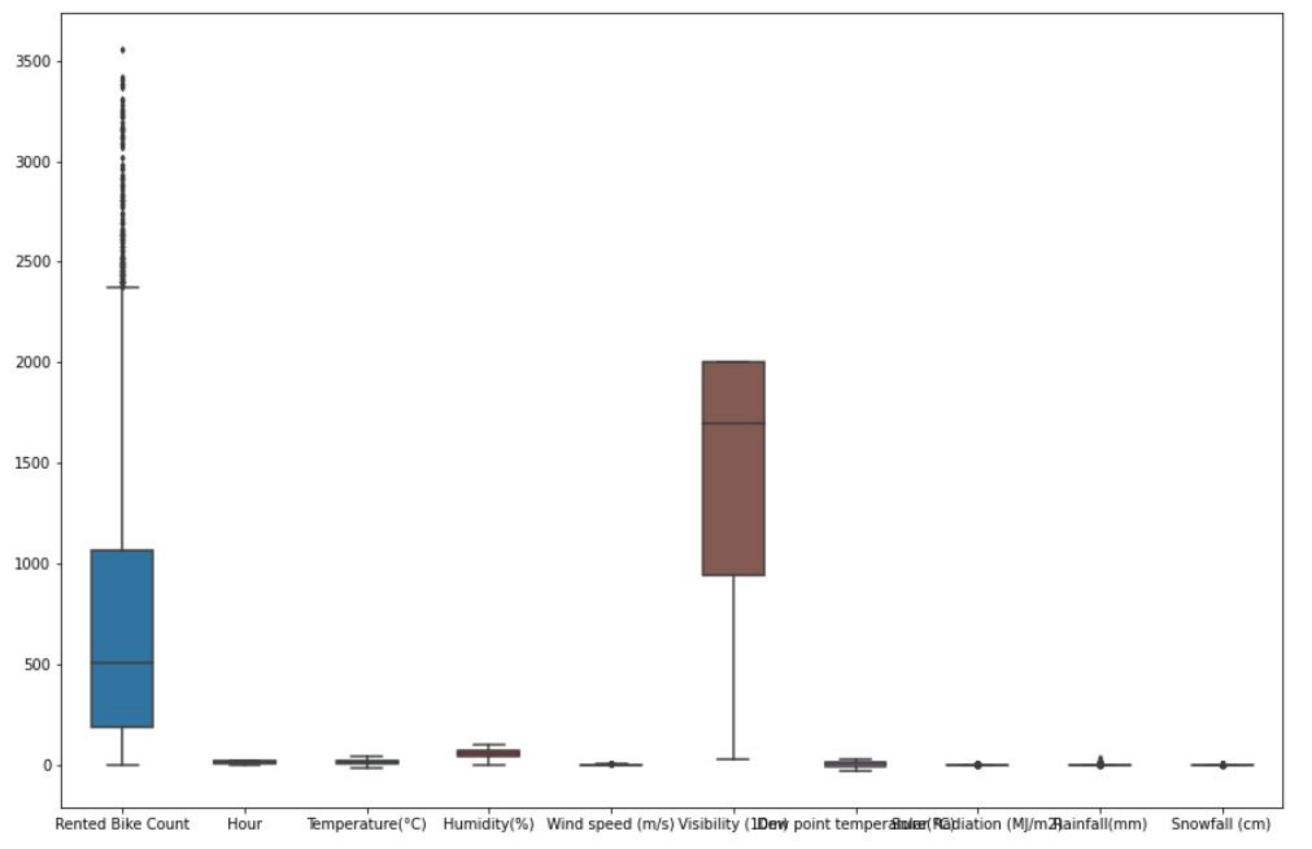
Data Summary

- 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 Hrs),Fun(Functional Hrs)

