Capstone Project - II

ML Supervised Regression

Seoul Bike Sharing Demand Prediction

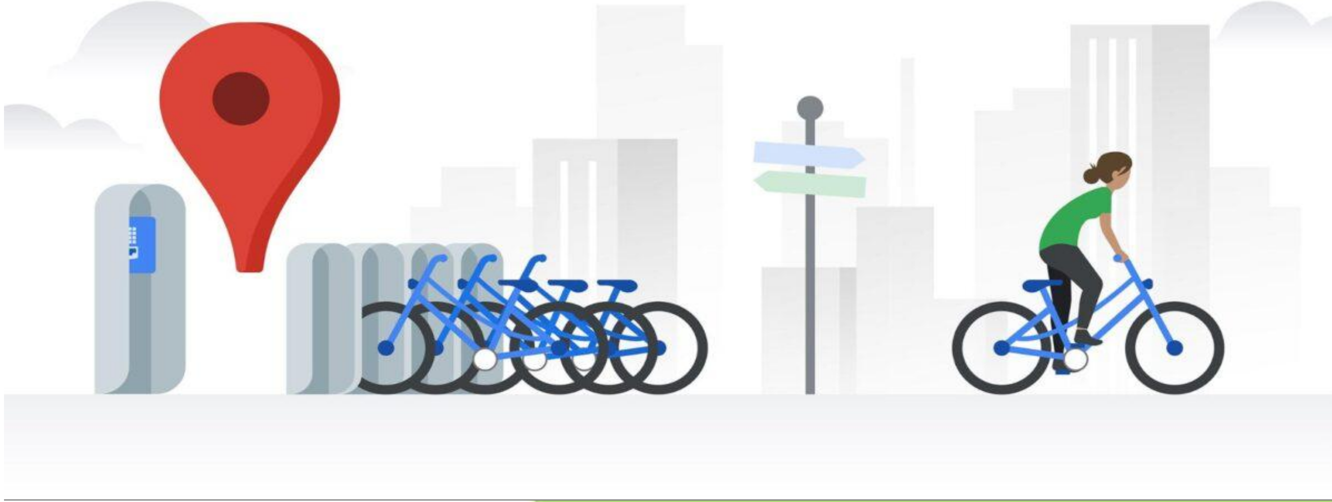
Presented By -
Kunika Gupta

TABLE OF CONTENTS

Problem Statements

- Prediction of bike count required at each hour.

- Reduce waiting time of public.



Dataset Attributes and their Description

- Date : year-month-day
- Rented Bike count at each hour
- Hour - Hour of the day
- Temperature- In Celsius
- Humidity - %
- Wind Speed - m/s
- Visibility - 10m
- Dew point temperature- In 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)

Data Exploration

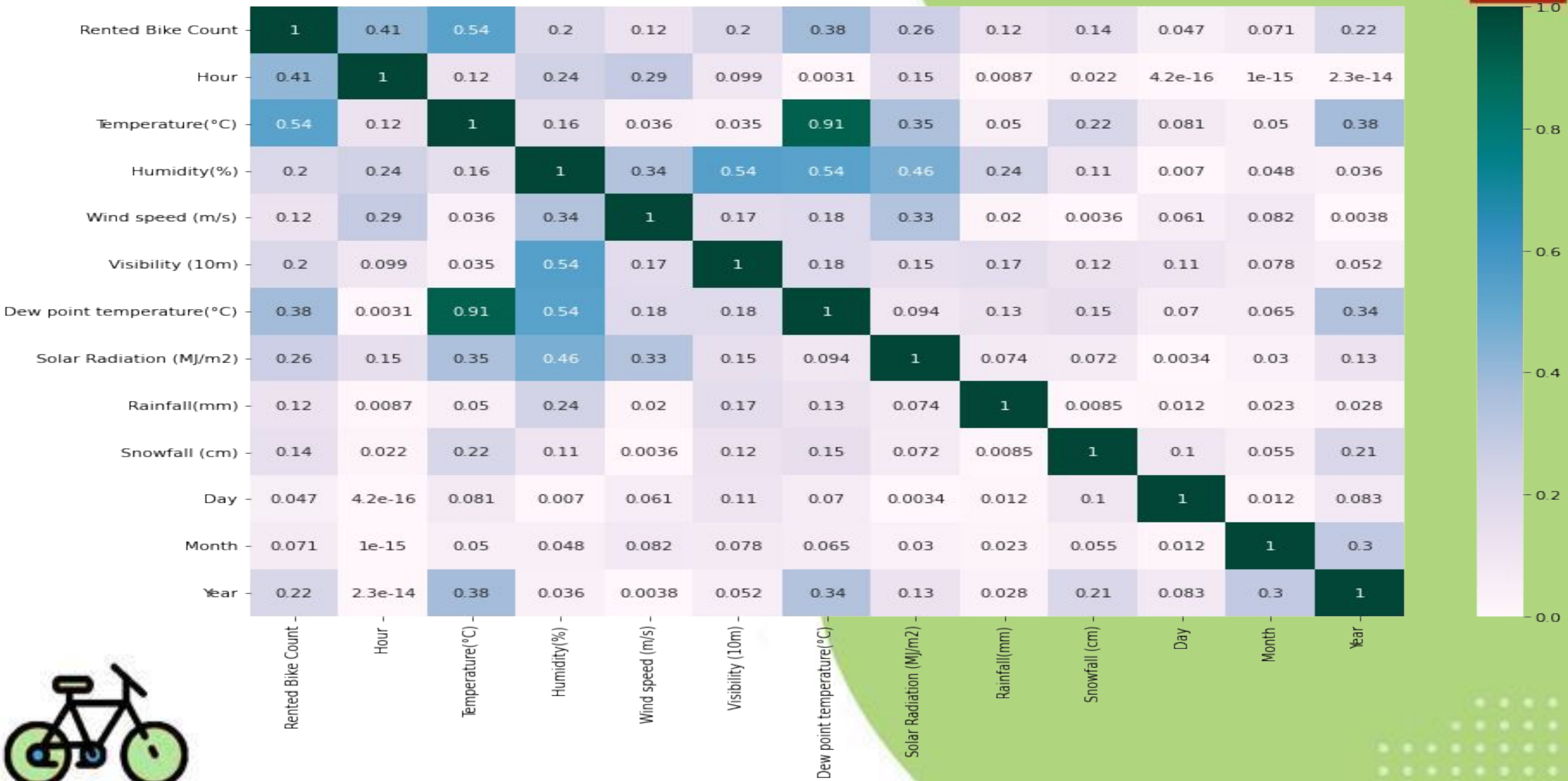
- ❖ The dataset has 8760 rows and 14 features(columns).
- ❖ Three categorical features Seasons, Holiday, & Functioning Day.
- ❖ One Datetime features.
- ❖ Outliers present only in dependent variable.
- ❖ No null values.
- ❖ No Duplicate values.
- ❖ No Missing Values.

EDA

- Exploratory data analysis or commonly known as EDA helps to explore data, and possibly formulate hypotheses that might cause new data collection and experiments. EDA build a robust understanding of the data, issues associated with either the info or process. it’s a scientific approach to get the story of the data.

- It focuses more narrowly on checking assumptions required for model fitting and hypothesis testing. It also helps while handling missing values and making transformations of variables as needed.

