Capstone Project-2

Bike Sharing Demand Prediction

By

Neeraja C

Data Science Trainee|AlmaBetter

.Problem Statement

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. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

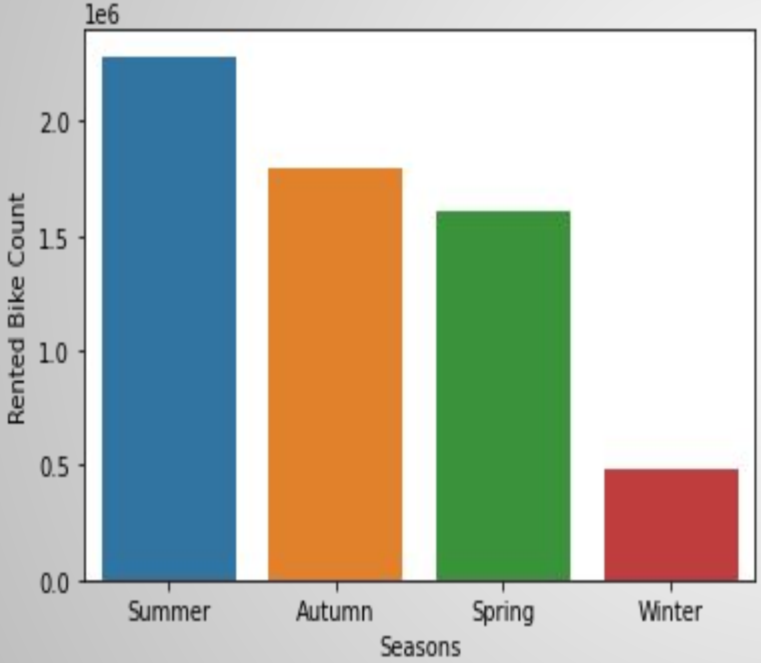
Column Details

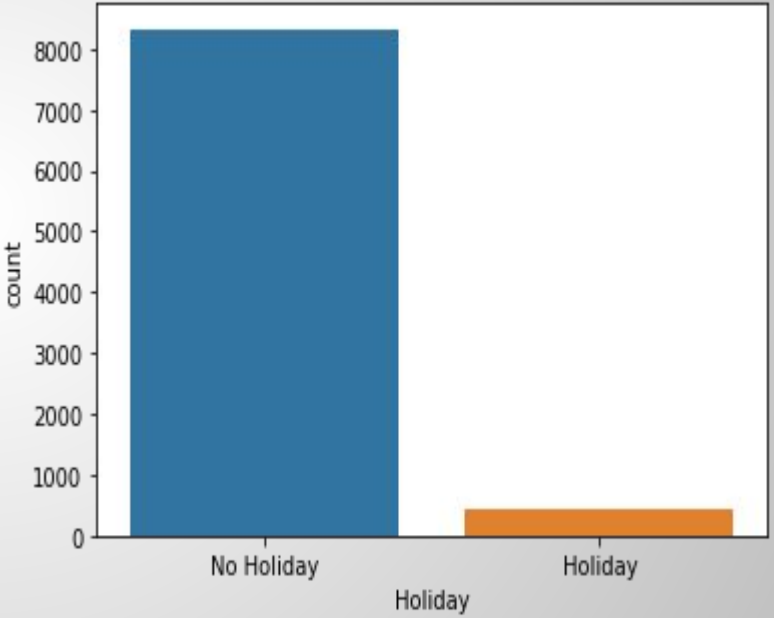
The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

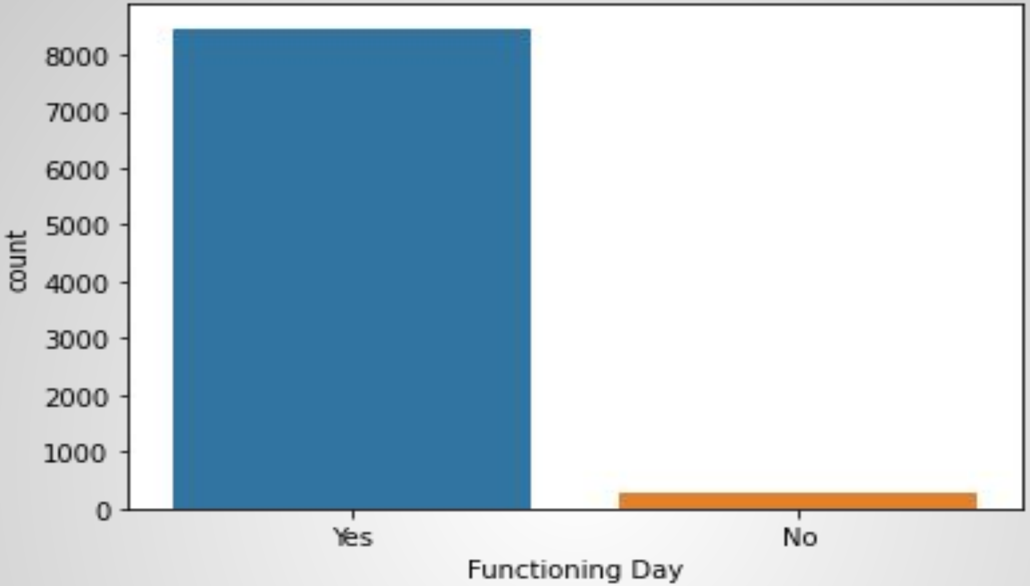
Attribute Information:
Date: year-month-day
Rented Bike count - Count of bikes rented at each hour
Hour - Hour of the day
Temperature-Temperature in Celsius
Humidity - %
Windspeed - m/s
Visibility - 10m

Dew point temperature - Celsius
Solar radiation - MJ/m2
Rainfall - mm
Snowfall - cm
Holiday - Holiday/No holiday
Functional Day – (Non Functional Hours), Fun(Functional hours)