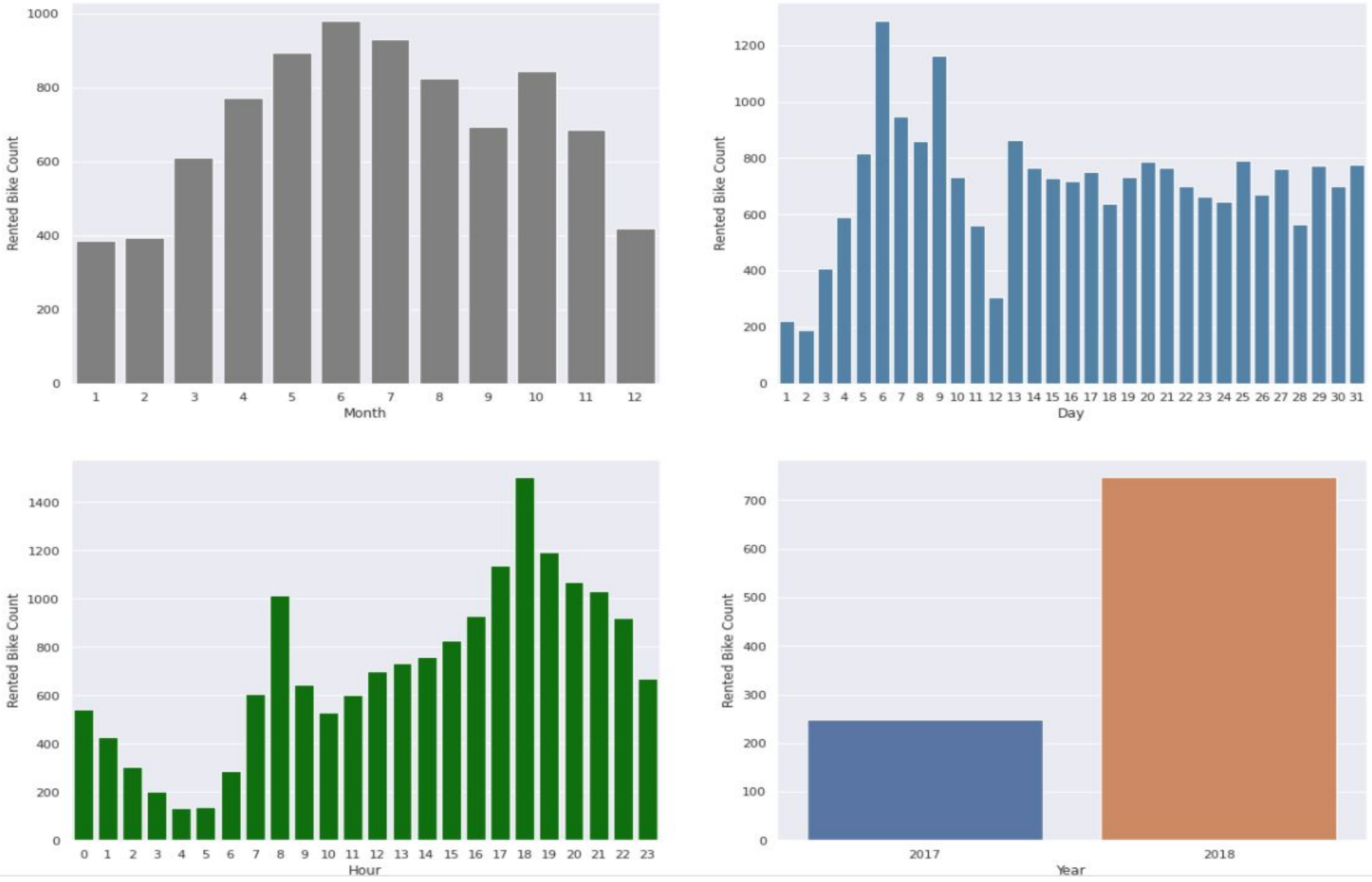
Basic Data Exploration

- The dataset has 8760 rows and 14 features(columns).
- Three categorical features ‘Seasons’, ‘Holiday’, & ‘Functioning Day’.
- Outliers present only in dependent variable.
- No Duplicated values.
- No null values.