Checking Multicollinearity



VIF

- VIF determines the strength of the correlation between the independent variables.
- VIF less than 5 will be included in the model. In some cases VIF of less than 10 is also acceptable.

	variables	VIF
0	Hour	4.456946
1	Temperature(°C)	188.757275
2	Humidity(%)	187.140788
3	Wind speed (m/s)	4.848147
4	Visibility (10m)	10.695216
5	Dew point temperature(°C)	127.016687
6	Solar Radiation (MJ/m2)	2.909493
7	Rainfall(mm)	1.103999
8	Snowfall (cm)	1.152549
9	Day	4.420676
10	Month	4.722327
11	Year	407.294385

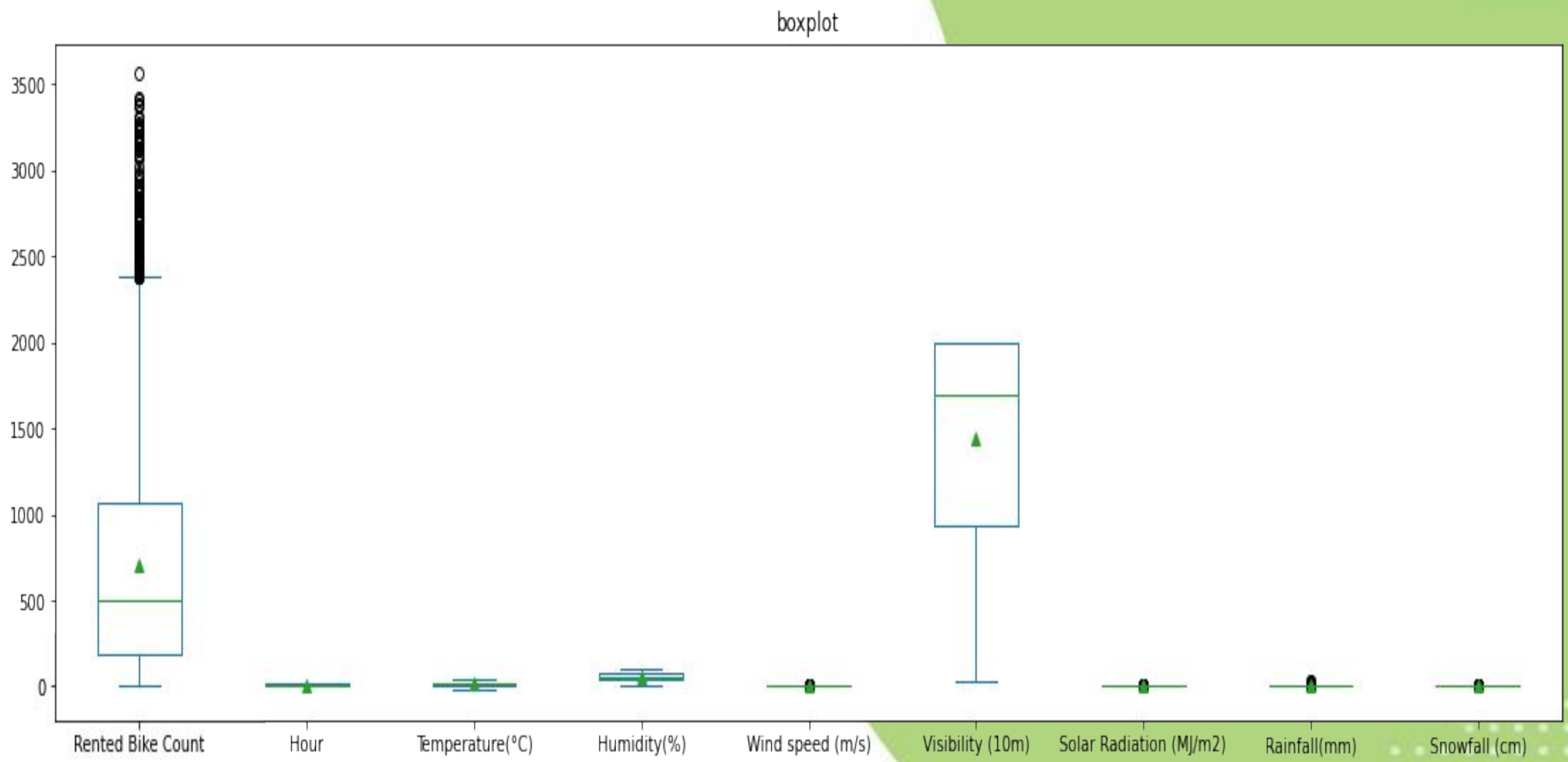
	variables	VIF
0	Hour	4.424883
1	Temperature(°C)	38.365560
2	Humidity(%)	8.326992
3	Wind speed (m/s)	4.836834
4	Visibility (10m)	9.425316
5	Dew point temperature(°C)	19.812251
6	Solar Radiation (MJ/m2)	2.905084
7	Rainfall(mm)	1.082979
8	Snowfall (cm)	1.141184
9	Day	4.346307
10	Month	4.692494

	variables	VIF
0	Hour	3.998419
1	Temperature(°C)	3.236167
2	Humidity(%)	6.757926
3	Wind speed (m/s)	4.621365
4	Visibility (10m)	5.455330
5	Solar Radiation (MJ/m2)	2.280208
6	Rainfall(mm)	1.081555
7	Snowfall (cm)	1.136671
8	Day	3.849545
9	Month	4.603431

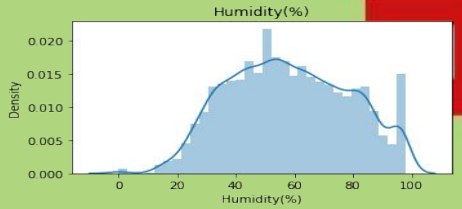
Checking multicollinearity after applying VIF & removing Variables that have high collinearity.

