CAPSTONE PROJECT - II

Bike Sharing Demand Prediction

(SEOUL BIKE PREDICTION)

By:

Shrinidhi Choragi

Data Science Trainee

Almabetter

Contents

- Introduction
- Problem statement
- Dataset
- Methodology
  - Exploratory Data Analysis
  - Regression Analysis
  - Data Modelling
  - Evaluation
  - Results
- Conclusion

Introduction

A bike rental or bike hire business rents out bicycles for short periods of time, usually for a few hours.

It is a service in which bicycles are made available for shared use to individuals on a short-term basis for a price or free.

The user enters payment information, and the computer unlocks a bike. The user returns the bike by placing it in the dock, which locks it in place.

Problem Statement

The objective of this project is to predict bike rental count/ forecast bike rental demand required at each hour based on bike usage patterns with the environmental and seasonal data history. It is a regression problem.

Some of the questions to be explored through this study:

- What is the relation between the features and the bike rental count?
- Which regressive model gives the most optimum predictions?
- What features influence the most in predicting the bike rental count?

The methodology of the project includes an exploratory data analysis, a predictive analysis using various regression algorithms and in the end, evaluating the models to decide on the most optimum model and influential features in predicting the bike rental count.

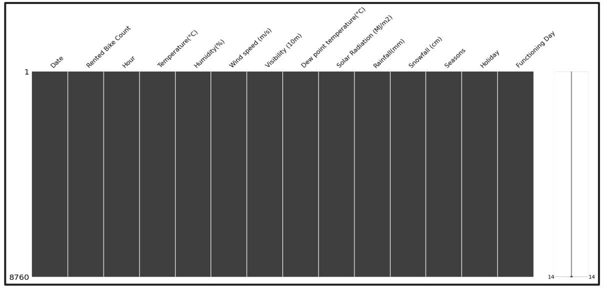
Dataset

The dataset contains **8760** observations, **13** predictors, and a target variable '**Rented Bike Count**' describing number of bikes that are rented per hour as a function of weather conditions. The predictors/features describe various environmental factors and weather information. The dataset presents the company's data between years 2017-18.

The features of the dataset are:

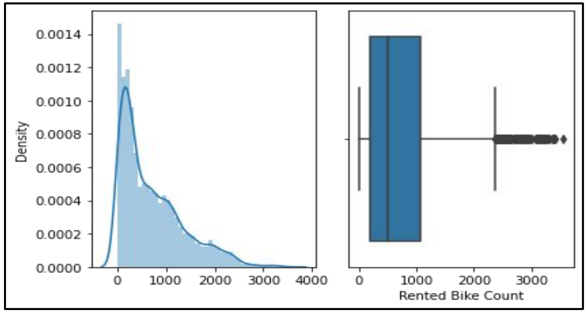
- **Date** : year-month-day
- **Hour**: hour of the day
- **Temperature**- celsius
- **Humidity** - %
- **Wind speed** - m/s
- **Visibility** - 10m
- **Dew point temperature** - celsius
- **Solar radiation** - MJ/m2
- **Rainfall** - mm
- **Snowfall** - cm
- **Seasons** - Winter, Spring, Summer, Autumn
- **Holiday** - Holiday/No holiday
- **Functional Day** - NoFunc(Non Functional Hours), Fun(Functional hours)

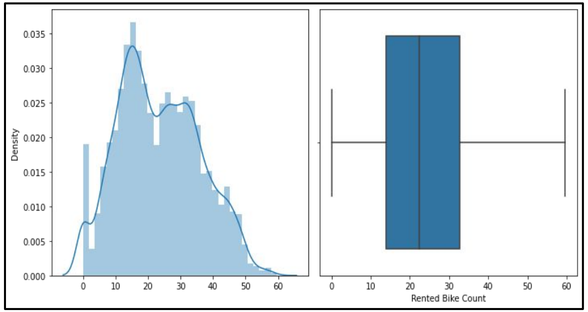
Exploratory Data Analysis



- **<u>Missing Value Analysis</u>**
  There were no missing values found in the dataset.