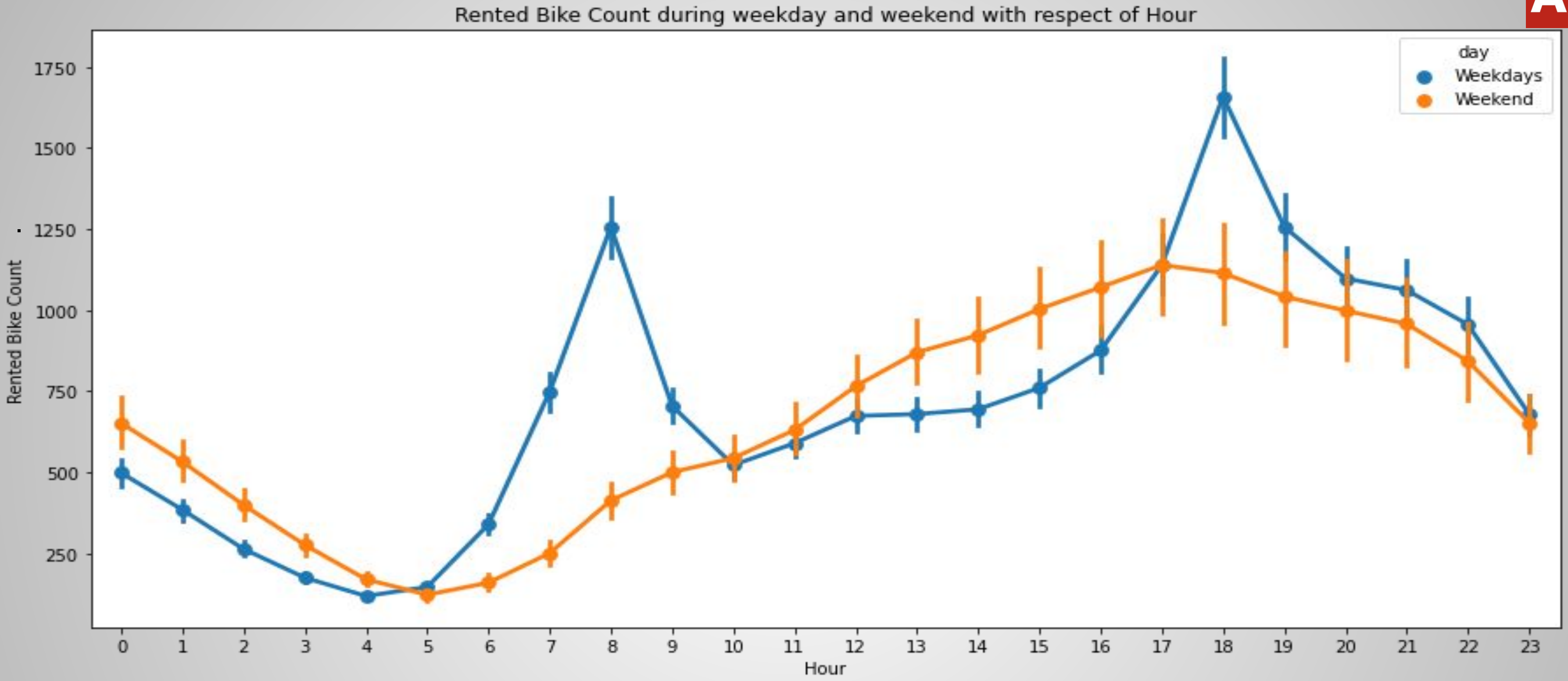
EDA


1e6
Rented Bike Count
0.0
0.5
1.0
1.5
2.0
Summer
Autumn
Spring
Winter
Seasons


count
0
1000
2000
3000
4000
5000
6000
7000
8000
No Holiday
Holiday
Holiday



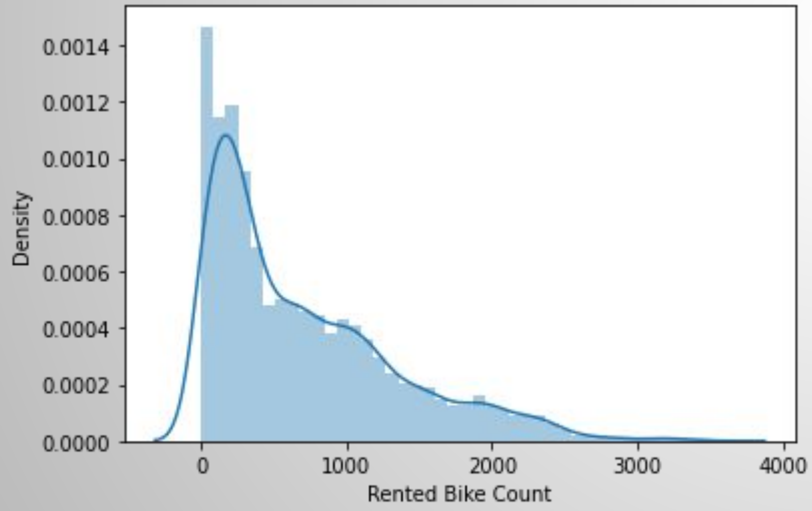
- There are high rental demand in summer.
- Rental count in working day is much more than that in holiday.

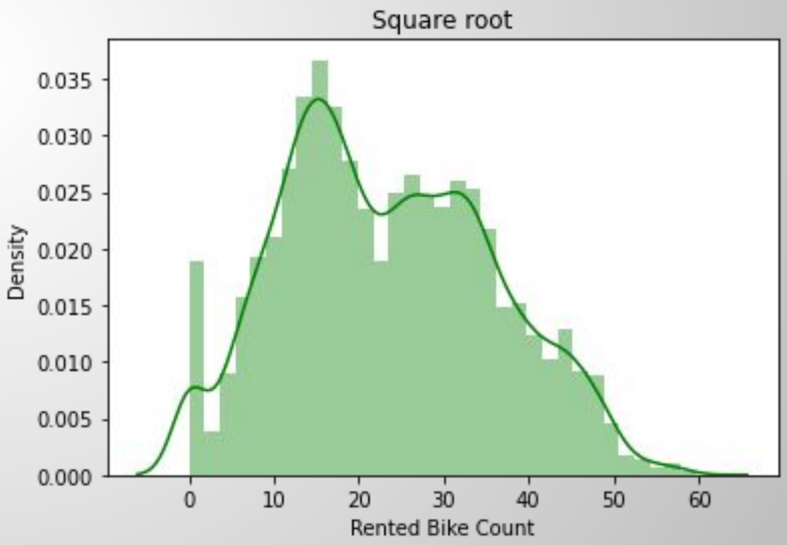


The demand is high in the afternoon on the weekend. While there is more demand during office hours in weekdays.

Normalization :

It is a technique applied during data preparation to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges or losing information. Here I have applied square root on our dependent variable.