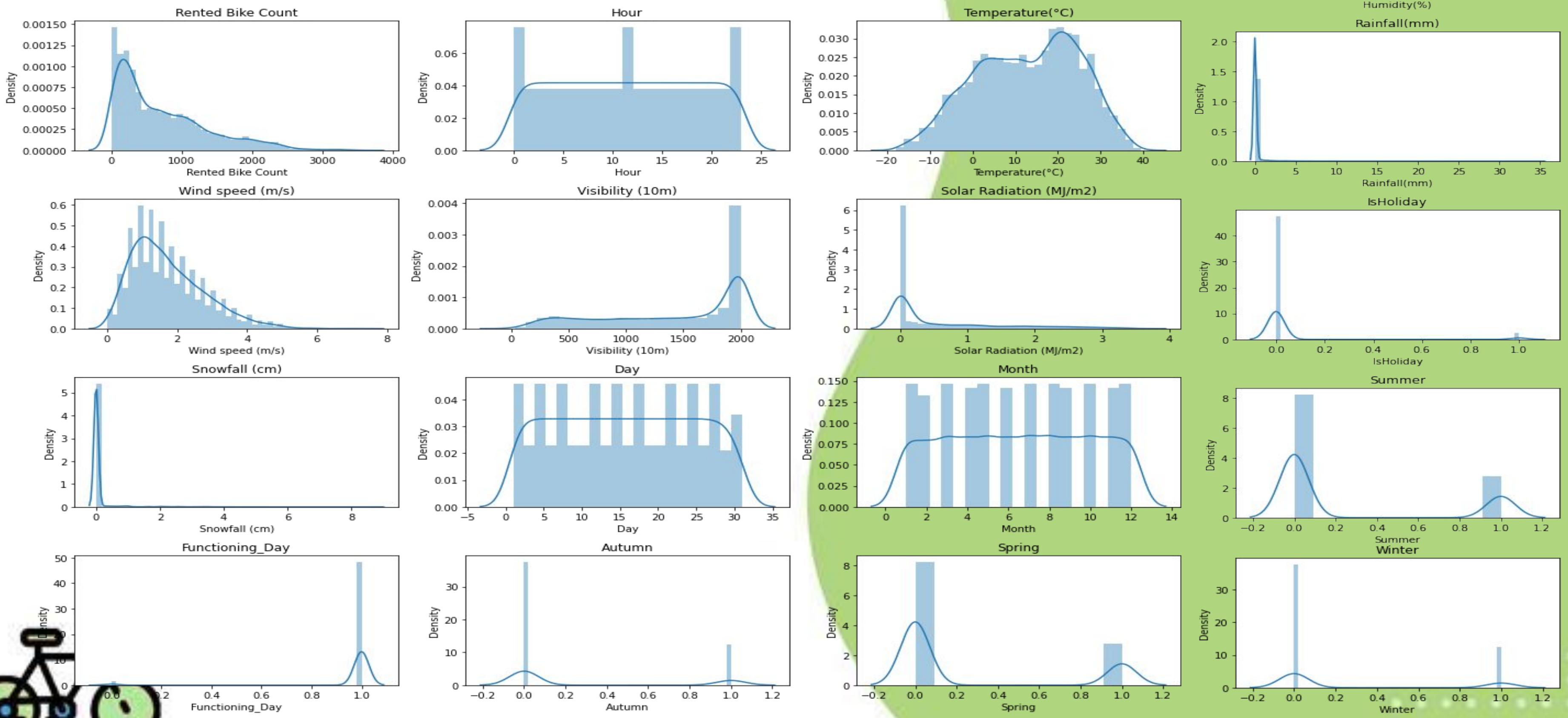
	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Day	Month
Rented Bike Count	1	0.41	0.54	0.2	0.12	0.2	0.26	0.12	0.14	0.047	0.071
Hour	0.41	1	0.12	0.24	0.29	0.099	0.15	0.0087	0.022	4.2e-16	1e-15
Temperature(°C)	0.54	0.12	1	0.16	0.036	0.035	0.35	0.05	0.22	0.081	0.05
Humidity(%)	0.2	0.24	0.16	1	0.34	0.54	0.46	0.24	0.11	0.007	0.048
Wind speed (m/s)	0.12	0.29	0.036	0.34	1	0.17	0.33	0.02	0.0036	0.061	0.082
Visibility (10m)	0.2	0.099	0.035	0.54	0.17	1	0.15	0.17	0.12	0.11	0.078
Solar Radiation (MJ/m2)	0.26	0.15	0.35	0.46	0.33	0.15	1	0.074	0.072	0.0034	0.03
Rainfall(mm)	0.12	0.0087	0.05	0.24	0.02	0.17	0.074	1	0.0085	0.012	0.023
Snowfall (cm)	0.14	0.022	0.22	0.11	0.0036	0.12	0.072	0.0085	1	0.1	0.055
Day	0.047	4.2e-16	0.081	0.007	0.061	0.11	0.0034	0.012	0.1	1	0.012
Month	0.071	1e-15	0.05	0.048	0.082	0.078	0.03	0.023	0.055	0.012	1

Boxplot for Numerical features



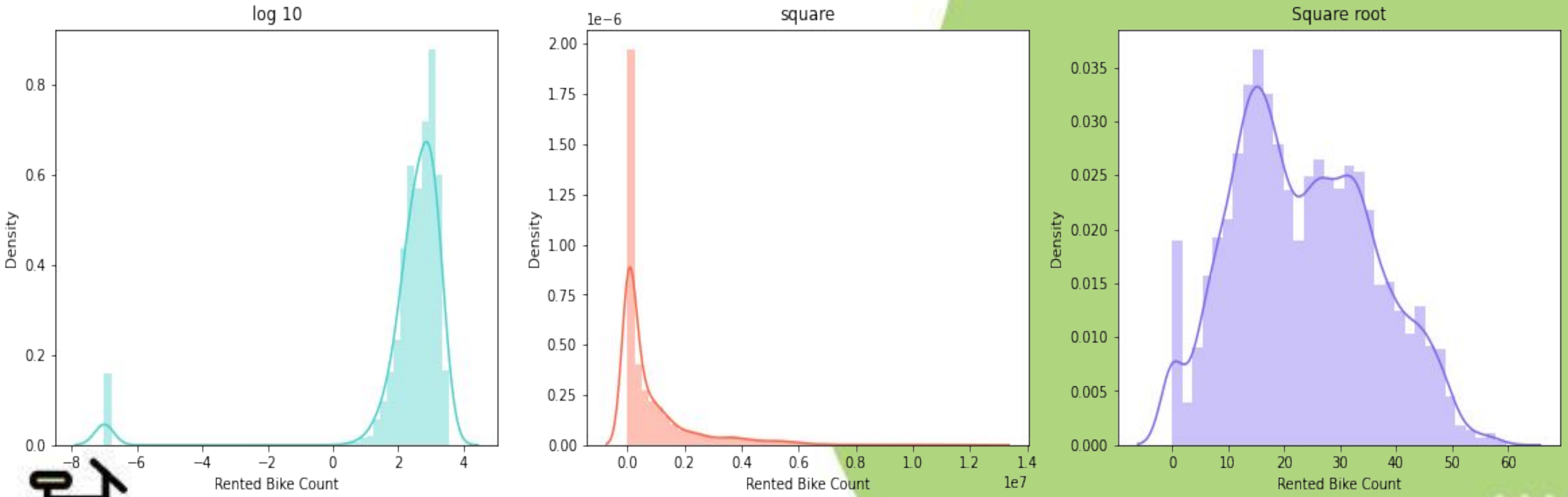
Checking the skewness of variable





Normalization

- As we saw, our dependent variable is right skewed. So we will try some transformations to normalize it.
- We observed that 'Square Root' transformation is normalizing the dependent variable so we will use this transformation while we will split our data in Train and Test.



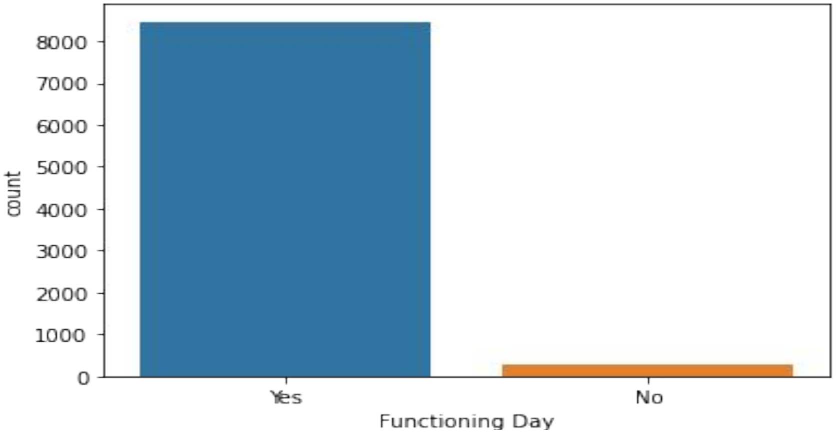
Feature Engineering

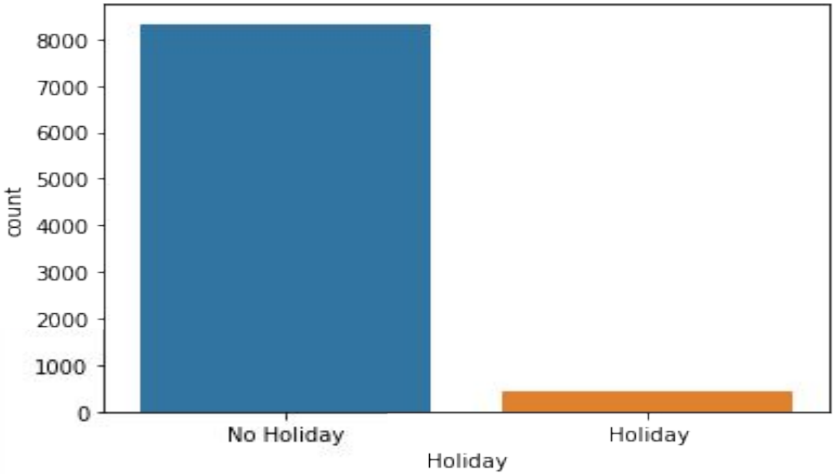
All machine learning algorithms use some input data to create outputs. Algorithms require features with some specific characteristics to work properly. Here, the need for feature engineering arises. Feature engineering mainly have two goals:

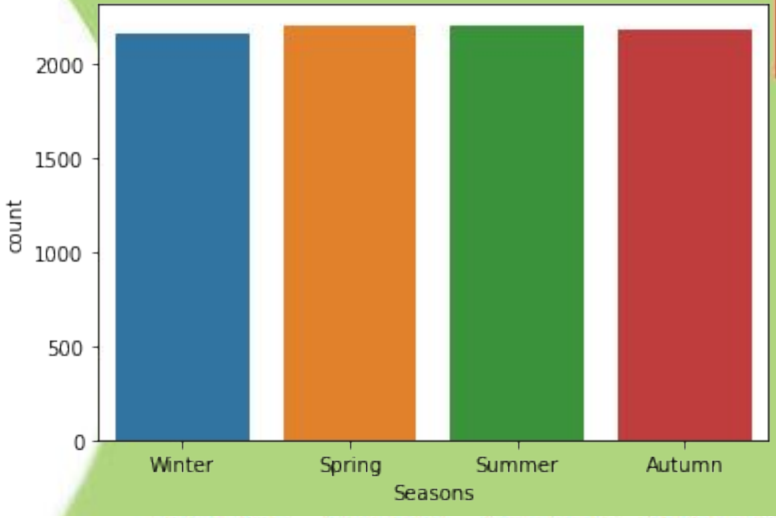
- ❖ Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
- ❖ Improving the performance of machine learning models.

We'll try adding and removing some features in this section in order to make a perfect data matrix we can pass to a machine learning model. We will try to interpret categorical features as numeric to be passed to the ML models.

Categorical Variable







	Autumn	Spring	Summer	Winter
0	0	0	0	1
1	0	0	0	1
2	0	0	0	1
3	0	0	0	1
4	0	0	0	1
...	...	...	...	...

8760 rows × 4 columns

```
df.head()
```

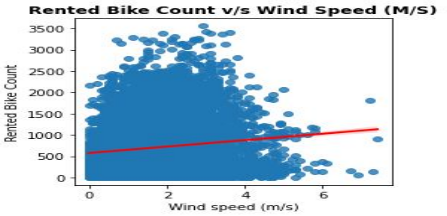
Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)	Day	Month	IsHoliday	Functioning_Day	Autumn	Spring	Summer	Winter
254	0	-5.2	37	2.2	2000	0.0	0.0	0.0	12	1	0	1	0	0	0	1
204	1	-5.5	38	0.8	2000	0.0	0.0	0.0	12	1	0	1	0	0	0	1
173	2	-6.0	39	1.0	2000	0.0	0.0	0.0	12	1	0	1	0	0	0	1
107	3	-6.2	40	0.9	2000	0.0	0.0	0.0	12	1	0	1	0	0	0	1
78	4	-6.0	36	2.3	2000	0.0	0.0	0.0	12	1	0	1	0	0	0	1

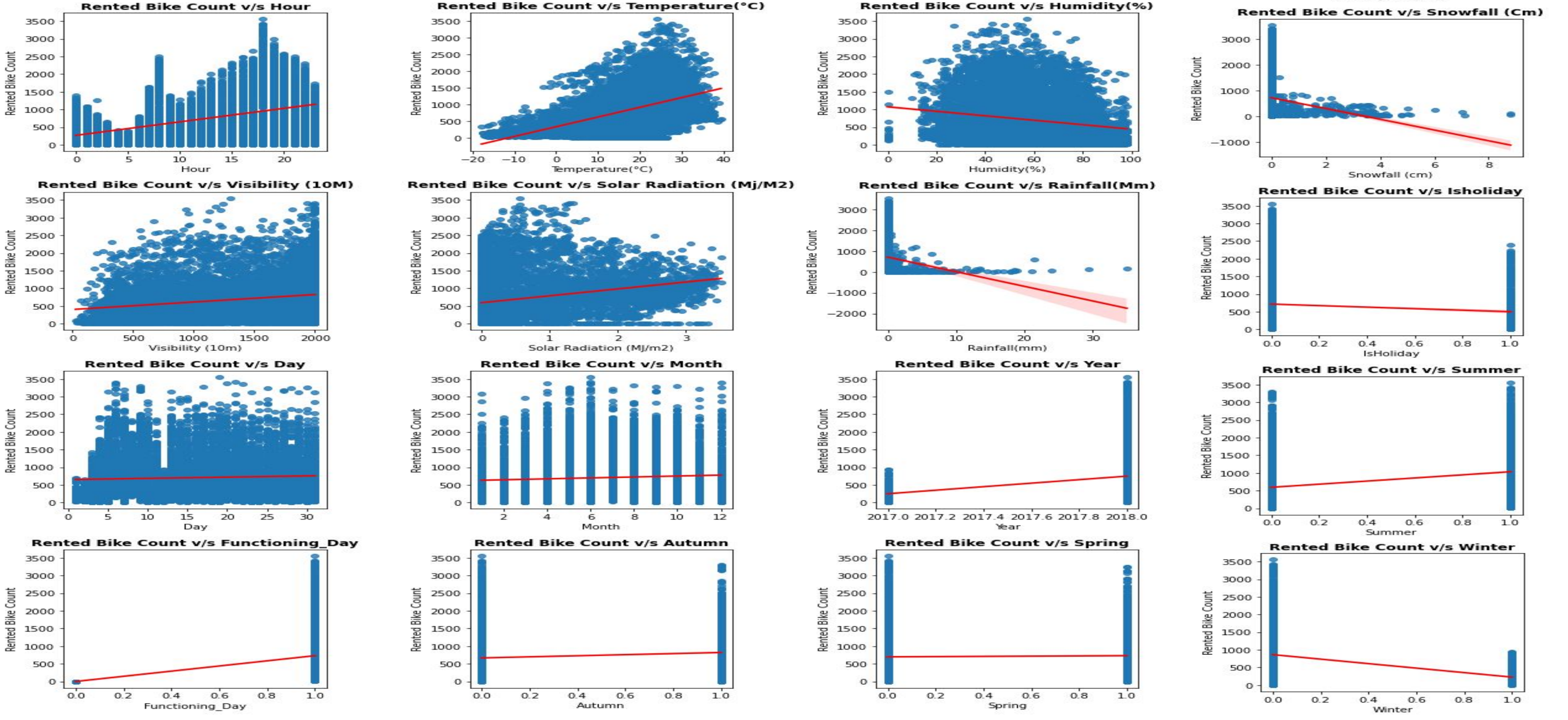
```
# Checking the shape of the updated dataframe.

df.shape
```

```
(8760, 17)
```