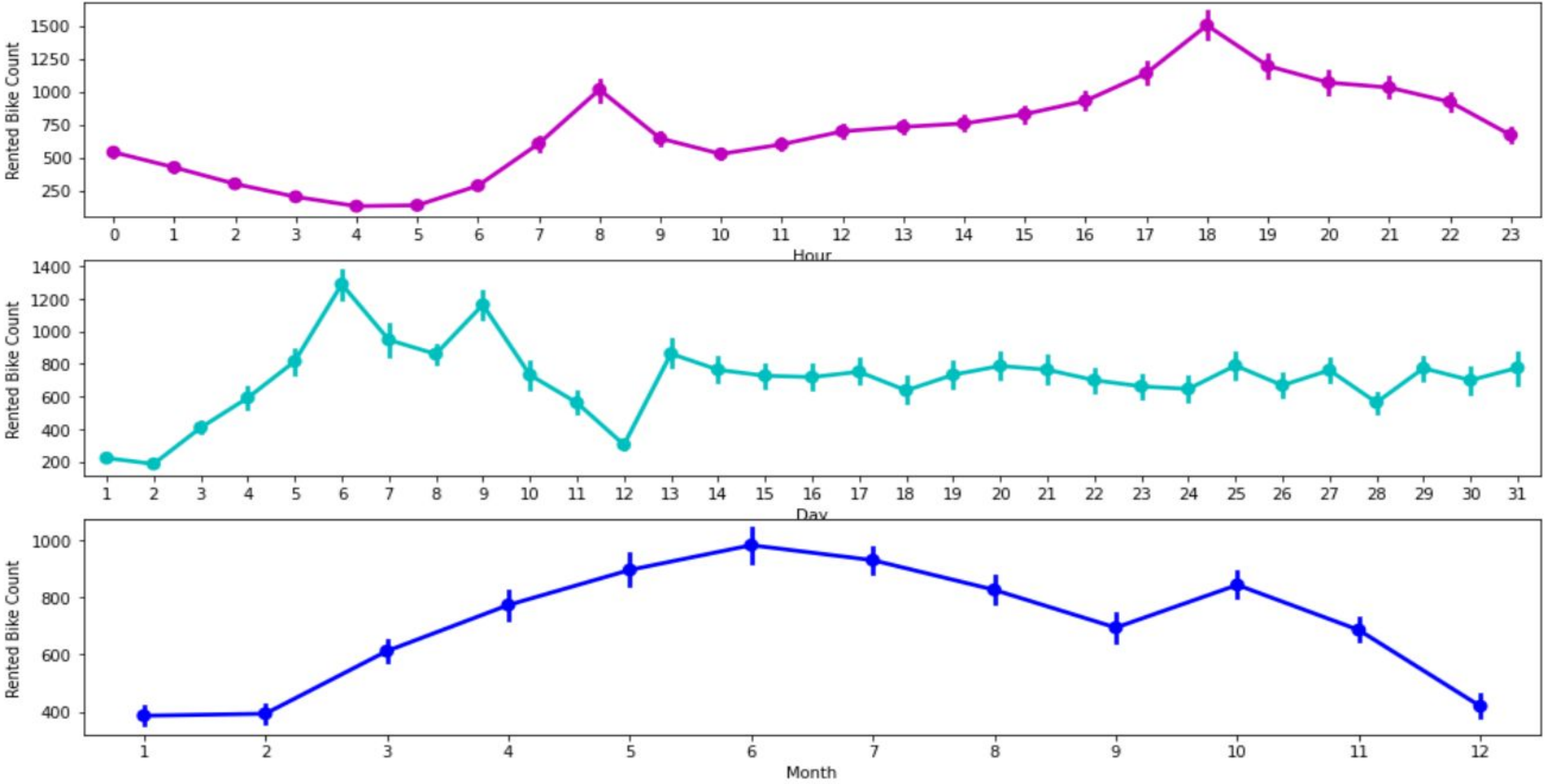
Outliers in the features



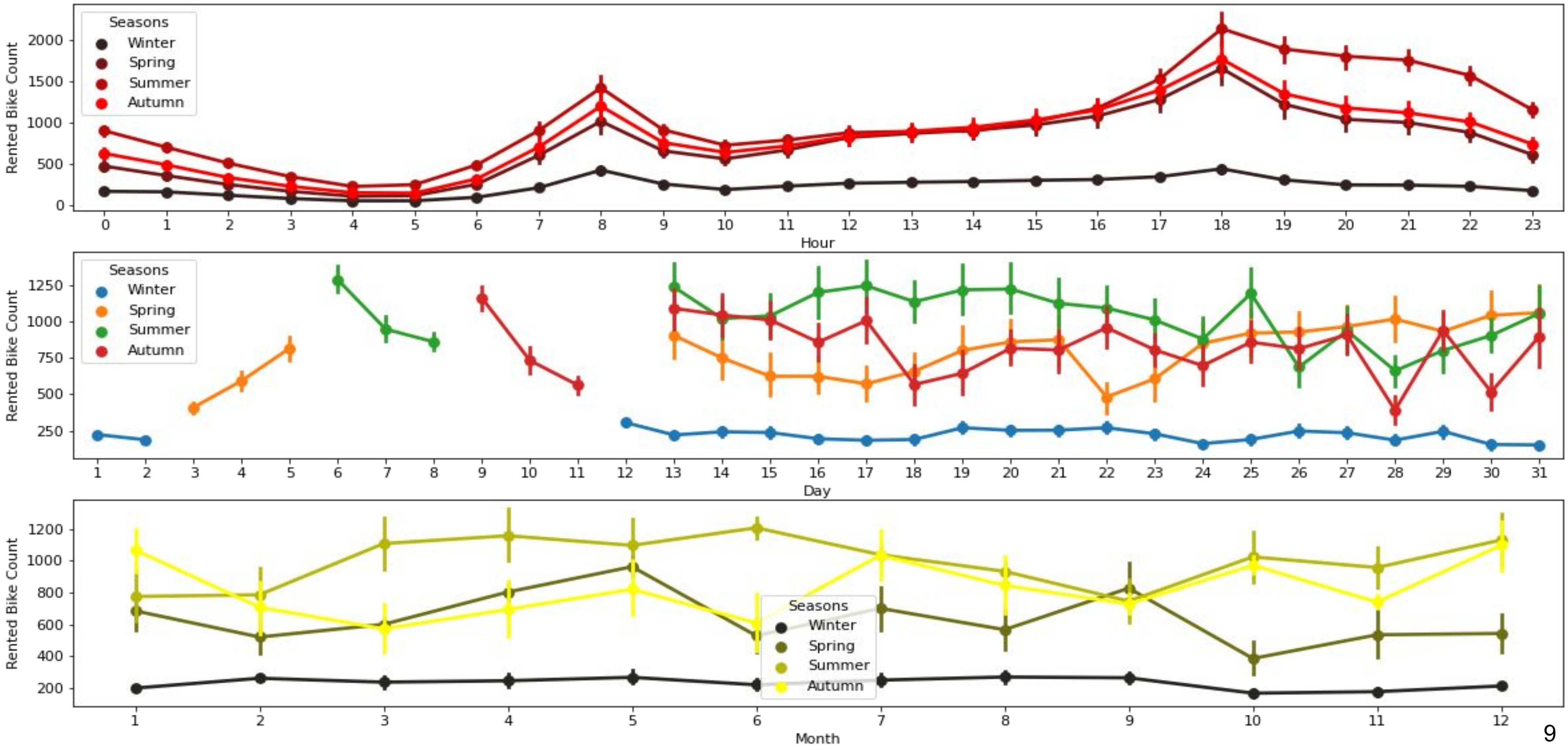
Mean Distribution of Rent Count



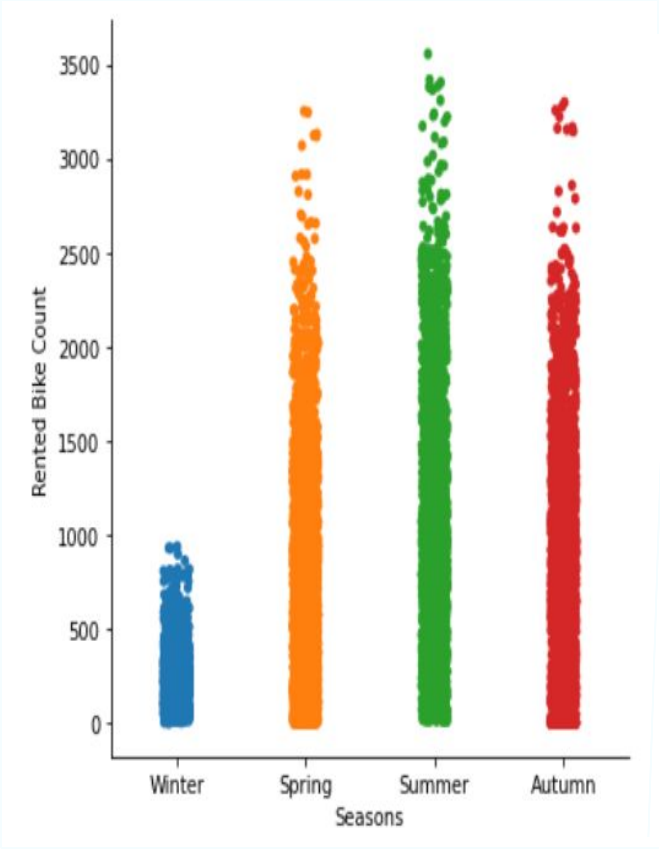
Rented Bike Spread over time

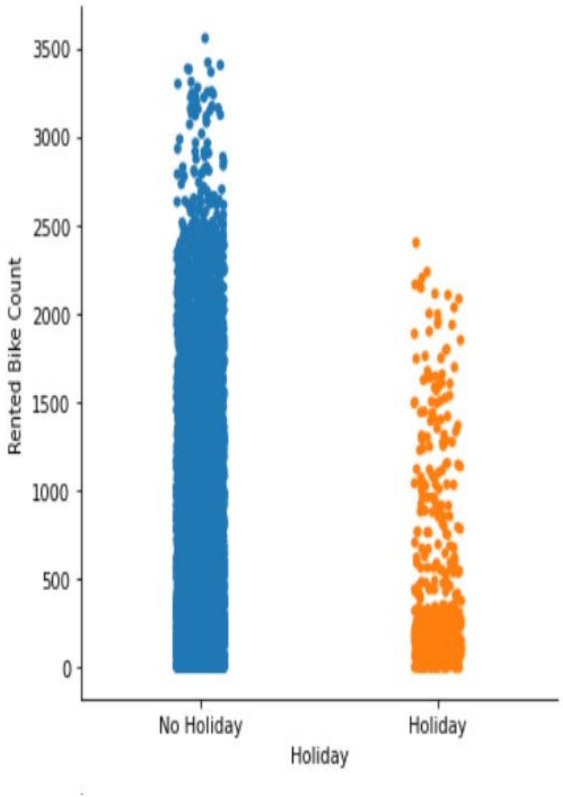


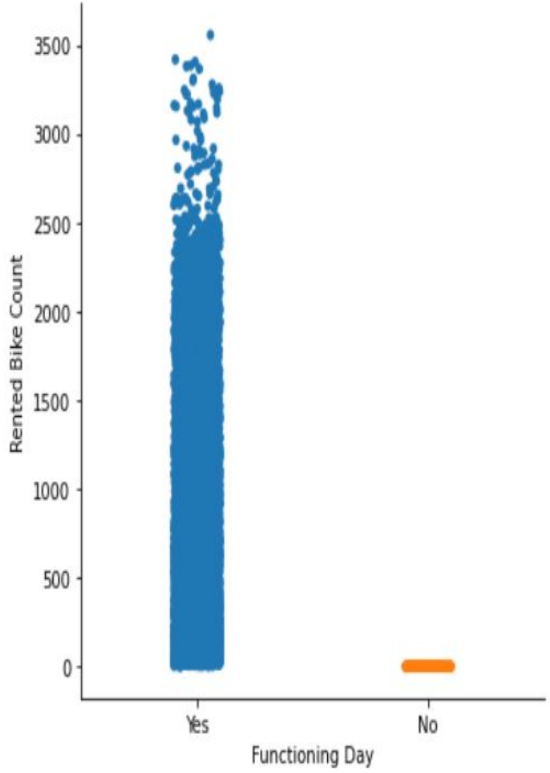
Rented Bike Spread over time and seasons