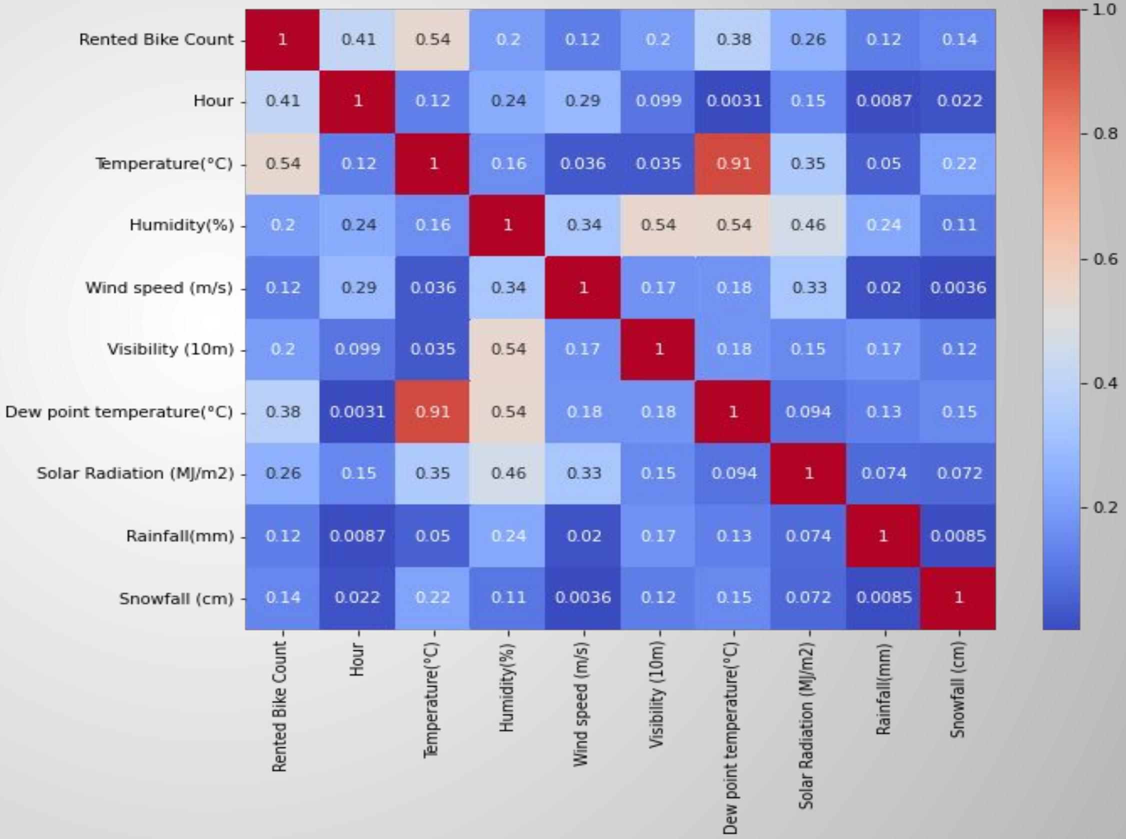


Feature Selection

Correlation Heatmap

It is used to understand which variables are related to each other and the strength of this relationship.

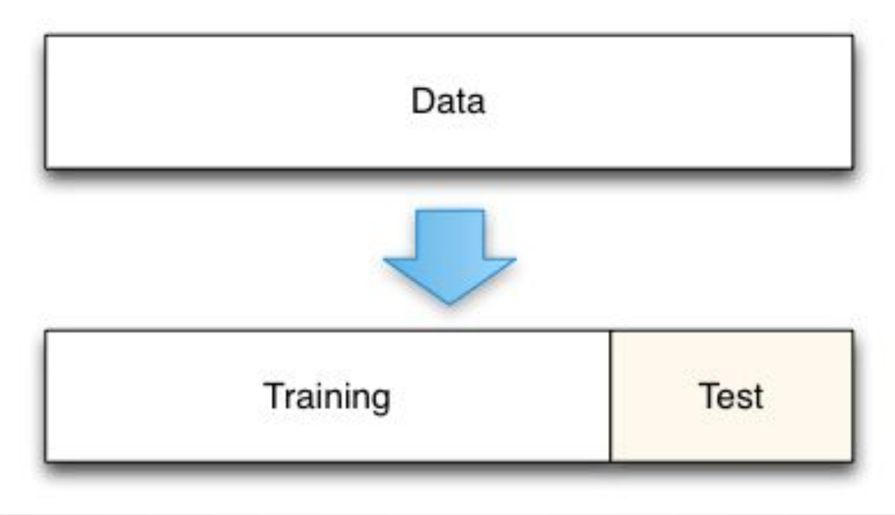
Here we can see that dew point temperature is highly correlated to Temperature.So we can remove that feature.



	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)
Rented Bike Count	1	0.41	0.54	0.2	0.12	0.2	0.38	0.26	0.12	0.14
Hour	0.41	1	0.12	0.24	0.29	0.099	0.0031	0.15	0.0087	0.022
Temperature(°C)	0.54	0.12	1	0.16	0.036	0.035	0.91	0.35	0.05	0.22
Humidity(%)	0.2	0.24	0.16	1	0.34	0.54	0.54	0.46	0.24	0.11
Wind speed (m/s)	0.12	0.29	0.036	0.34	1	0.17	0.18	0.33	0.02	0.0036
Visibility (10m)	0.2	0.099	0.035	0.54	0.17	1	0.18	0.15	0.17	0.12
Dew point temperature(°C)	0.38	0.0031	0.91	0.54	0.18	0.18	1	0.094	0.13	0.15
Solar Radiation (MJ/m2)	0.26	0.15	0.35	0.46	0.33	0.15	0.094	1	0.074	0.072
Rainfall(mm)	0.12	0.0087	0.05	0.24	0.02	0.17	0.13	0.074	1	0.0085
Snowfall (cm)	0.14	0.022	0.22	0.11	0.0036	0.12	0.15	0.072	0.0085	1

Train-Test Split

(8760, 25)

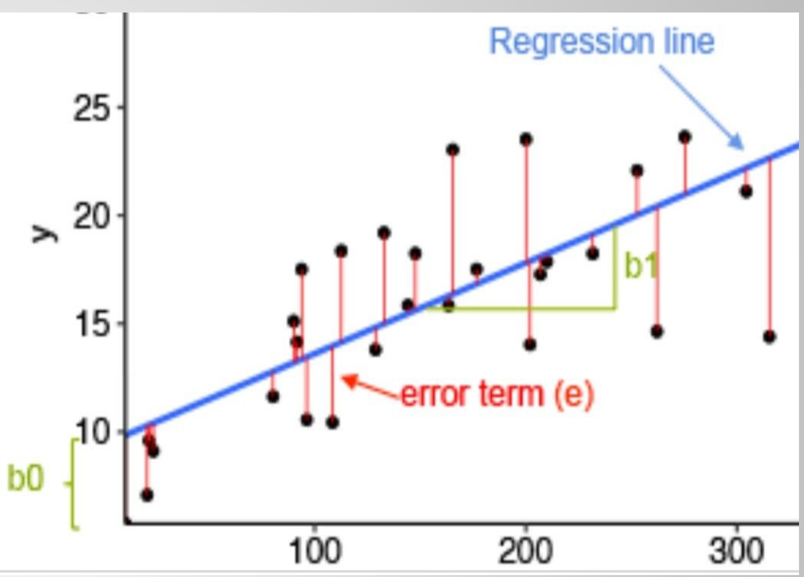


(6570, 25)

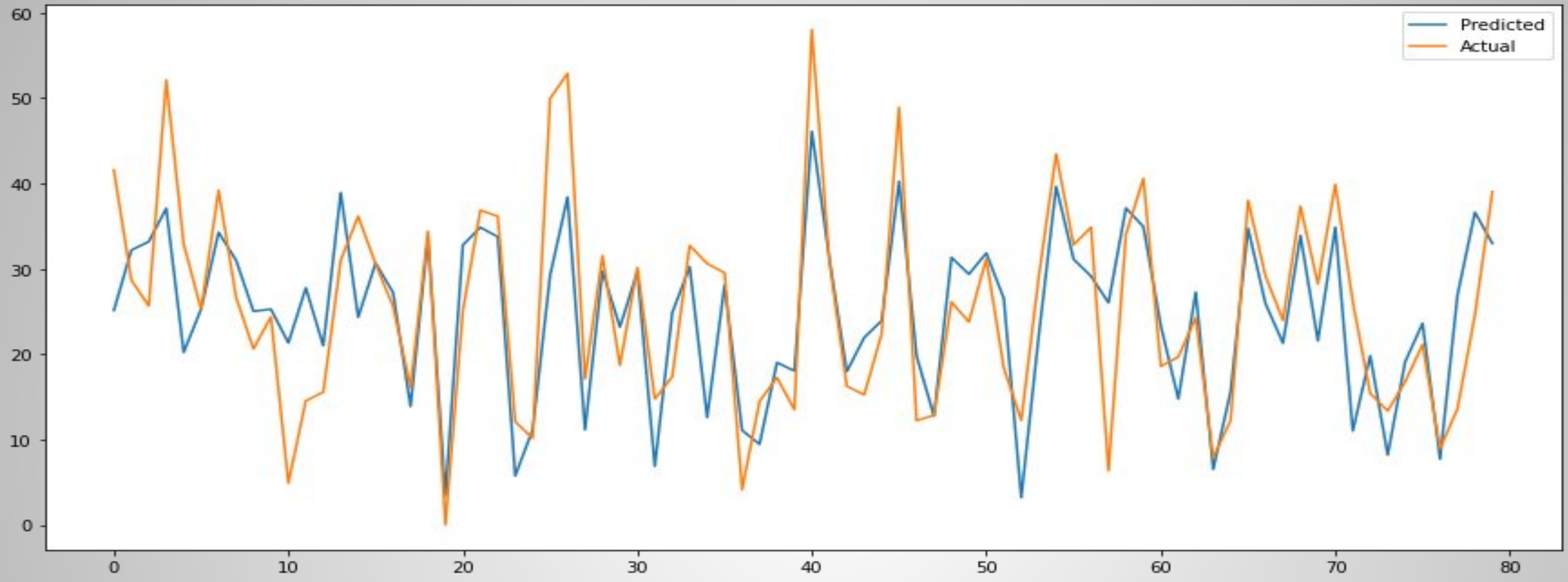
(2190, 25)