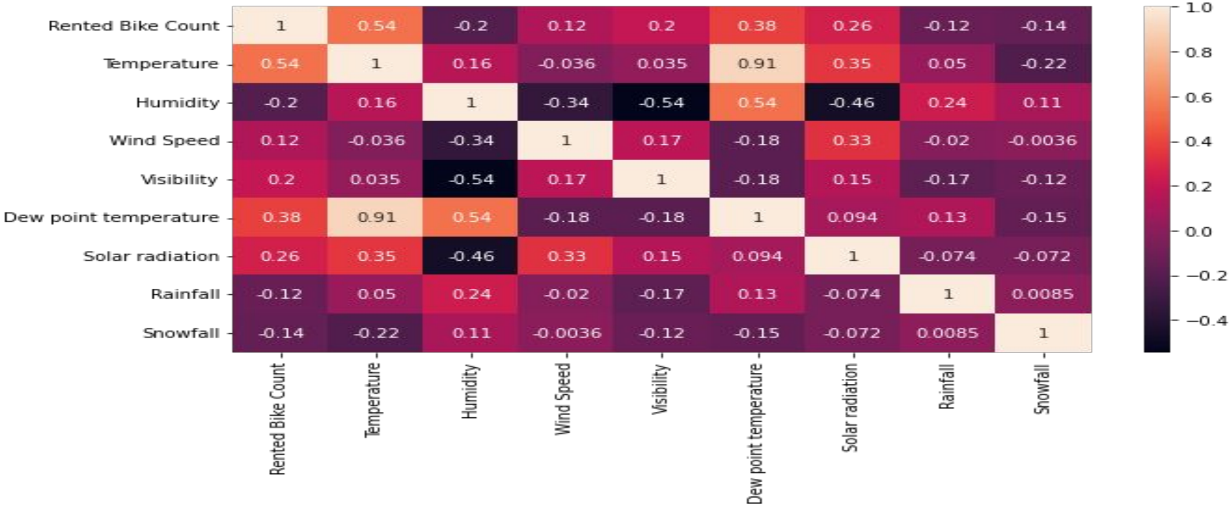
- **<u>Outlier Analysis</u>**
  - The outliers of the features are handled during data modeling using *Robust scaler*.
  - The outliers of the target variable were treated using *Square Root Transformation.*



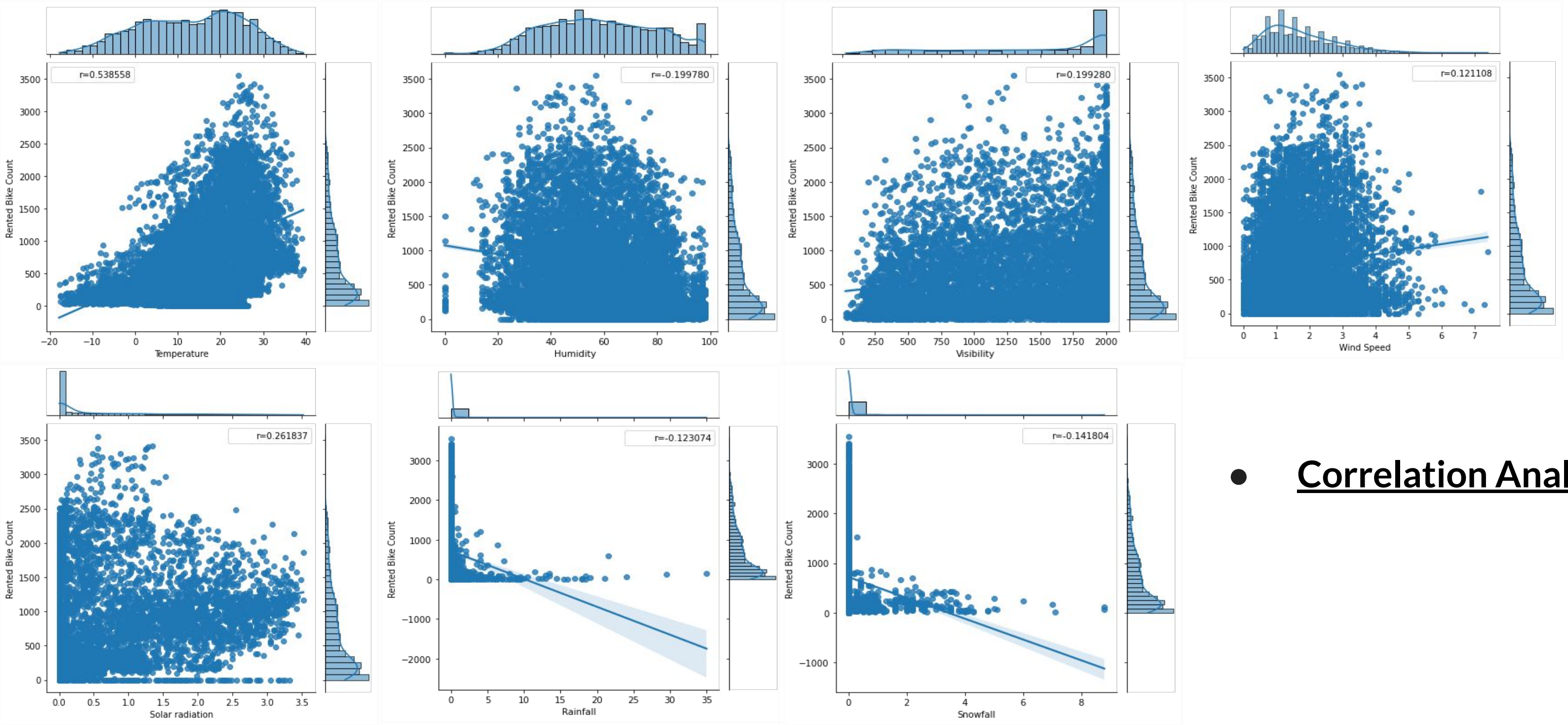


Exploratory Data Analysis

- <u>**Correlation Analysis**</u>
    - The temperature correlates (0.54) with the count of bike rents.
    - Temperature and dew-point temperature are highly correlated. One of the features could be dropped later.