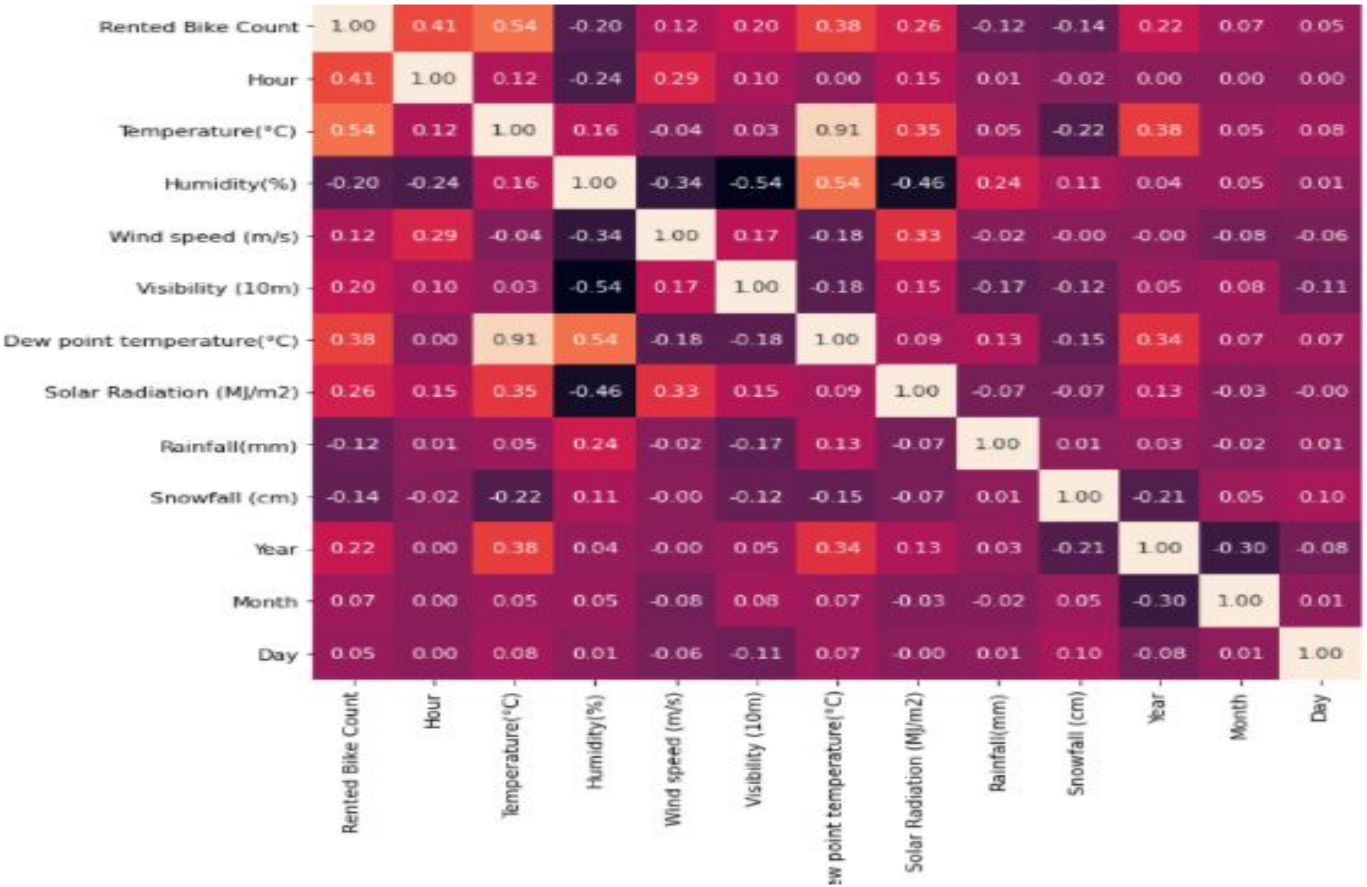
Spread of Categorical Variables







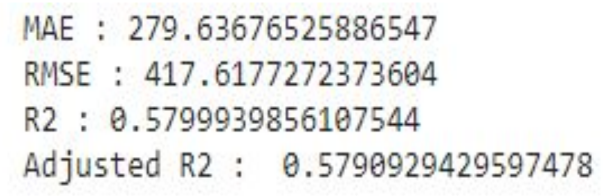
Correlation Matrix



	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	ew point temperature(°C)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Year	Month	Day
Rented Bike Count	1.00	0.41	0.54	-0.20	0.12	0.20	0.38	0.26	-0.12	-0.14	0.22	0.07	0.05
Hour	0.41	1.00	0.12	-0.24	0.29	0.10	0.00	0.15	0.01	-0.02	0.00	0.00	0.00
Temperature(°C)	0.54	0.12	1.00	0.16	-0.04	0.03	0.91	0.35	0.05	-0.22	0.38	0.05	0.08
Humidity(%)	-0.20	-0.24	0.16	1.00	-0.34	-0.54	0.54	-0.46	0.24	0.11	0.04	0.05	0.01
Wind speed (m/s)	0.12	0.29	-0.04	-0.34	1.00	0.17	-0.18	0.33	-0.02	-0.00	-0.00	-0.08	-0.06
Visibility (10m)	0.20	0.10	0.03	-0.54	0.17	1.00	-0.18	0.15	-0.17	-0.12	0.05	0.08	-0.11
Dew point temperature(°C)	0.38	0.00	0.91	0.54	-0.18	-0.18	1.00	0.09	0.13	-0.15	0.34	0.07	0.07
Solar Radiation (MJ/m2)	0.26	0.15	0.35	-0.46	0.33	0.15	0.09	1.00	-0.07	-0.07	0.13	-0.03	-0.00
Rainfall(mm)	-0.12	0.01	0.05	0.24	-0.02	-0.17	0.13	-0.07	1.00	0.01	0.03	-0.02	0.01
Snowfall (cm)	-0.14	-0.02	-0.22	0.11	-0.00	-0.12	-0.15	-0.07	0.01	1.00	-0.21	0.05	0.10
Year	0.22	0.00	0.38	0.04	-0.00	0.05	0.34	0.13	0.03	-0.21	1.00	-0.30	-0.08
Month	0.07	0.00	0.05	0.05	-0.08	0.08	0.07	-0.03	-0.02	0.05	-0.30	1.00	0.01
Day	0.05	0.00	0.08	0.01	-0.06	-0.11	0.07	-0.00	0.01	0.10	-0.08	0.01	1.00