Linear Regression

Linear Regression is an algorithm that belongs to supervised Machine Learning. It tries to apply relations that will predict the outcome of an event based on the independent variable data points. The relation is usually a straight line that best fits the different data points as close as possible. The output is of a continuous form, i.e., numerical value. By applying this model I got the value of R Square as **0.52.**



R-Squared determines the proportion of variance in the dependent variable that can be explained by the independent variable. The higher the R-Squared, the better the model fits your data.



MSE: It measures how close a regression line is to a set of data ponts. Here I got MSE as 52.13.

MAE: It is the arithmetic average of absolute errors. Here MAE = 5.41.

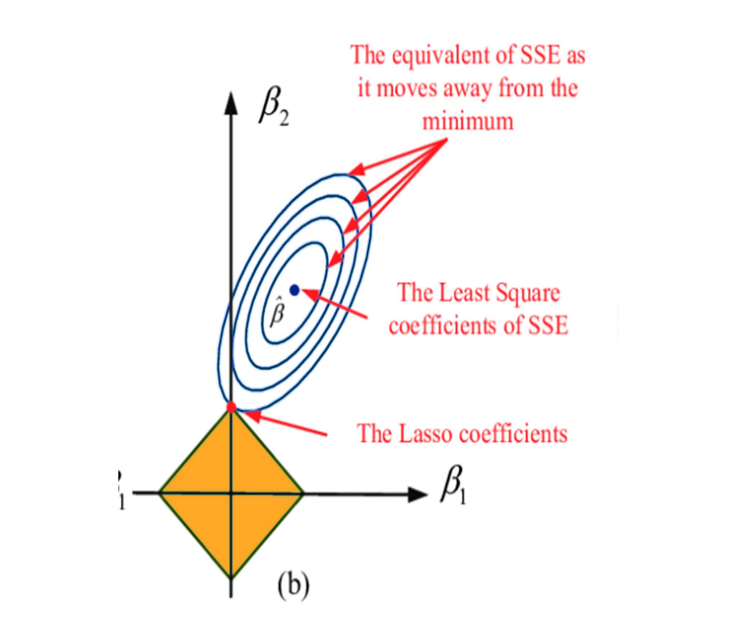
The closer the value of the MAE is to 0, the better.

Lasso Regression

Lasso regression is a regularization technique used over regression methods for a more accurate prediction. It allows us to shrink or regularize coefficients to avoid overfitting and make them work better on different datasets.

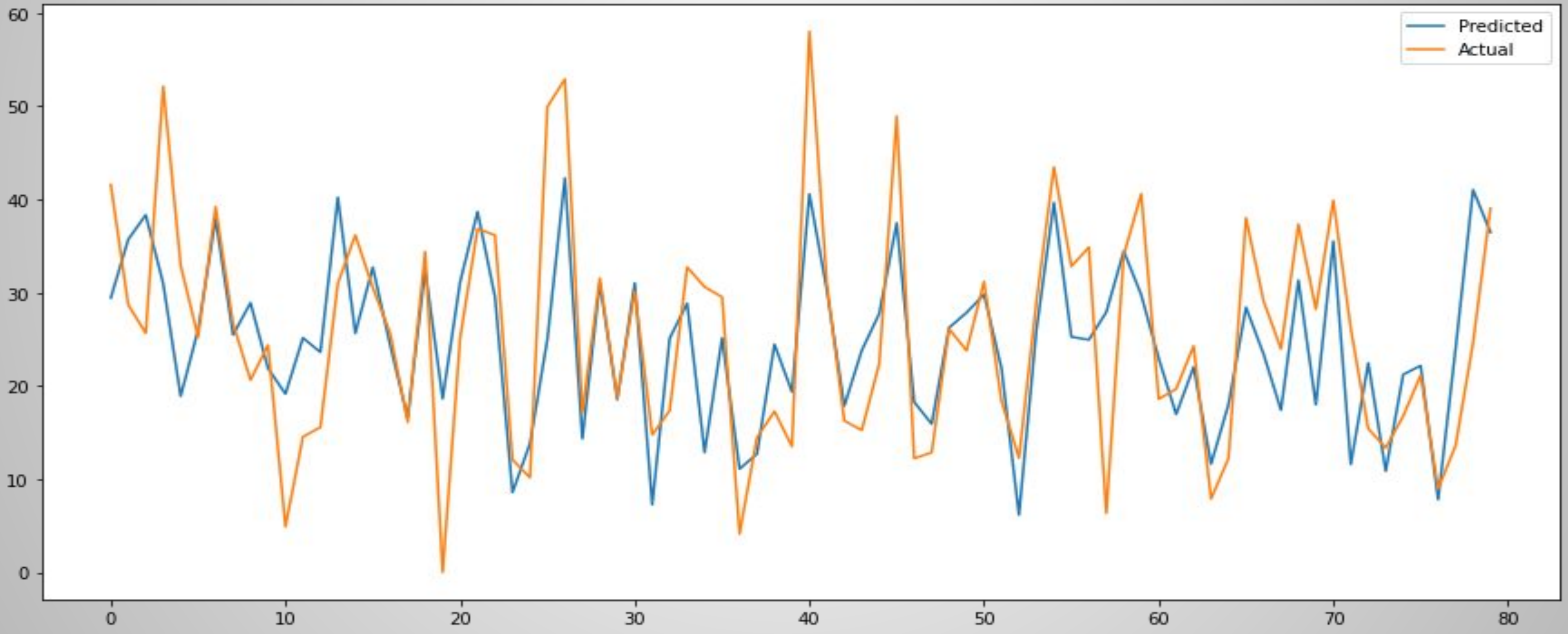
By applying this model, i got very low accuracy.