Exploratory Data Analysis

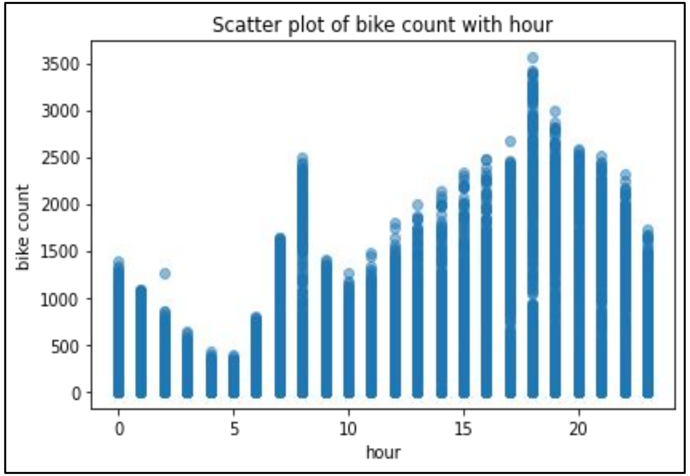


- **<u>Correlation Analysis</u>**

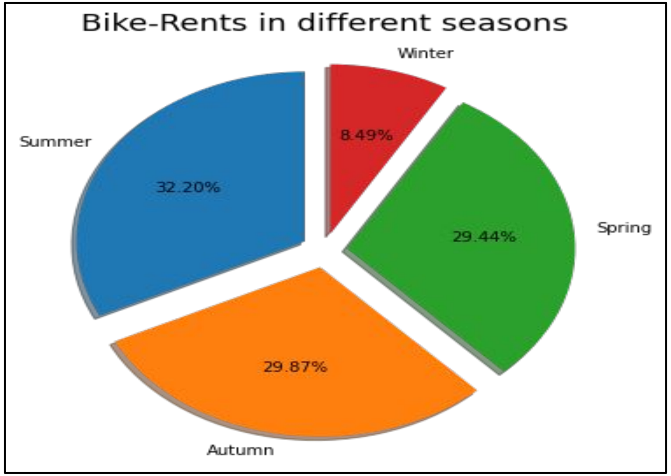
Exploratory Data Analysis

- **Variable Analysis**

Bike Rental Count Analysis: Count v/s Hour of the day



Bike Rental Count Analysis: Different seasons



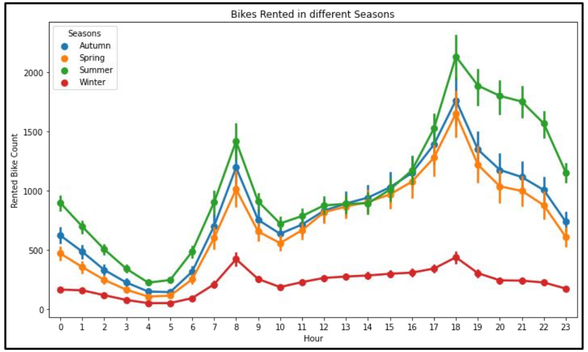
Exploratory Data Analysis

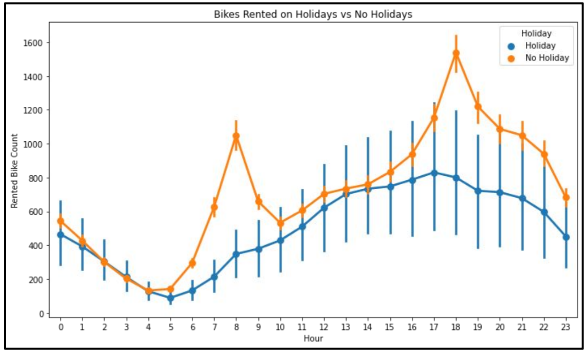
- <u>**Variable Analysis: Bike Rental Count Analysis- Throughout the day**</u>