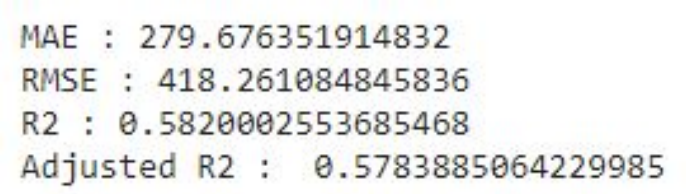
Sklearn Linear Regression

Train Set Metrics



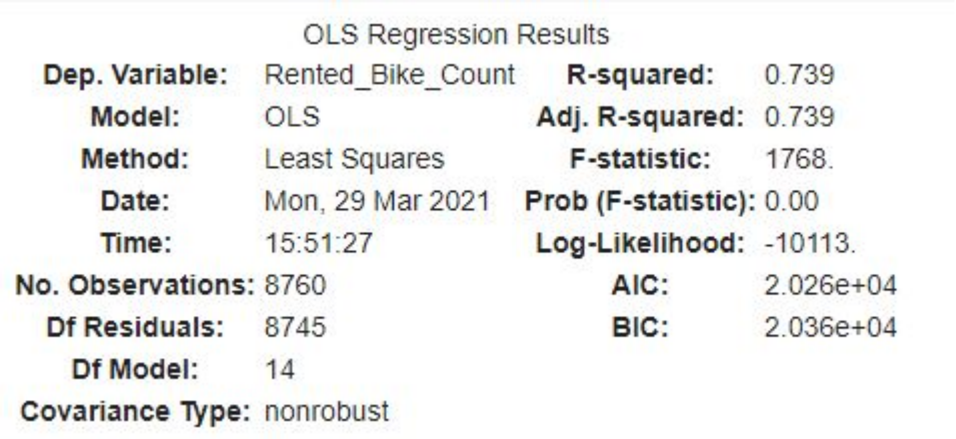
```
MAE : 279.63676525886547
RMSE : 417.6177272373604
R2 : 0.5799939856107544
Adjusted R2 :  0.5790929429597478
```

Test Set Metrics



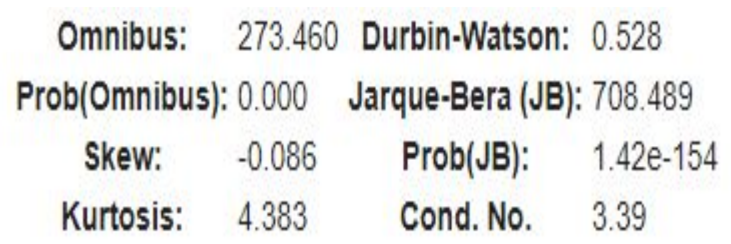
```
MAE : 279.676351914832
RMSE : 418.261084845836
R2 : 0.5820002553685468
Adjusted R2 :  0.5783885064229985
```

StatsModel Linear Regression



OLS Regression Results

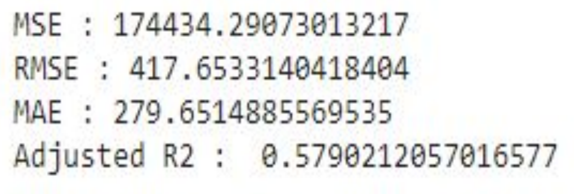
Dep. Variable:	Rented_Bike_Count	**R-squared:**	0.739
Model:	OLS	**Adj. R-squared:**	0.739
Method:	Least Squares	**F-statistic:**	1768.
Date:	Mon, 29 Mar 2021	**Prob (F-statistic):**	0.00
Time:	15:51:27	**Log-Likelihood:**	-10113.
No. Observations:	8760	**AIC:**	2.026e+04
Df Residuals:	8745	**BIC:**	2.036e+04
Df Model:	14		
Covariance Type:	nonrobust		



Omnibus:	273.460	**Durbin-Watson:**	0.528
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	708.489
Skew:	-0.086	**Prob(JB):**	1.42e-154
Kurtosis:	4.383	**Cond. No.**	3.39

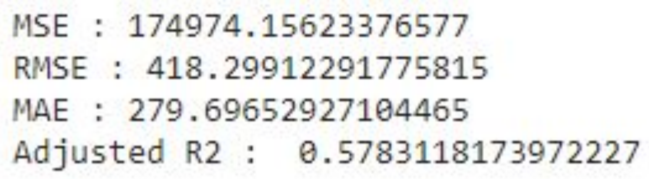
Lasso Regression

Train Set Metrics



```
MSE : 174434.29073013217
RMSE : 417.6533140418404
MAE : 279.6514885569535
Adjusted R2 :  0.5790212057016577
```

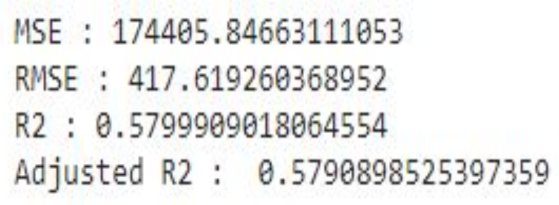
Test Set Metrics



```
MSE  : 174974.15623376577
RMSE  : 418.29912291775815
MAE  : 279.69652927104465
Adjusted R2 :  0.5783118173972227
```

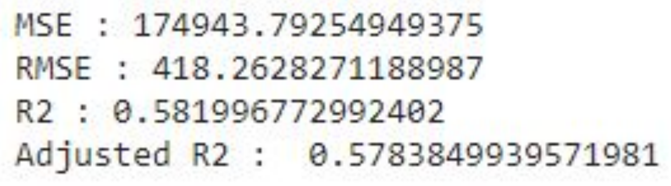
Ridge Regression

Train Set Metrics



```
MSE : 174405.84663111053
RMSE : 417.619260368952
R2 : 0.5799909018064554
Adjusted R2 :  0.5790898525397359
```