R squared Value on test set : 0.14
Mean Absolute error on test set : 6.24
Mean squared error : 65.94



β_2
The equivalent of SSE as it moves away from the minimum
$\hat{\beta}$
The Least Square coefficients of SSE
The Lasso coefficients
β_1
(b)

The graph showing Actual value and Predicted value

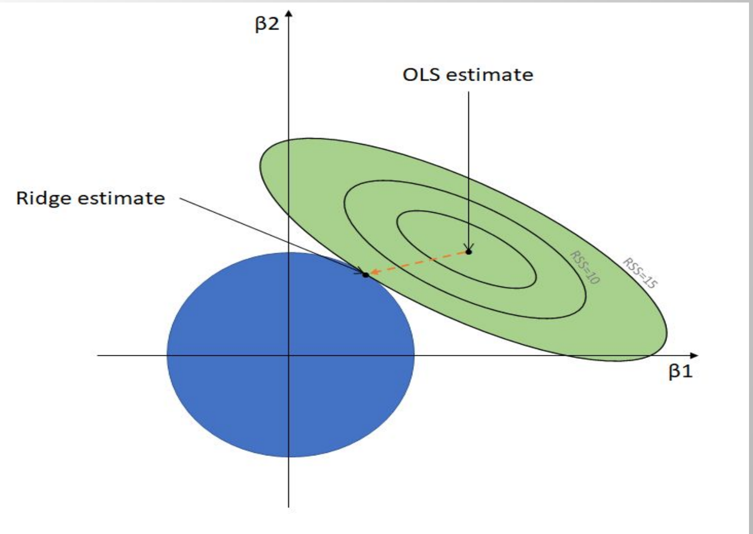


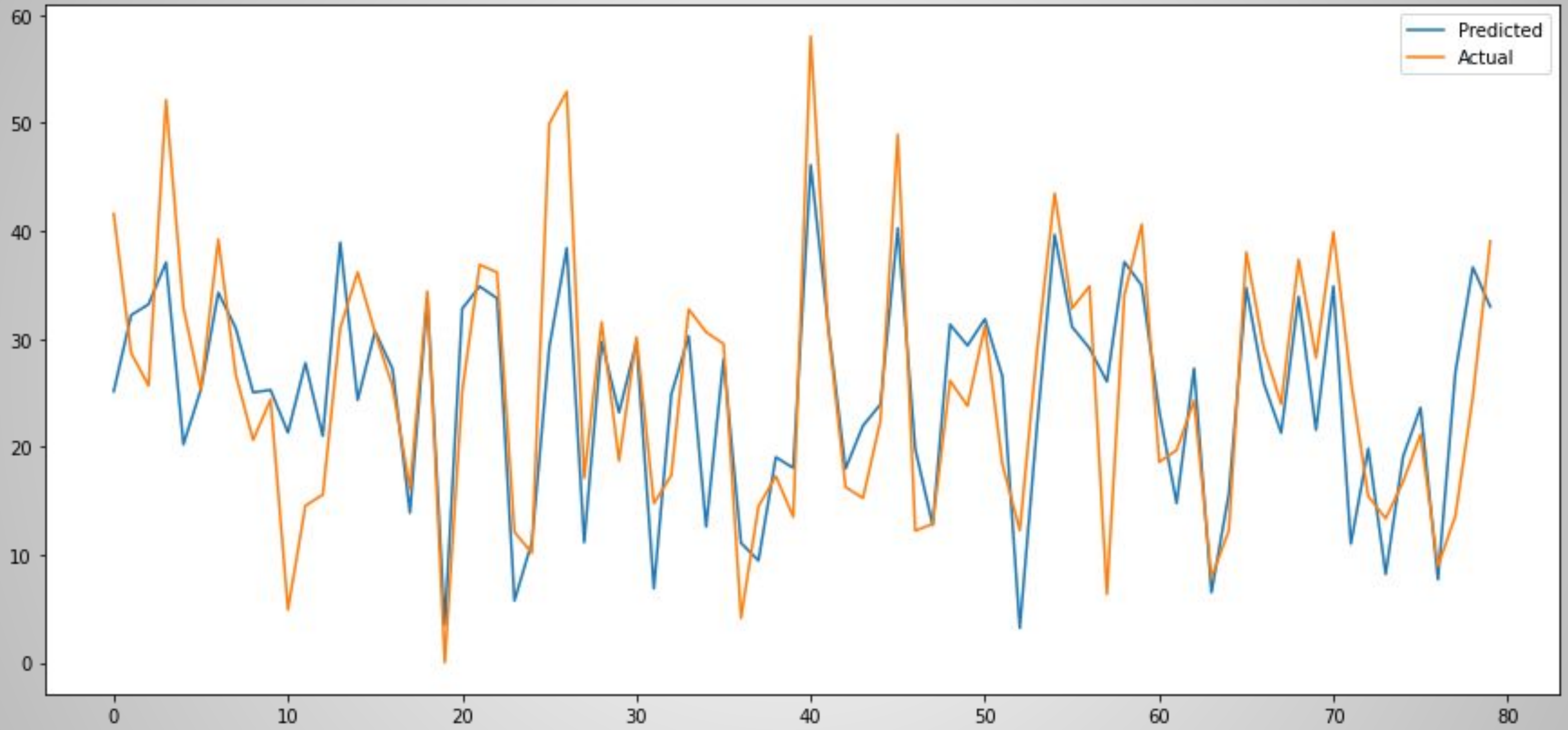
Ridge Regression

A Ridge regressor is basically a regularized version of a Linear Regressor. i.e to the original cost function of linear regressor we add a regularized term that forces the learning algorithm to fit the data and helps to keep the weights lower as possible. The regularized term has the parameter ‘alpha’ which controls the regularization of the model i.e helps in reducing the variance of the estimates.

Evaluation Metrics

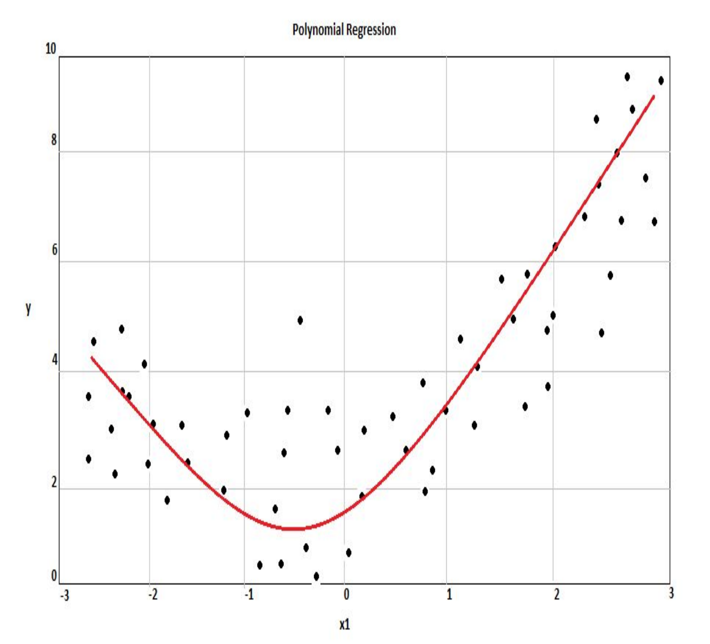
R squared Value on test set : 0.52
Mean Absolute error on test set : 5.4
Mean squared error : 52.12