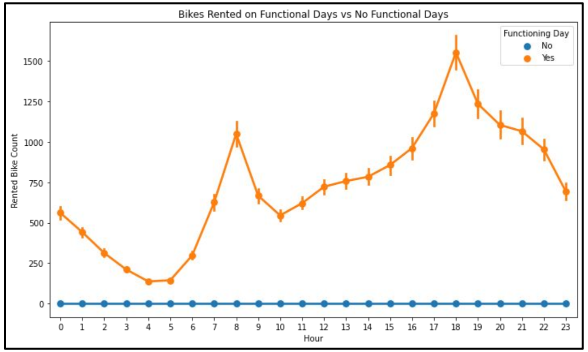
All seasons

Holiday: Yes or No

Functioning Day : Yes or No



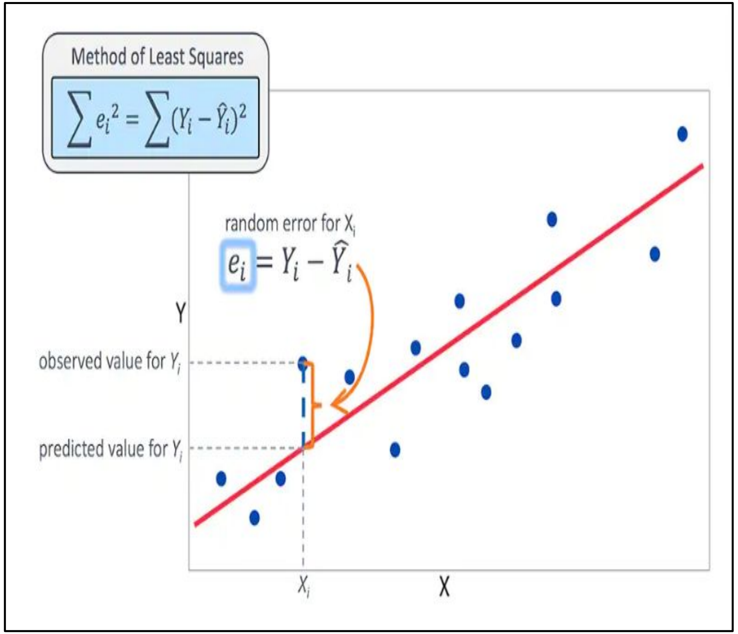




Regression Analysis

<u>Assumptions of a Regression Model.</u>

- There should be a linear and additive relationship between the dependent variable and the independent variable(s).
- No Autocorrelation: There should be no correlation between the residual (error) terms.
- No Multicollinearity: There shouldn't be a correlation between independent variables.
- Homoscedasticity: The error terms must have constant variance.
- The error terms must be normally distributed.



Data Modeling

Feature Selection: Multicollinearity Test

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables

- VIF is always greater or equal to 1.
- If VIF = 1 ➪ Not correlated to any of the variables.
- If 1 < VIF < 5 ➪ Moderately correlated.
- VIF > 5 ➪ Highly correlated.
- If there are multiple variables with VIF greater than 5, then remove one of them and repeat the process.

Encoding categorical columns

One Hot Encoding is used to produce binary integers- 0 and 1 to encode the categorical features. The categorical features namely season, hour, month, holiday, and functioning day are encoded.