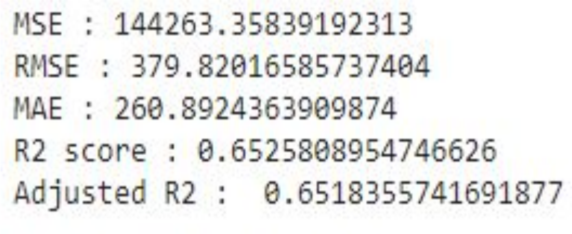
Test Set Metrics



```
MSE  : 174943.79254949375
RMSE  : 418.2628271188987
R2  : 0.581996772992402
Adjusted R2 :  0.5783849939571981
```

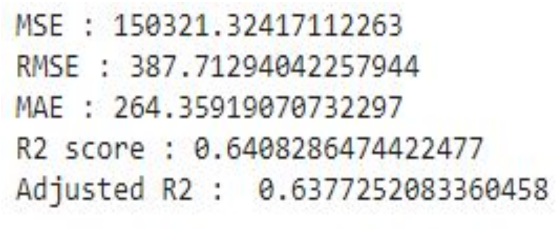
Decision Tree

Train Set Metrics



```
MSE : 144263.35839192313
RMSE : 379.82016585737404
MAE : 260.8924363909874
R2 score : 0.6525808954746626
Adjusted R2 :  0.6518355741691877
```

Test Set Metrics



```
MSE : 150321.32417112263
RMSE : 387.71294042257944
MAE : 264.35919070732297
R2 score : 0.6408286474422477
Adjusted R2 :  0.6377252083360458
```

Parameters:

i) Criterion : mse

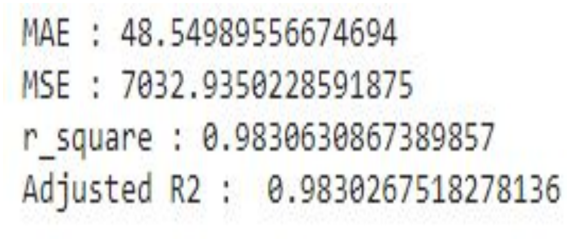
ii) max_leaf node = 9

iii) min_sample_leaf = 1

iv) min_sample_split = 2

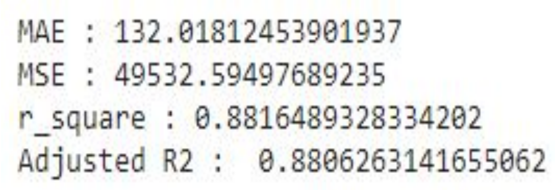
Random Forest

Train Set Metrics



```
MAE : 48.54989556674694
MSE : 7032.9350228591875
r_square : 0.9830630867389857
Adjusted R2 :  0.9830267518278136
```

Test Set Metrics



```
MAE : 132.01812453901937
MSE : 49532.59497689235
r_square : 0.8816489328334202
Adjusted R2 :  0.8806263141655062
```

Parameters:

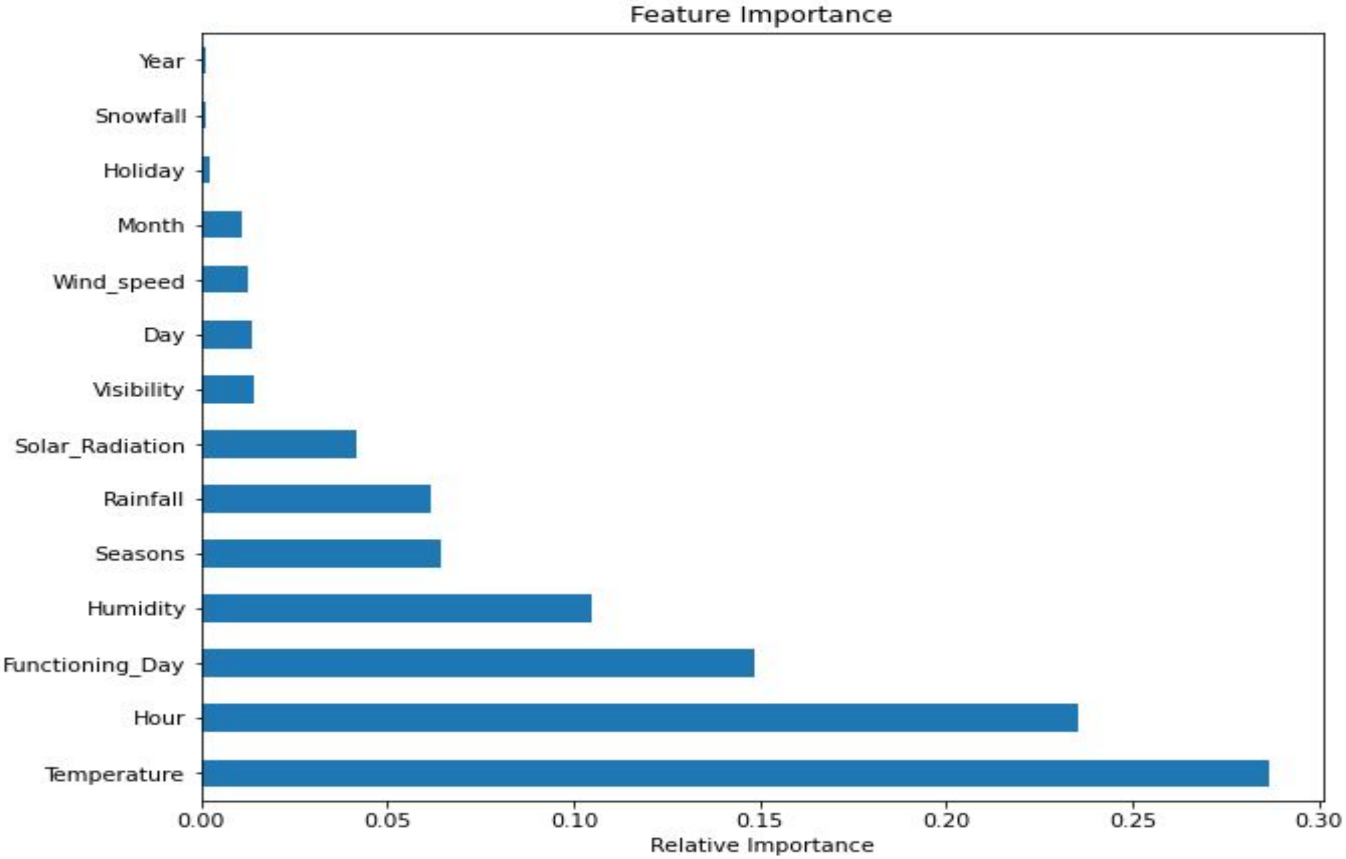
i) Criterion : mse

ii) n_estimators = 100

iii) min_sample_leaf = 1

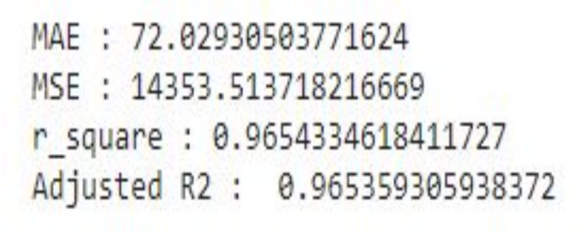
iv) min_sample_split = 2

Random Forest Feature Importance



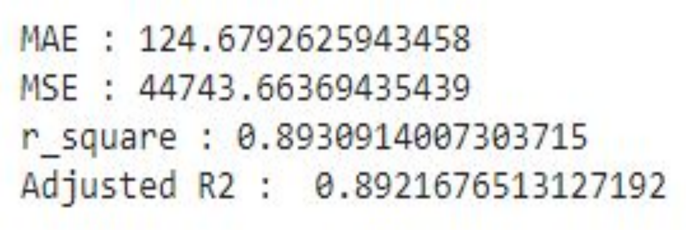
Xtreme Gradient Boosting

Train Set Metrics



```
MAE : 72.02930503771624
MSE : 14353.513718216669
r_square : 0.9654334618411727
Adjusted R2 :  0.965359305938372
```

Test Set Metrics



```
MAE : 124.6792625943458
MSE : 44743.66369435439
r_square : 0.893091400730371 5
Adjusted R2 :  0.8921676513127192
```

Parameters:

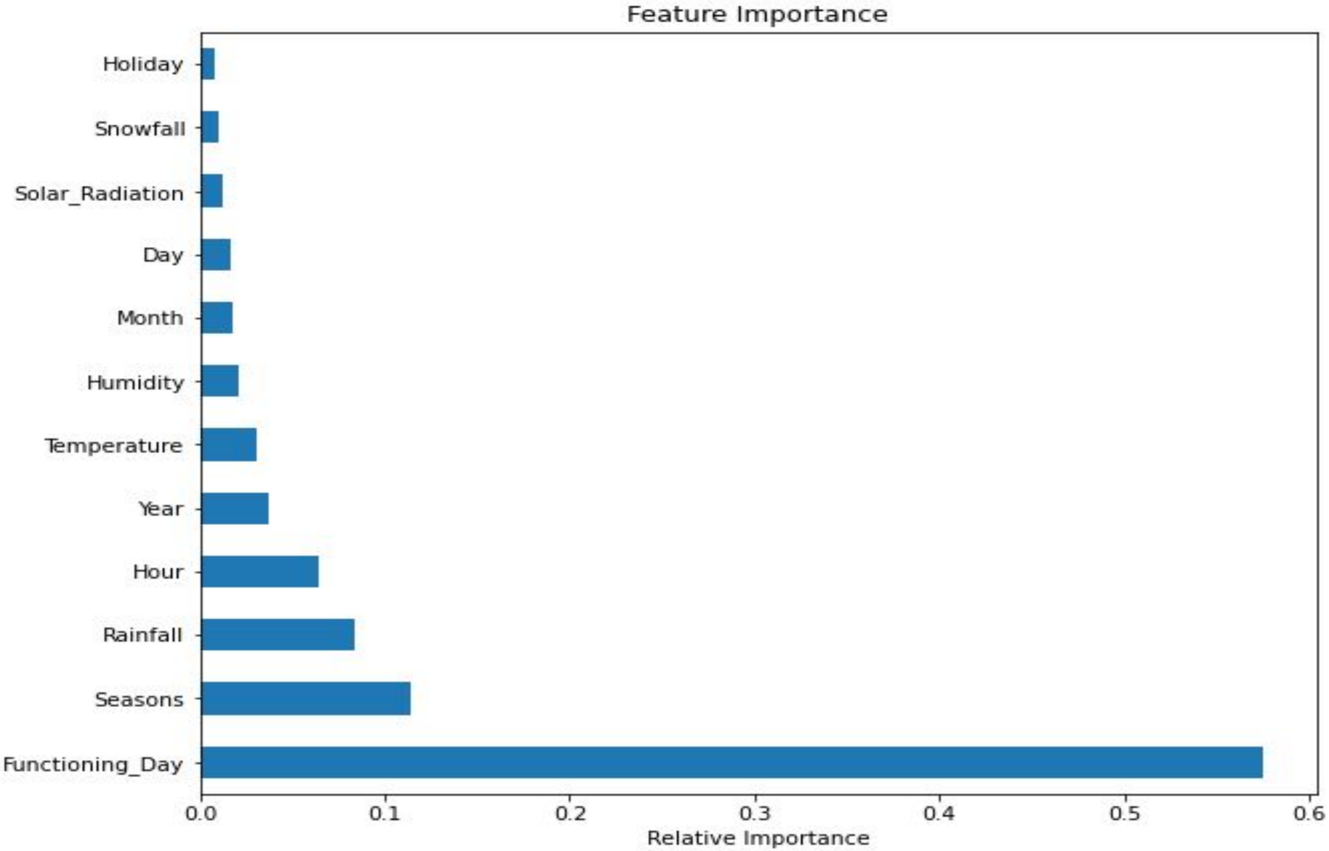
i) learning rate = 0.1

ii) n_estimators = 100

iii) max_depth = 9

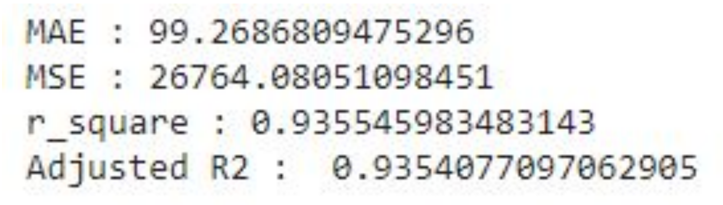
iv) min_sample_split = 2

XGboost Feature Importance



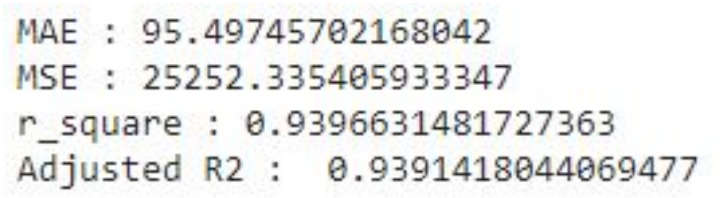
Xtreme Gradient Boosting- Grid SearchCv

Train Set Metrics



```
MAE : 99.2686809475296
MSE : 26764.08051098451
r_square : 0.935545983483143
Adjusted R2 :  0.9354077097062905
```