60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
Predicted
Actual

Polynomial Regression

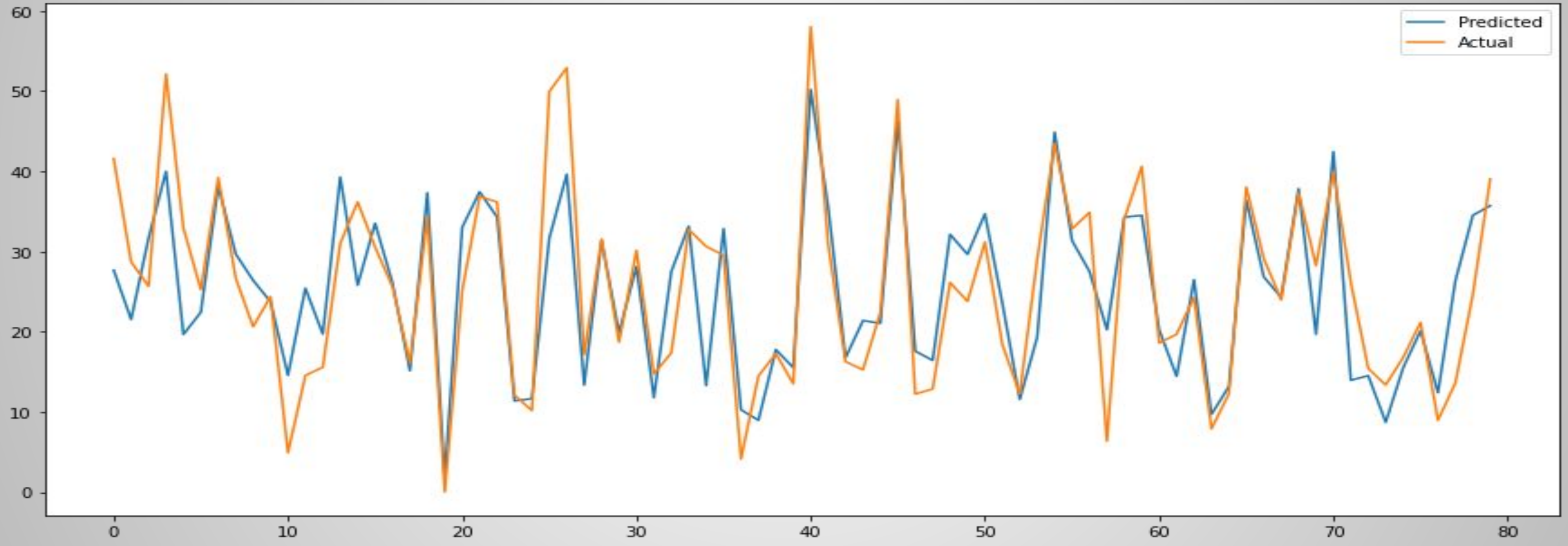
Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression.



Evaluation Metrics

R squared Value on test set : 0.60
Mean Absolute error on test set : 4.48
Mean squared error : 51.57

The graph showing Actual value and Predicted value of our model.



Decision Tree

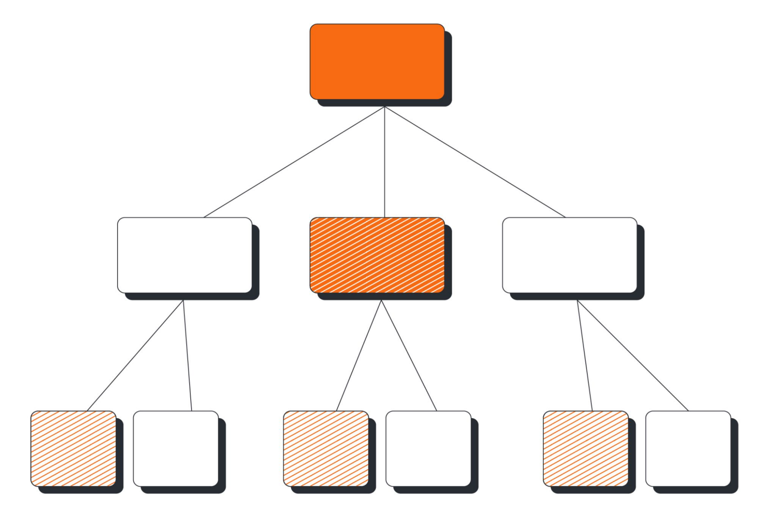
A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

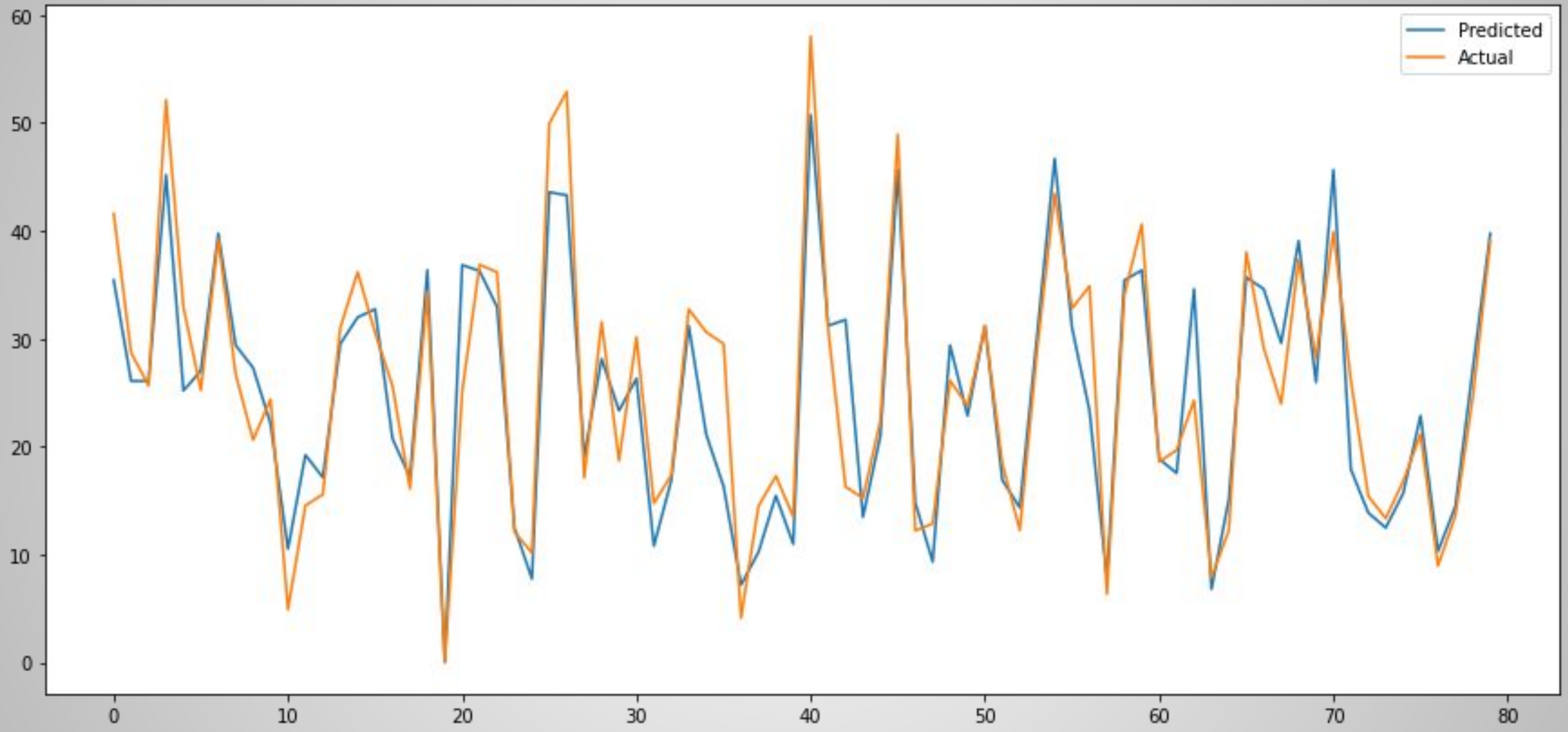
Here I got a better model because multi-collinearty deos not affect tree base models.