Independent features linearity with Target feature

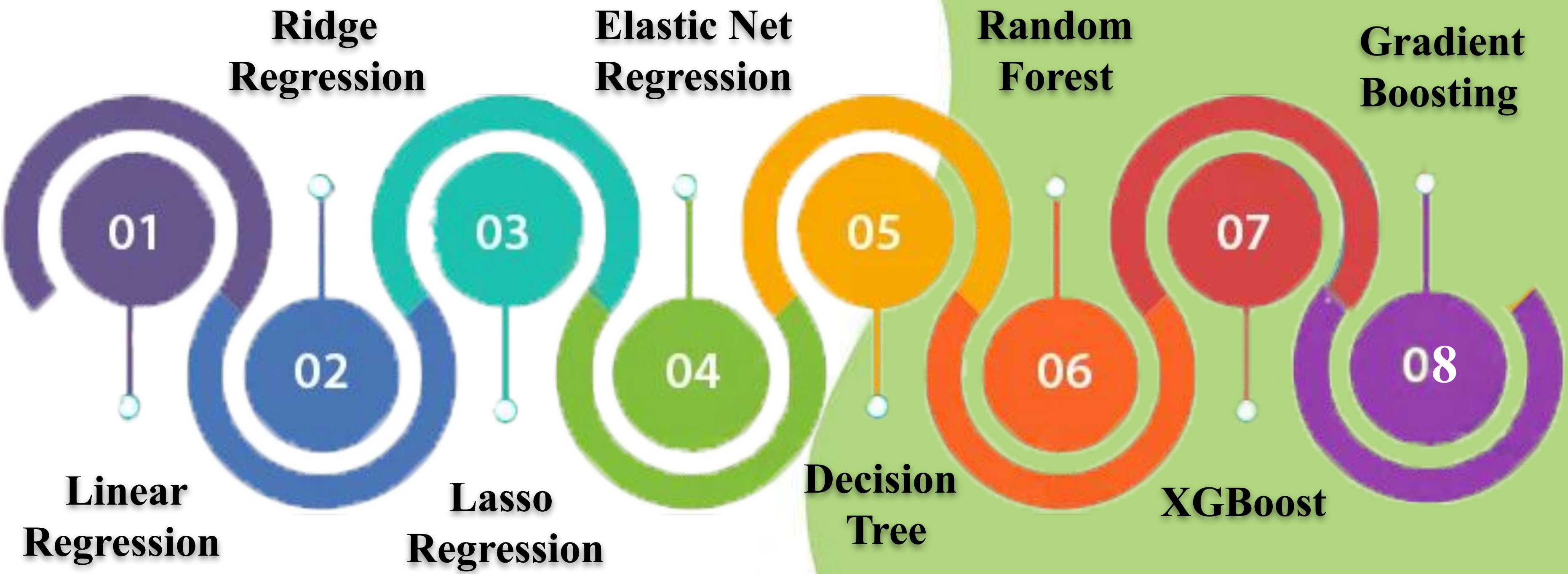




Data Preparation

- ❖ Now that the Dataset is cleaned and we have added all the necessary features along with some conversions of categorical features via.,
  - Label Encoding
  - One Hot Encoding
- ❖ Then, We used MinMaxscaler for transforming data
- ❖ So, now we have split the data into training and testing sets.
  - Train Test Split (Test size = “0.2”   Random state = “42”)

Types of Models Used



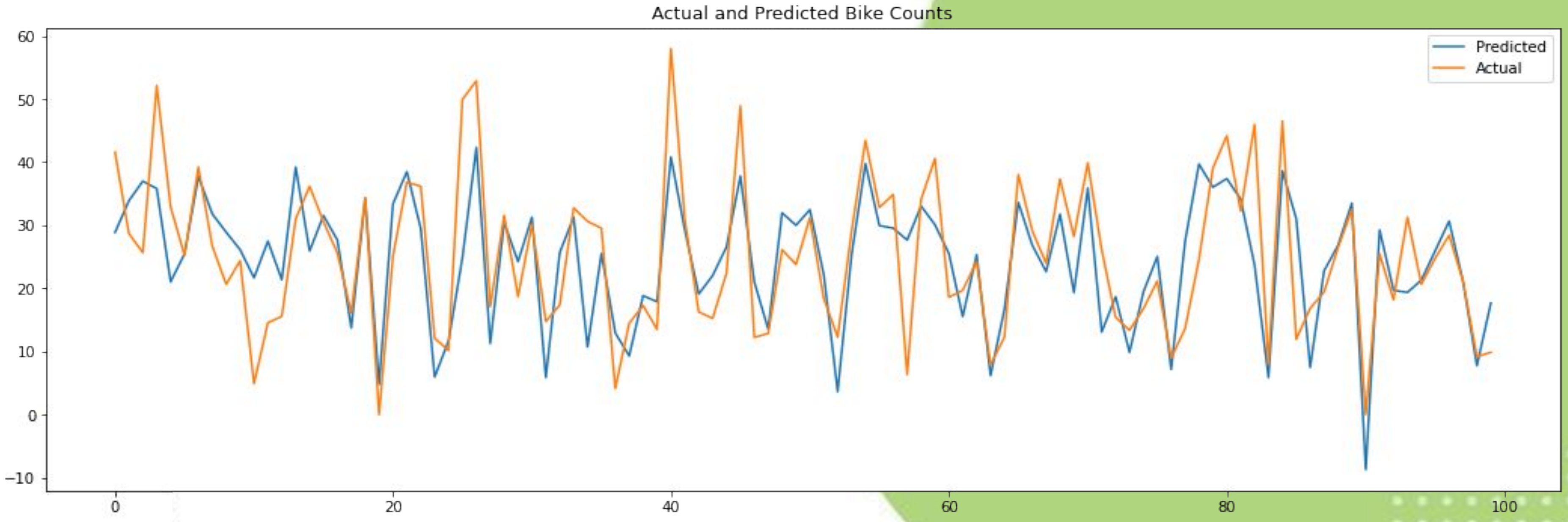
Linear Regression

Train Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
5.612	**53.298**	**7.301**	**0.657**	**0.65**

Test Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
5.692	**56.054**	**7.487**	**0.636**	**0.63**



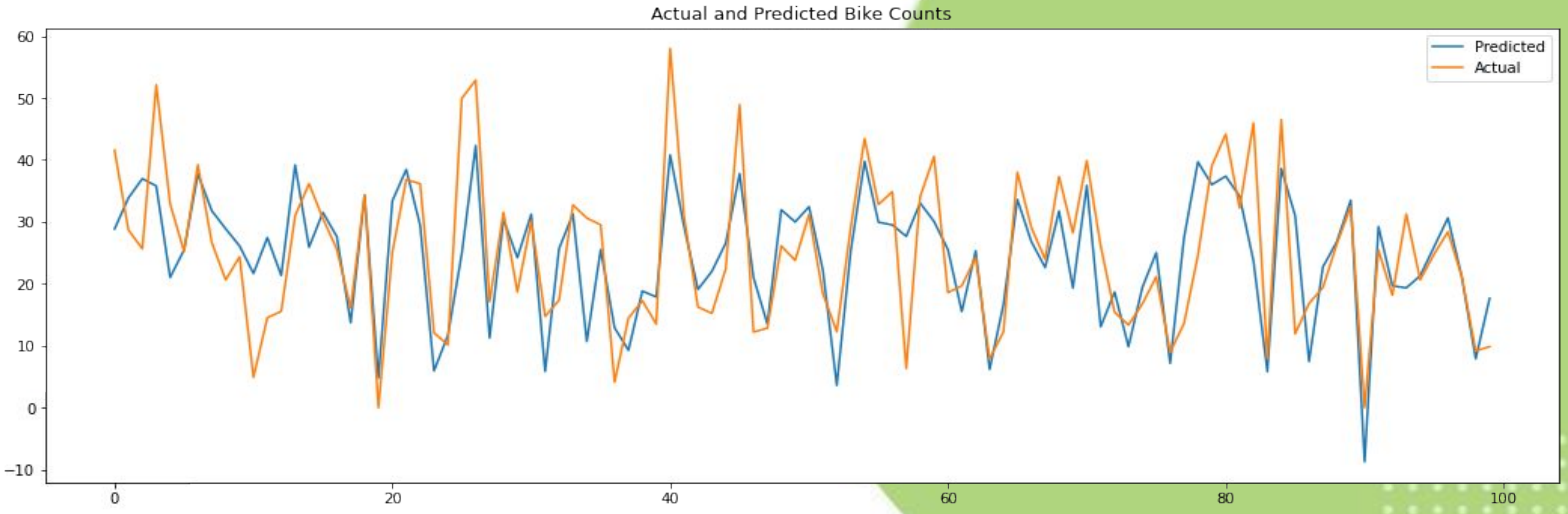
Ridge Regression

Train Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
5.613	**53.298**	**7.301**	**0.657**	**0.65**