Data Modeling

Data Split

The dataset is split into train and test data in the ratio of 75:25 resp, using sklearn's *train_test_split*.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42,shuffle= True)
```

Train: ***6560** observations and **51** features.*

Test: ***2190** observations and **51** features.*

Hyperparameter Tuning

GridSearchCV is used along with cross validation to get the best values for the specified hyperparameters.

It takes a dictionary with parameter names as keys and lists of parameter values, a performance measure and an integer that is the number of folds for K-fold cross-validation.

Data Modeling

Feature Scaling

Feature scaling is a method used to normalize the range of independent variables or features of data.

Why feature scaling?

- To facilitate fair comparison of features of different units based on standardized coefficients.
- Regularization techniques manipulate the value of the coefficients, this makes the model performance sensitive to the scale of features.
- To handle the outliers in the predictors.

Therefore, the split data is subjected to:

Robust Scaler - handles the outliers in the features, due to its insensitivity to outliers.

Minmax Scalar- normalizes the feature values.

Data Modeling

Model fitting

The following models have been studied and implemented on the given dataset:

- Linear Regression
- Regularized Regression
  - Lasso Regression
  - Ridge Regression
  - Elastic Net Regression
- Decision Tree regression
- Random Forest Regression
- Gradient Boosting Regression
- Light Gradient Boosting Regression
- CatBoost Regression

Evaluation

Evaluation Metrics

R_Squared

$$\text{R-Square} = 1 - \frac{\Sigma(\text{Y\_actual} - \text{Y\_predicted})^2}{\Sigma(\text{Y\_actual} - \text{Y\_mean})^2}$$

Mean Squared Error

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

Adjusted R_Squared

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

where:

n = number of observations

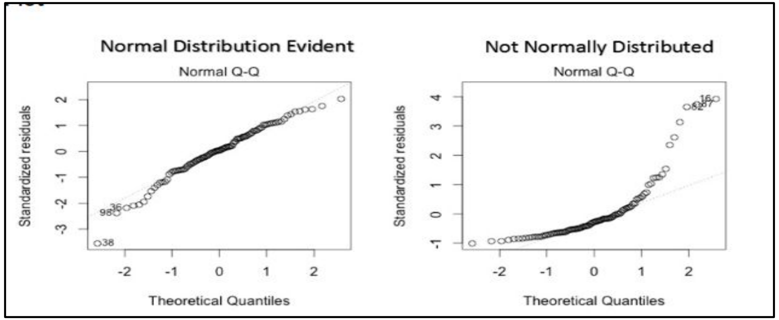
k = number of independent variables

R_a^2 = adjusted R^2

Evaluation

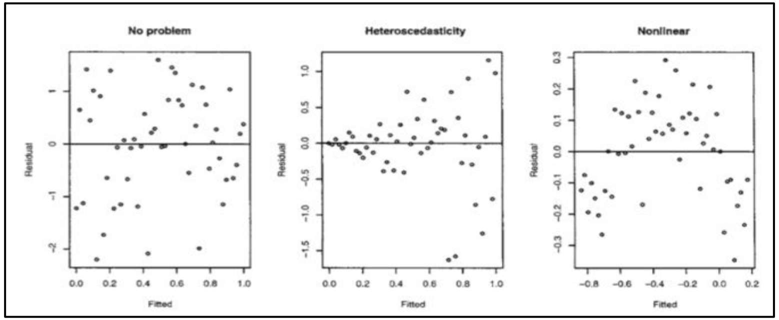
Evaluation Plots: Q-Q Residual Plot

- The Q-Q or quantile-quantile is a scatter plot that helps in validating the assumption of normal distribution in a data set.
- Fairly straight line aligning with the $45°$ line indicates normal distribution of errors.