R squared Value on test set : 0.86

Mean Absolute error on test set : 3.16

Mean squared error : 21




Predicted
Actual
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80

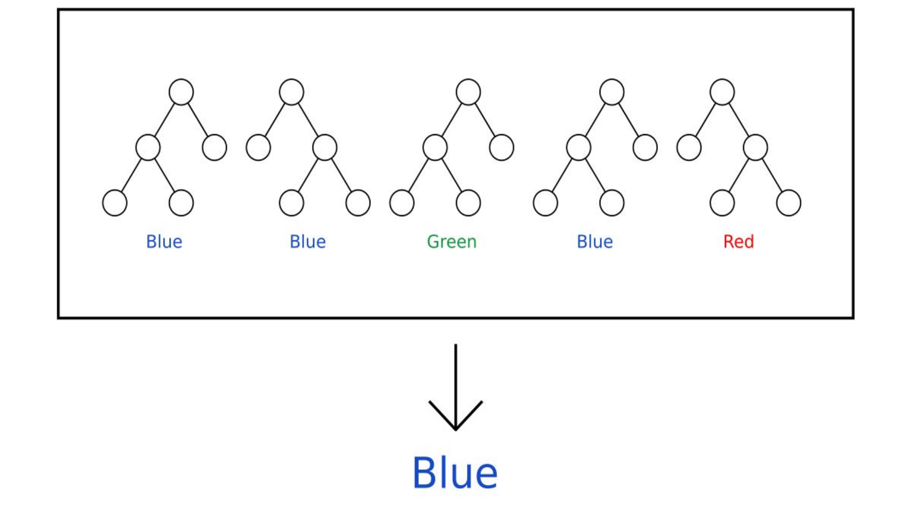
Random Forest

Random Forest is a supervised machine learning algorithm made up of decision trees
It is used for both classification and regression.

RandomizedSearchCV

RandomizedSearchCV is very useful when we have many parameters to try and the training time is very long.

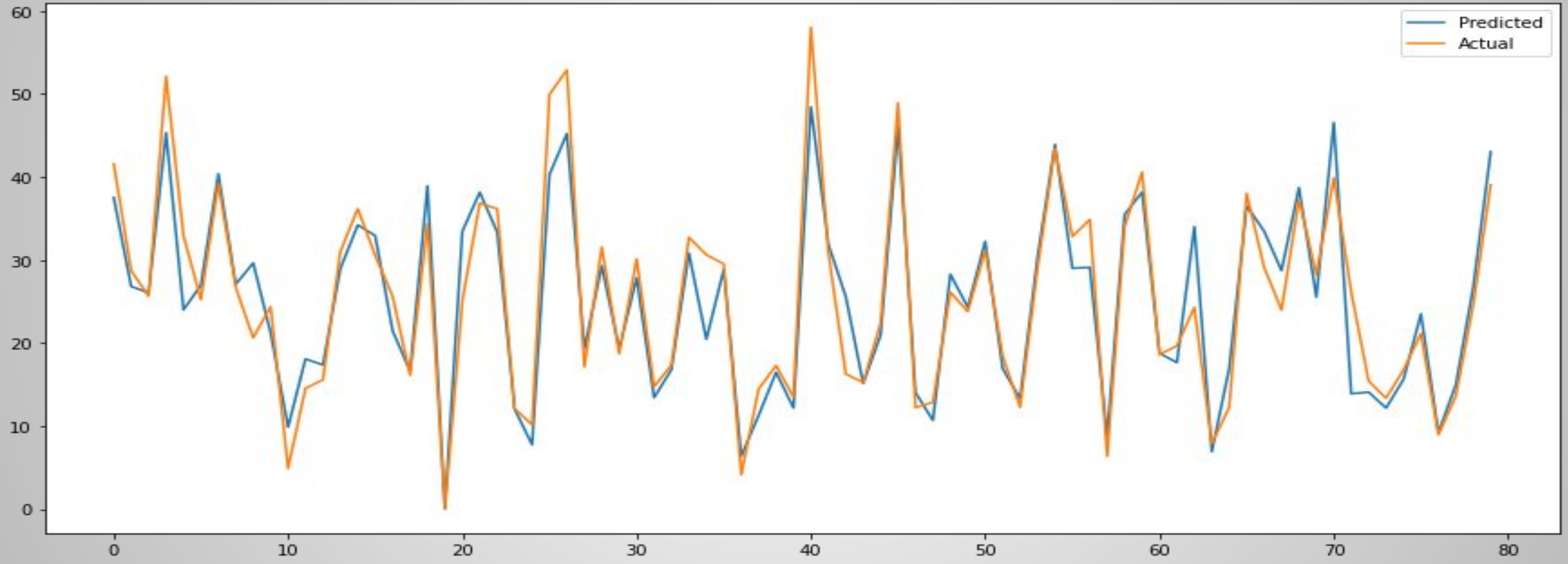
If the number of parameters to consider is particularly high and the magnitudes of influence are imbalanced, the better choice is to use the Random Search.



R squared Value on test set : 0.89

Mean Absolute error on test set : 2.68

Mean squared error : 15.58

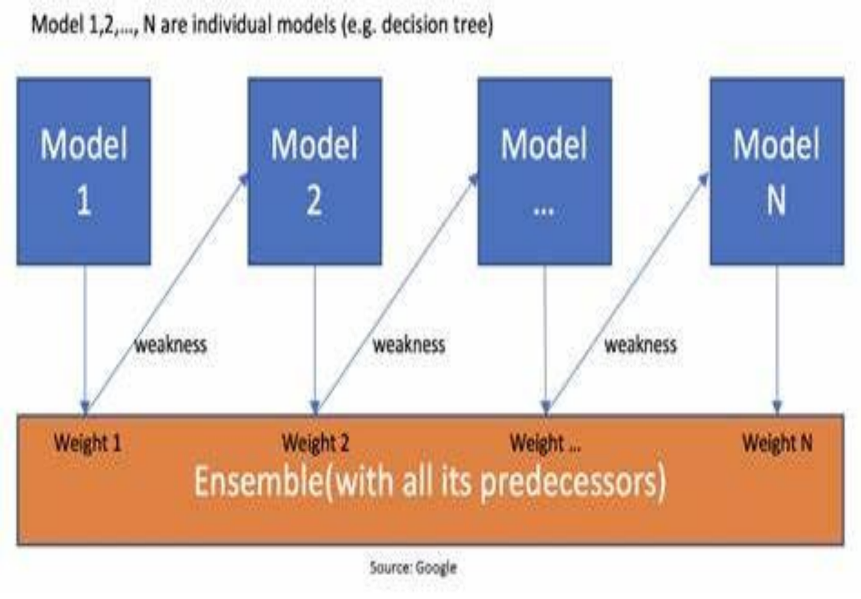


AdaBoost Regressor

AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split.