Test Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
5.692	**56.024**	**7.485**	**0.636**	**0.63**



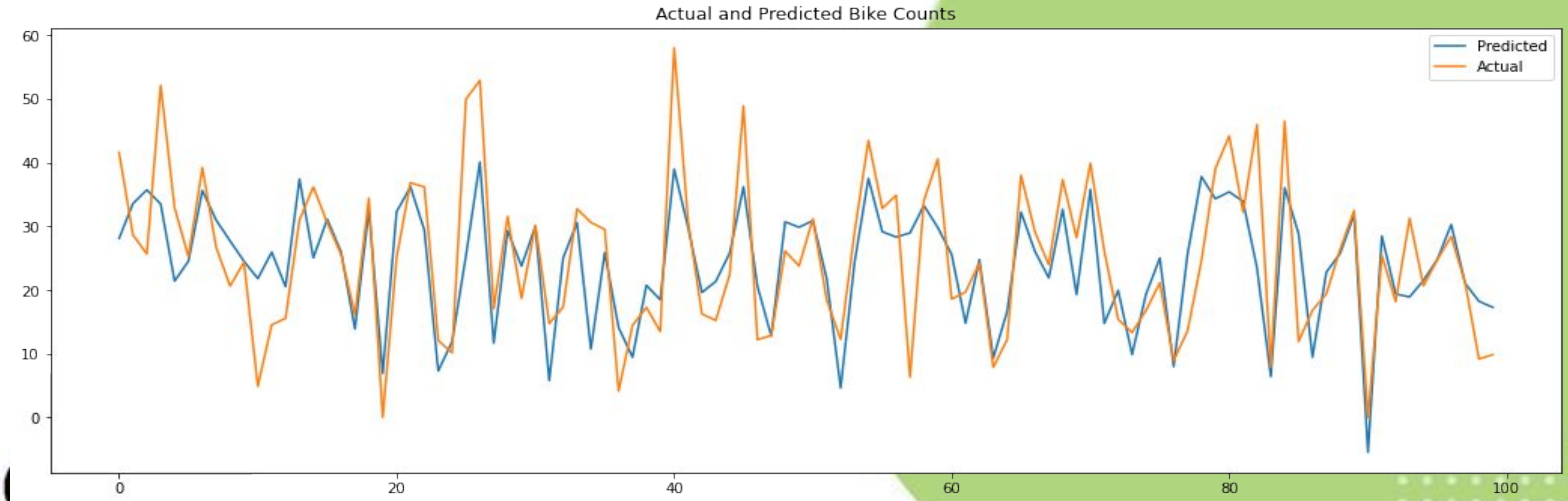
Lasso Regression

Train Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
5.847	57.825	7.604	0.628	0.62

Test Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
5.899	59.921	7.741	0.611	0.61



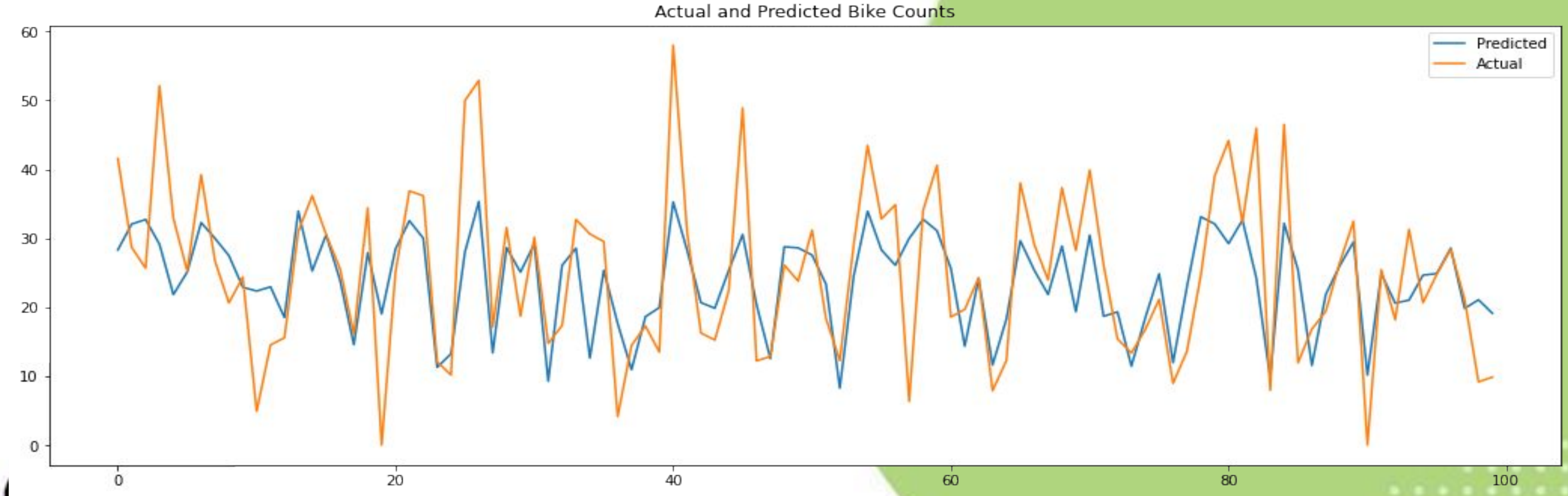
Elastic Net Regression

Train Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
6.816	76.037	8.720	0.510	0.51

Test Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
6.821	77.227	8.788	0.498	0.49



Optimization

- Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set.

- GridSearchCV is a technique to search through the best parameter values from the given set of the grid of parameters. It is basically a cross-validation method.