Evaluation Plots: Residual Plot

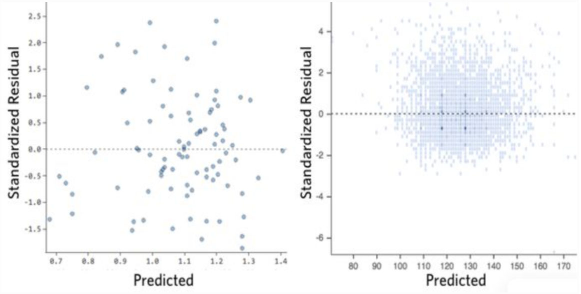
- The presence of non-constant variance in the error terms results in heteroscedasticity.

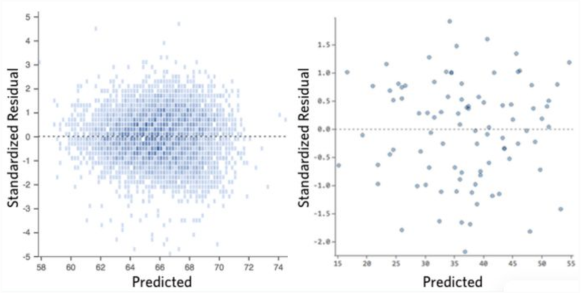


Evaluation

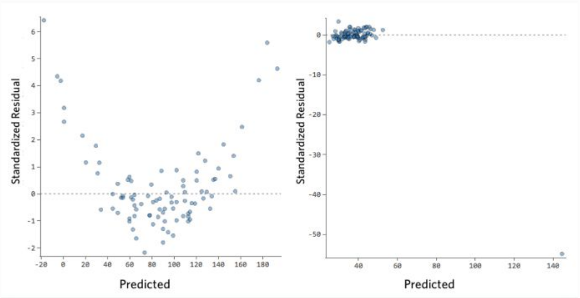
<u>Evaluation Plots: Residual Plot continued..</u>

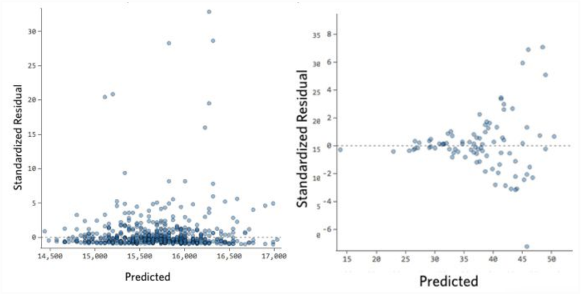
1.





2.





1. Ideal Plots
2. Non-Ideal Plots

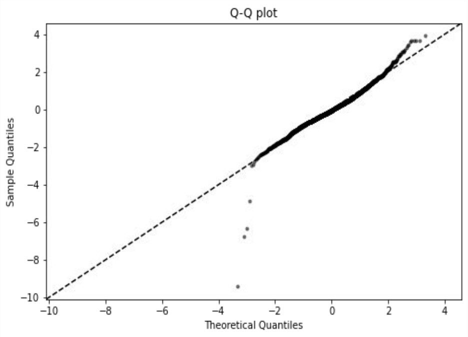
Results- Linear Regression

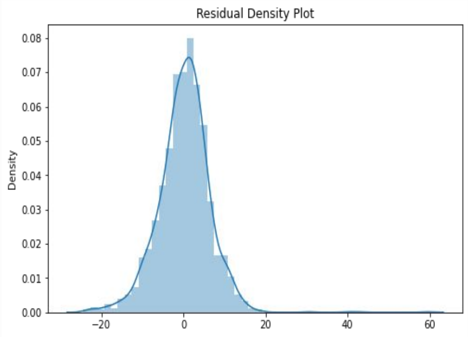
Train

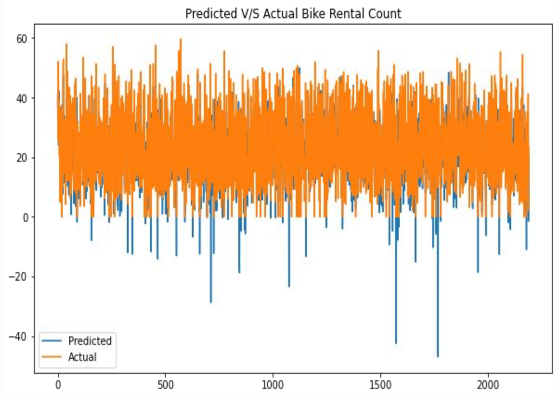
MSE	RMSE	R2_score	Adjusted R2_score
37.083	6.0896	0.7626	0.7627

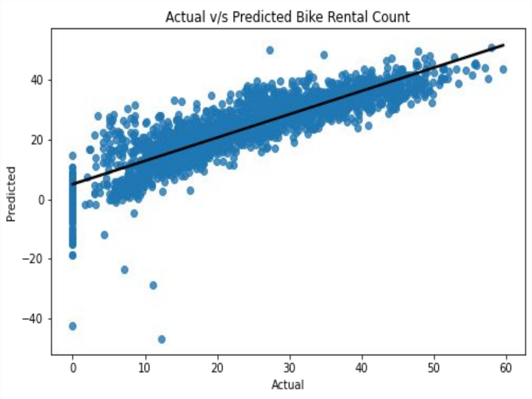
Test

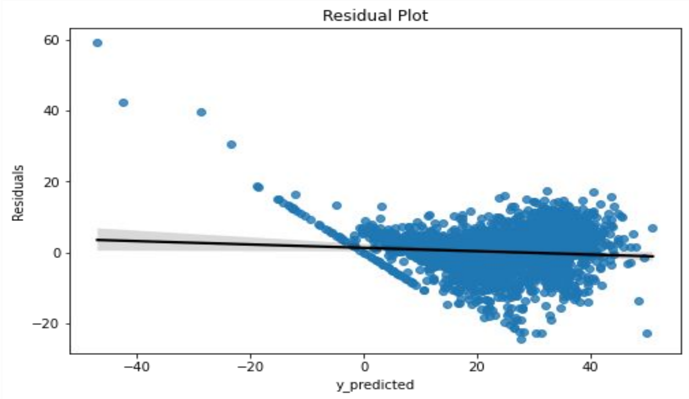
MSE	RMSE	R2_score	Adjusted R2_score
39.2326	6.2636	0.7408	0.7409











Results- Regularized Regression [Lasso](GridsearchCV)

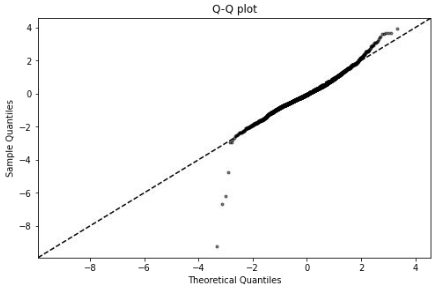
```
lasso_grid = GridSearchCV(lasso_reg, parameters, scoring='neg_mean_squared_error', cv=5)  'alpha': 0.0015
```

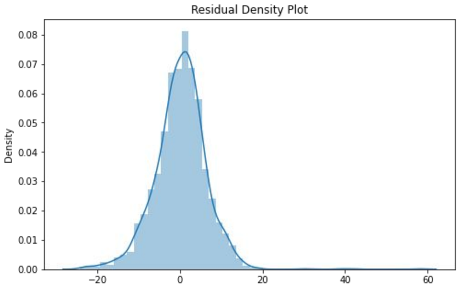
Train

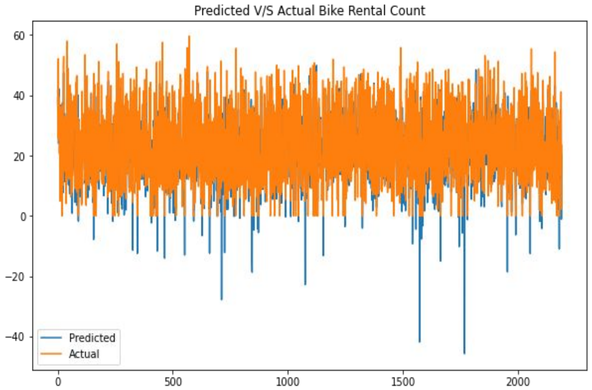
MSE	RMSE	R2_score	Adjusted R2_score