Grid Search

Grid Search is an effective method for adjusting the parameters in supervised learning and improve the generalization performance of a model. With Grid Search, we try all possible combinations of the parameters of interest and find the best ones.

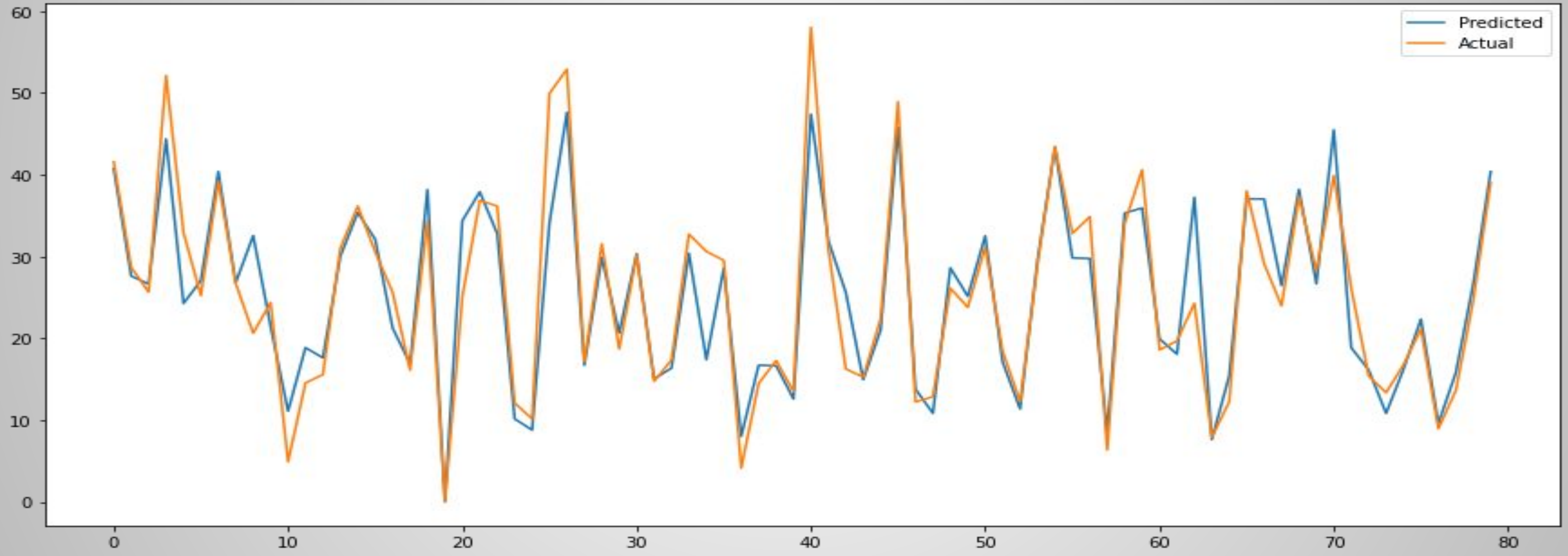

Model 1,2,..., N are individual models (e.g. decision tree)
Model 1
Model 2
Model ...
Model N
weakness
weakness
weakness
Weight 1
Weight 2
Weight ...
Weight N
Ensemble(with all its predecessors)

Source: Google

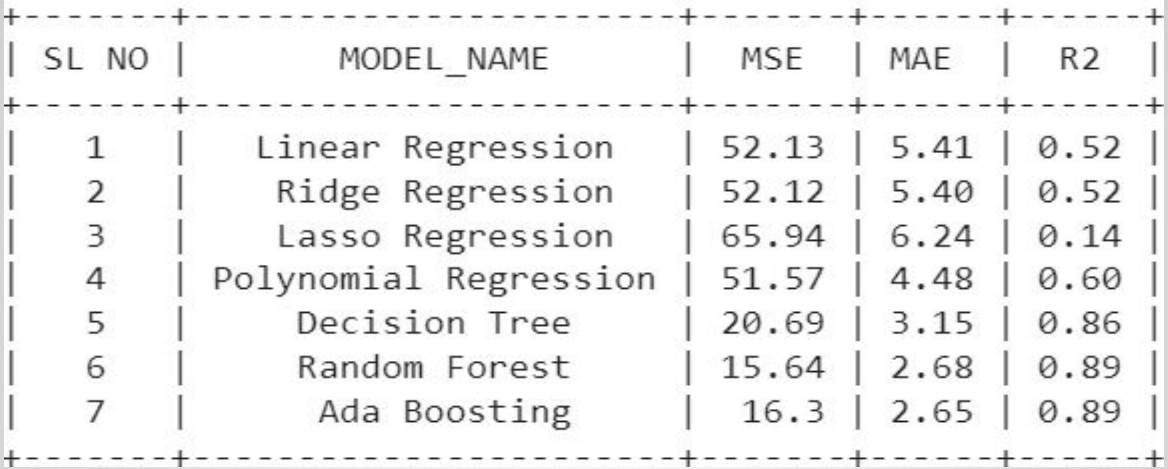
R squared Value on test set : 0.89

Mean Absolute error on test set : 2.62

Mean squared error : 16.06



Conclusions



```
+-------+---------------------+-------+------+------+
| SL NO |      MODEL_NAME     |  MSE  | MAE  |  R2  |
+-------+---------------------+-------+------+------+
|   1   |  Linear Regression  | 52.13 | 5.41 | 0.52 |
|   2   |   Ridge Regression  | 52.12 | 5.40 | 0.52 |
|   3   |   Lasso Regression  | 65.94 | 6.24 | 0.14 |
|   4   | Polynomial Regression | 51.57 | 4.48 | 0.60 |
|   5   |    Decision Tree    | 20.69 | 3.15 | 0.86 |
|   6   |    Random Forest    | 15.64 | 2.68 | 0.89 |
|   7   |     Ada Boosting    |  16.3 | 2.65 | 0.89 |
+-------+---------------------+-------+------+------+
```

- Ensemble techniques are performing well for this regression problem.
- Random Forest and AdaBoost are performing well. Both have an accuracy of 89%.
- In the case of regularization, Lasso is not performing well.

Conclusion

- First, I have gone through simple but effective pre-processing steps and then dig deeper into the data and apply various machine learning regression techniques like Decision Trees, Random Forest and Ada boost regressor.
- During EDA we used plotly, seaborn and matplotlib to do the visualizations. During the data pre-processing part, we converted features into numeric ones.
- I have find the best hyperparameters by parameter tuning using GridSearchCV and RandomSearcgCV.
- I have used different models like Linear Regression, Ridge, Lasso, Polynomial, Decision Tree and ensemble techniques such as Random Forest and AdaBoost.
- Random Forest and AdaBoost are performing well. Both have an accuracy of 89% and have less error rate.