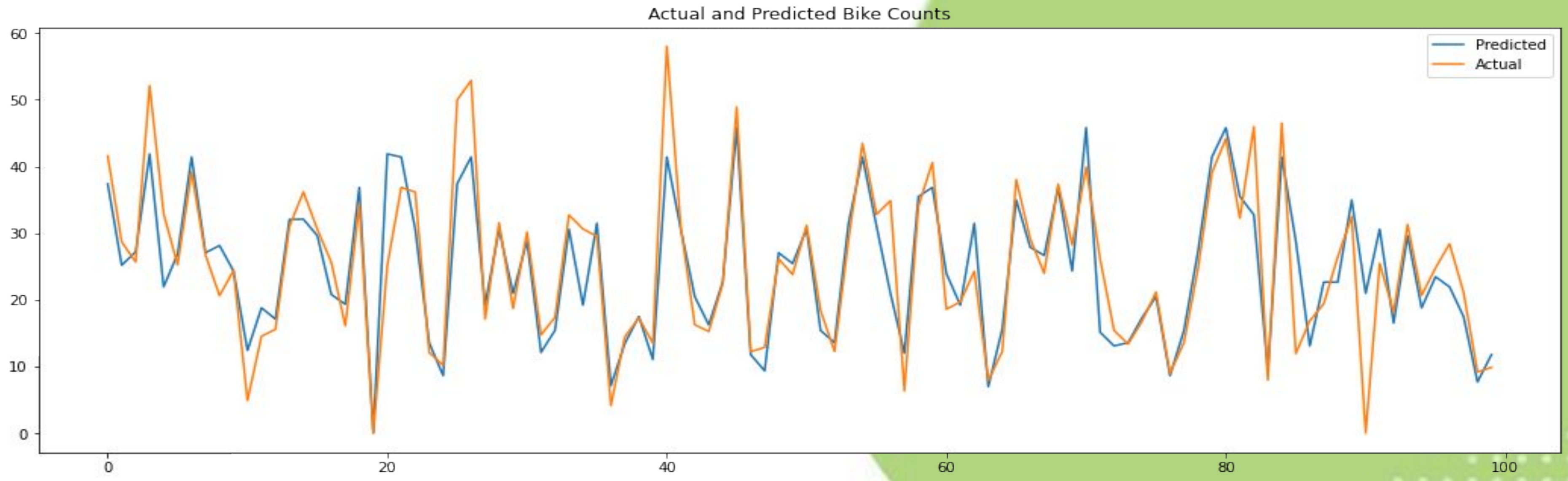
Decision Tree

Train Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
3.337	23.271	4.824	0.850	0.85

Test Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
3.579	27.385	5.233	0.822	0.82



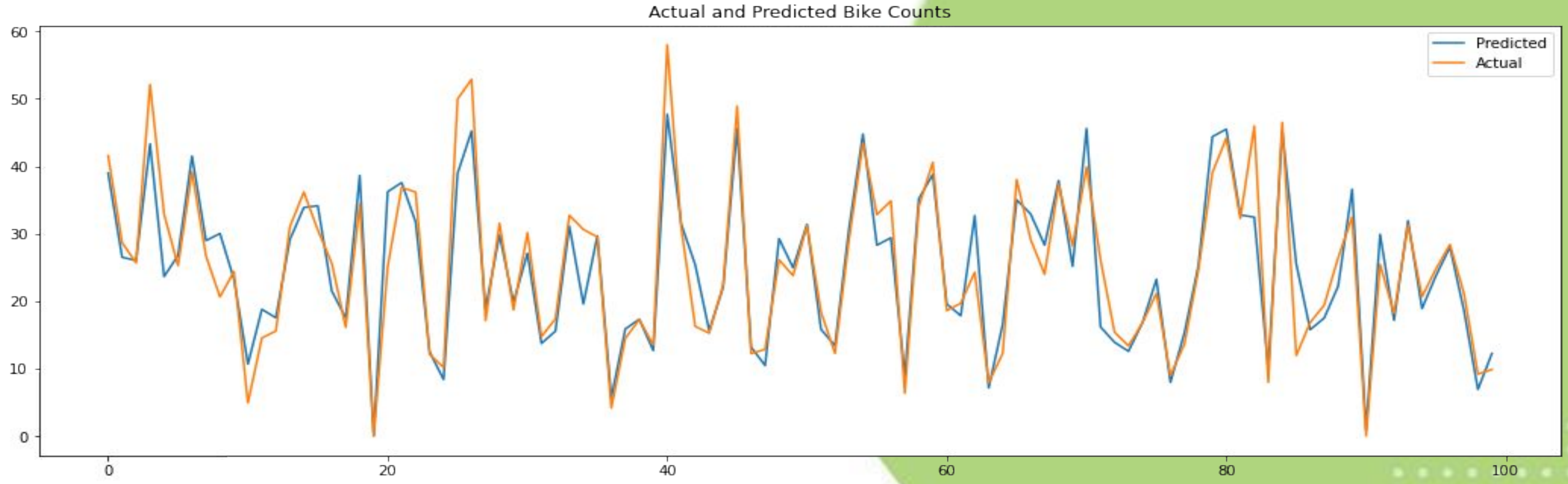
Random Forest

Train Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
1.998	**8.275**	**2.877**	**0.947**	**0.95**

Test Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
2.741	**16.216**	**4.027**	**0.895**	**0.89**



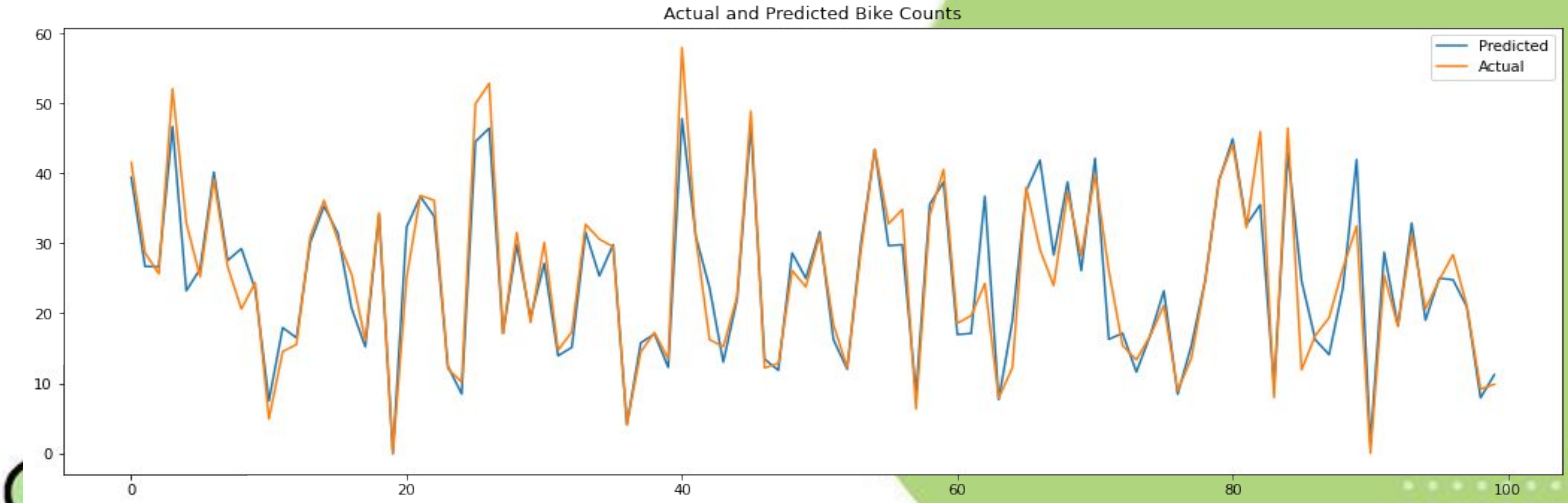
XGBoost

Train Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
1.022	**2.255**	**4.824**	**0.985**	**0.99**

Test Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
2.336	**13.077**	**4.824**	**0.915**	**0.91**