Test Set Metrics



```
MAE : 95.49745702168042
MSE : 25252.335405933347
r_square : 0.9396631481727363
Adjusted R2 :  0.9391418044069477
```

Parameters:

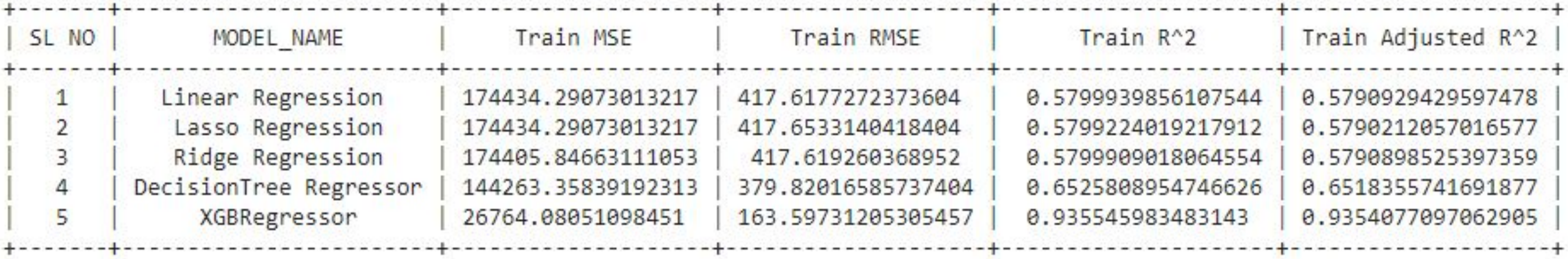
i) learning rate = 0.1

ii) n_estimators = 100

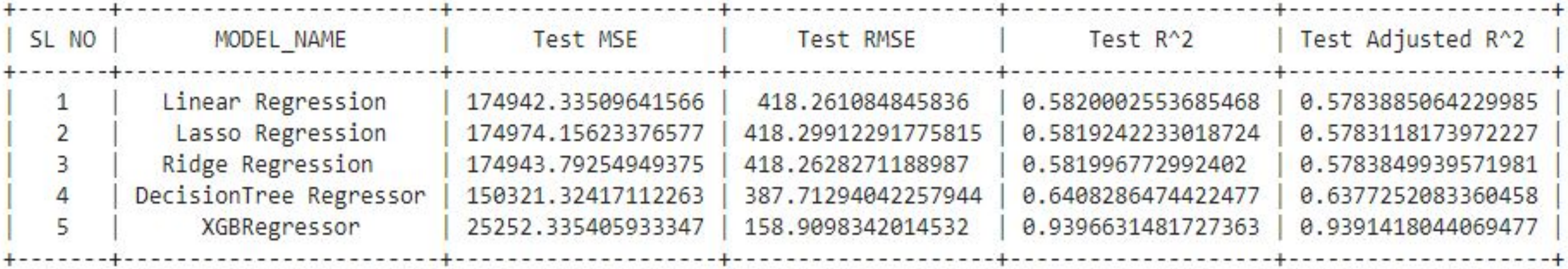
iii) max_depth = 7

iv) min_child_weight = 10

All Model Summary- Metrics



```
+-------+--------------------------+--------------------+--------------------+--------------------+--------------------+
| SL NO |        MODEL_NAME        |     Train MSE      |     Train RMSE     |     Train R^2      | Train Adjusted R^2 |
+-------+--------------------------+--------------------+--------------------+--------------------+--------------------+
|   1   |    Linear Regression     | 174434.29073013217 | 417.6177272373604  | 0.5799939856107544 | 0.5790929429597478 |
|   2   |     Lasso Regression     | 174434.29073013217 | 417.6533140418404  | 0.5799224019217912 | 0.579021205701657  |
|   3   |     Ridge Regression     | 174405.84663111053 |  417.619260368952  | 0.5799909018064554 | 0.5790898525397359 |
|   4   | DecisionTree Regressor   | 144263.35839192313 | 379.82016585737404 | 0.6525808954746626 | 0.6518355741691877 |
|   5   |       XGBRegressor       | 26764.08051098451  | 163.59731205305457 | 0.935545983483143  | 0.9354077097062905 |
+-------+--------------------------+--------------------+--------------------+--------------------+--------------------+
```



```
+-------+--------------------------+--------------------+--------------------+--------------------+--------------------+
| SL NO |        MODEL_NAME        |      Test MSE      |     Test RMSE      |      Test R^2      | Test Adjusted R^2  |
+-------+--------------------------+--------------------+--------------------+--------------------+--------------------+
|   1   |    Linear Regression     | 174942.33509641566 |  418.261084845836  | 0.5820002553685468 | 0.5783885064229985 |
|   2   |     Lasso Regression     | 174974.15623376577 | 418.29912291775815 | 0.5819242233018724 | 0.5783118173972227 |
|   3   |     Ridge Regression     | 174943.79254949375 | 418.2628271188987  | 0.581996772992402  | 0.5783849939571981 |
|   4   |  DecisionTree Regressor  | 150321.32417112263 | 387.71294042257944 | 0.6408286474422477 | 0.6377252083360458 |
|   5   |       XGBRegressor       | 25252.335405933347 | 158.9098342014532  | 0.9396631481727363 | 0.9391418044069477 |
+-------+--------------------------+--------------------+--------------------+--------------------+--------------------+
```

Challenges

- Handling the size of Large dataset.
- Feature Engineering.
- Removing the overfitting .
- Optimising the model.

Conclusion

- Comparing to all other algorithms XGboost has less Mean Squared error and Mean Absolute error, with model score of 95% and R-Squared value 93%.

- Total amount of Bike rentals increases with increase in temperature.

- Features like Functioning Day and Temperature has the higher importance for the model.

- There exists higher correlation between temperature and rental bike count. This could be a stepping stone for new bike rental stations.

Thank You