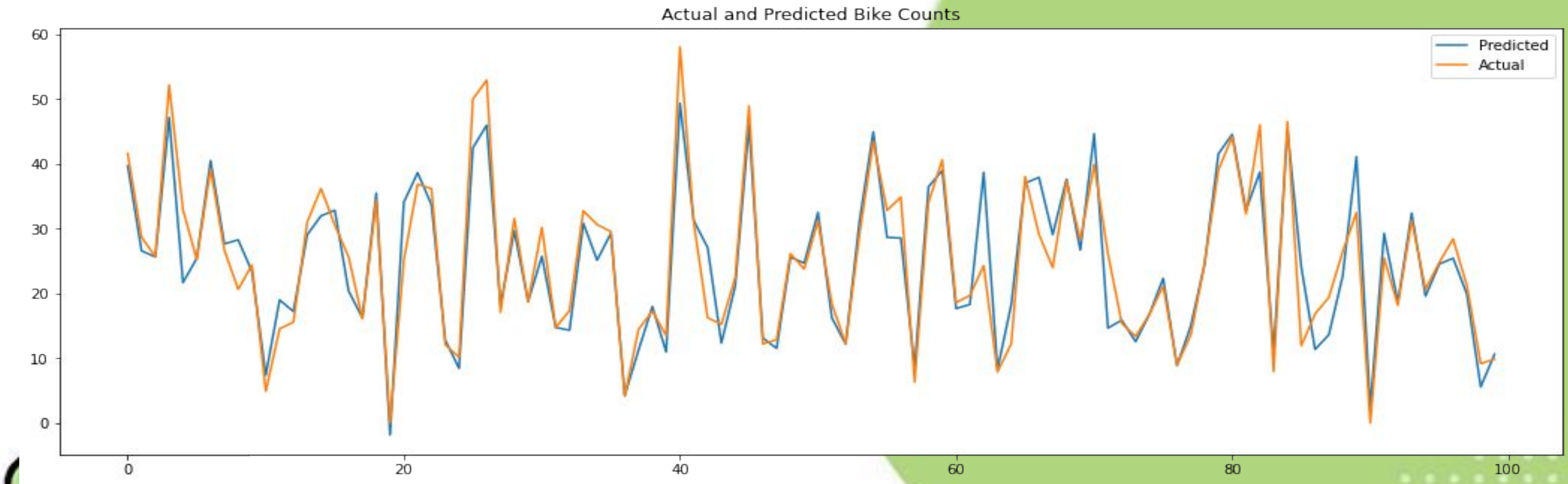
Gradient Boosting

Train Set Metrics

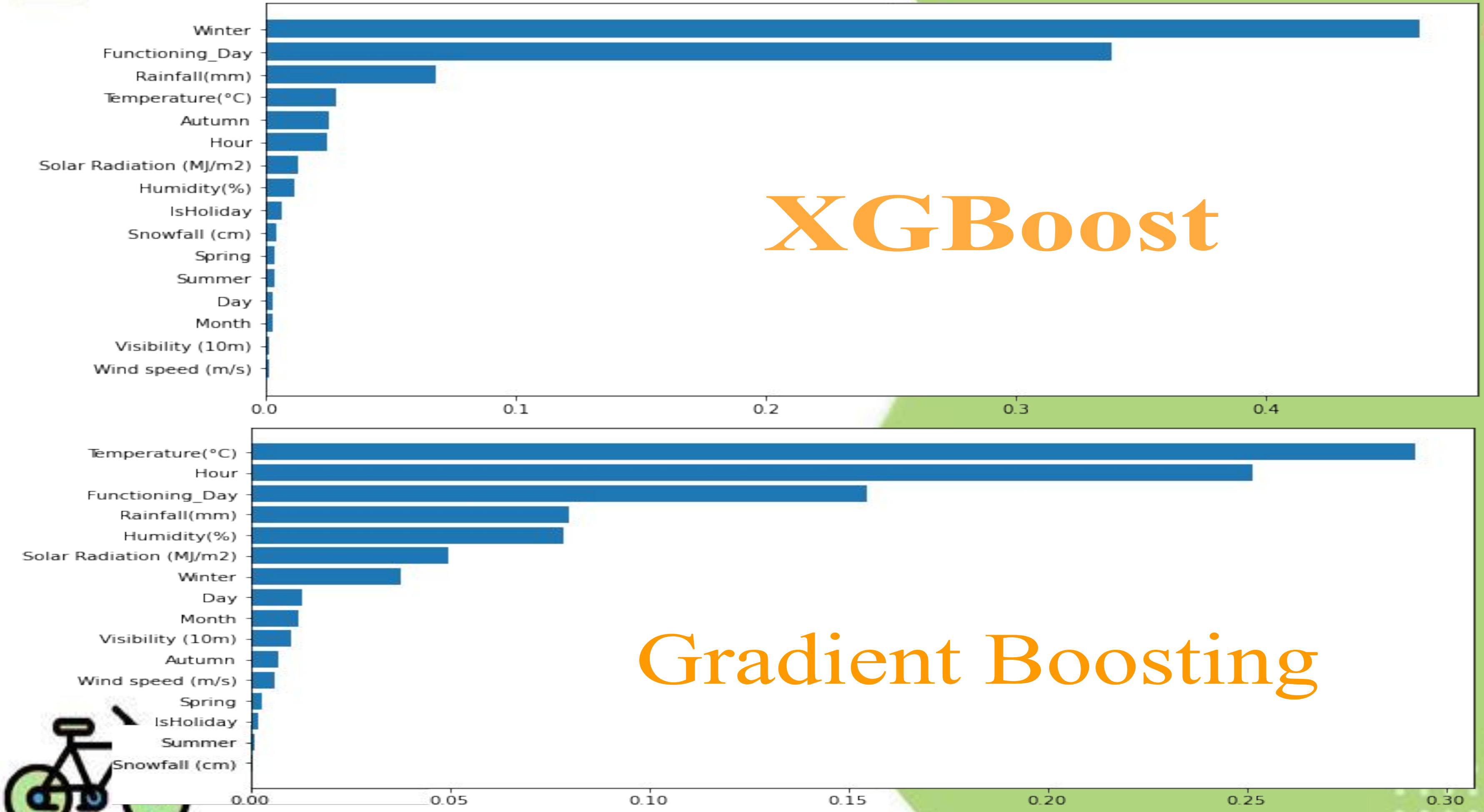
MAE	MSE	RMSE	R2 score	Adjusted R2
1.593	**5.353**	**4.824**	**0.966**	**0.97**

Test Set Metrics

MAE	MSE	RMSE	R2 score	Adjusted R2
2.418	**12.974**	**4.824**	**0.916**	**0.91**



Feature Importance



Evaluation Metrices For All Models

Test Set Metrices

	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear regression	5.692	56.054	7.487	0.636	0.63
1	Ridge regression	5.692	56.024	7.485	0.636	0.63
2	Lasso regression	5.899	59.921	7.741	0.611	0.61
3	Elastic net regression Test	6.821	77.227	8.788	0.498	0.49
4	Decision tree regression	3.579	27.385	5.233	0.822	0.82
5	Random forest regression	2.741	16.216	4.027	0.895	0.89
6	XGBoost regression	2.336	13.077	4.824	0.915	0.91
7	Gradient Boosting Regressor	2.418	12.974	4.824	0.916	0.91

Challenges

- ❖ Large Dataset to handle.
- ❖ Needs to plot lot of Graphs to analyse.
- ❖ Carefully handled Feature selection part as it affects the R2 score.
- ❖ Carefully tuned Hyperparameters as it affects the R2 score.
- ❖ Handled the positive skewness of the target variable.
- ❖ Handled the high correlation between various features.
- ❖ Need to convert categorical features into numerical features using feature engineering.

Conclusions

- **RMSE values for Test Data as lower the RMSE better the model performance:**

  Lowest RMSE values Model
  - RandomForestRegressor RMSE: **4.027**
  - XGBoost Regressor RMSE: **4.824**
  - GradientBoostingRegressor RMSE: **4.824**

- XGBoost Regressor and Gradient Boost Regressor gives the highest R2 score of 98% and 96% respectively for Train dataset and 91% for both regressor's Test dataset. So, We can deploy these models.

- The Temperature, Hour & Functioning Day are the most important features that positively drive the total rented bikes count.

- In conclusion, the demand prediction for the given Seoul bike sharing dataset can be accurately predicted using XGBoost Regressor and Gradient Boost Regressor.

Thank You