37.0875	6.0899	0.7626	0.7626

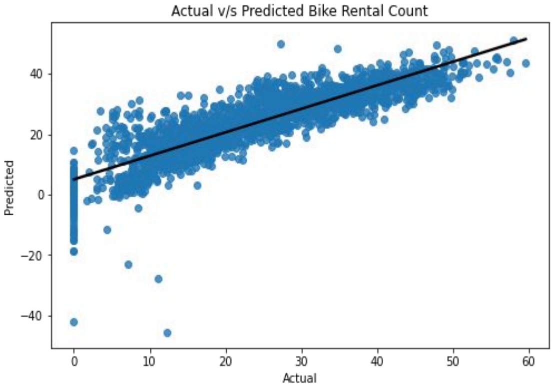
Test

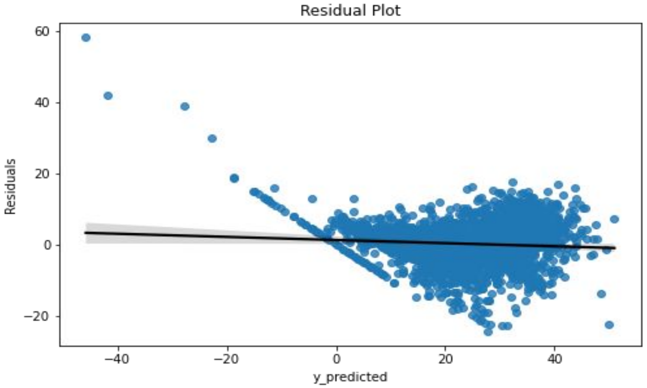
MSE	RMSE	R2_score	Adjusted R2_score
39.0978	6.2528	0.7417	0.7418











Results- Regularized Regression [Ridge](GridsearchCV)

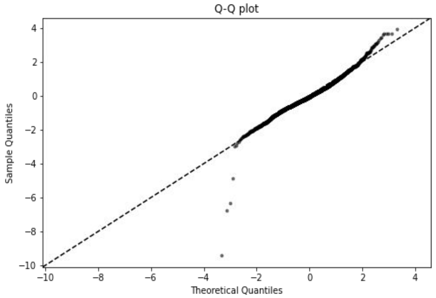
```
ridge_grid = GridSearchCV(ridge_lg, parameters, scoring='neg_mean_squared_error', cv=3)   'alpha': 0.01
```

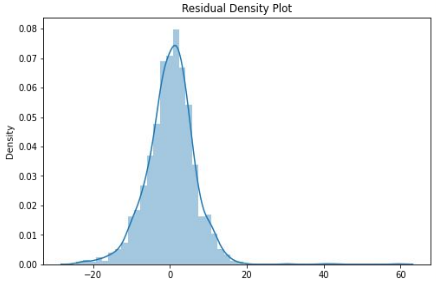
Train

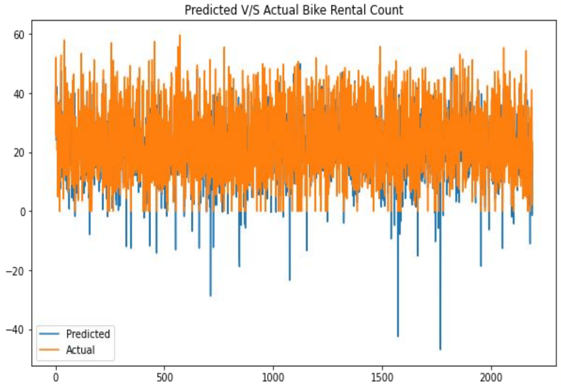
MSE	RMSE	R2_score	Adjusted R2_score
37.083	6.0896	0.7626	0.7627

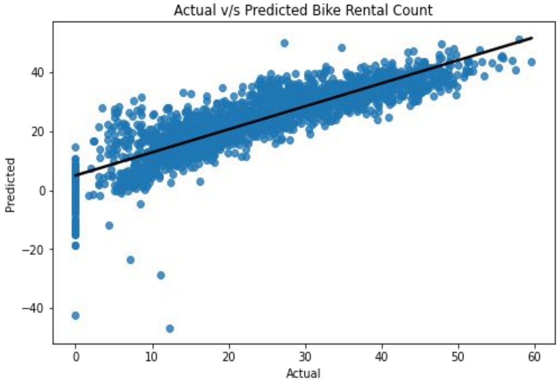
Test

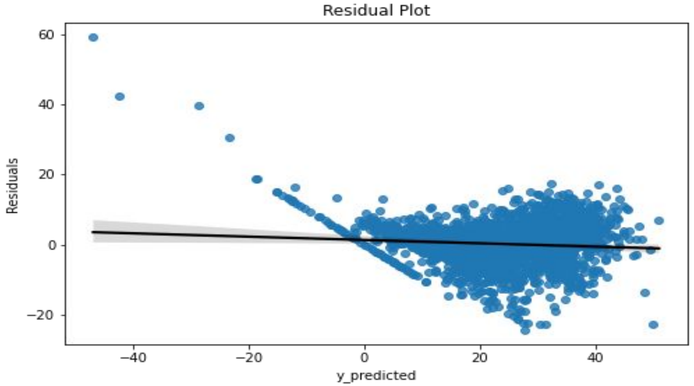
MSE	RMSE	R2_score	Adjusted R2_score
39.2264	6.2631	0.7408	0.7409











Results- Regularized Regression [Elastic Net](GridsearchCV)

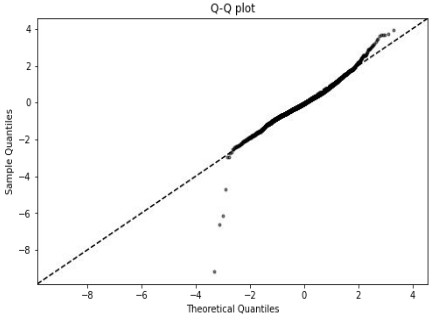
ElasticNet(alpha=0.0001, l1_ratio=0.6)

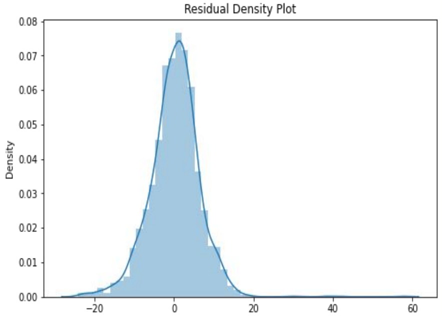
Train

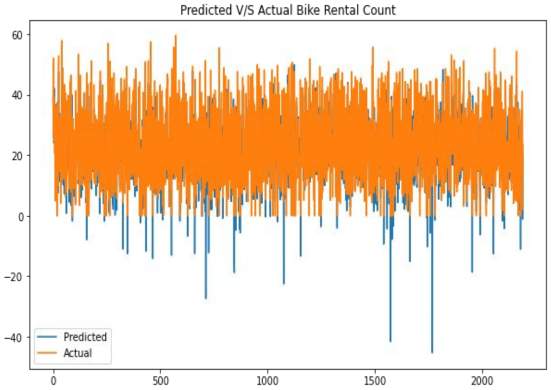
MSE	RMSE	R2_score	Adjusted R2_score
37.0854	6.0898	0.7626	0.7627

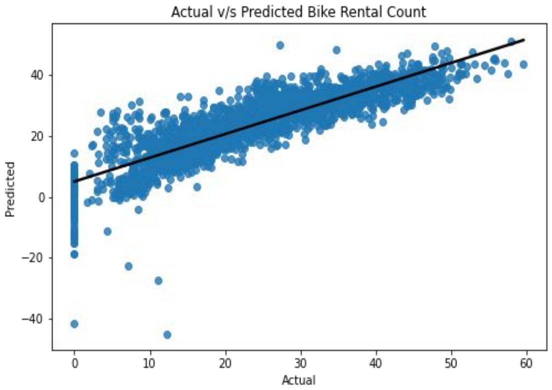
Test

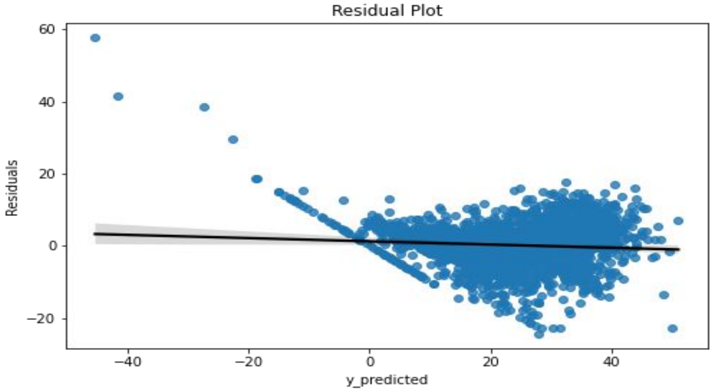
MSE	RMSE	R2_score	Adjusted R2_score
39.069	6.2505	0.7419	0.742





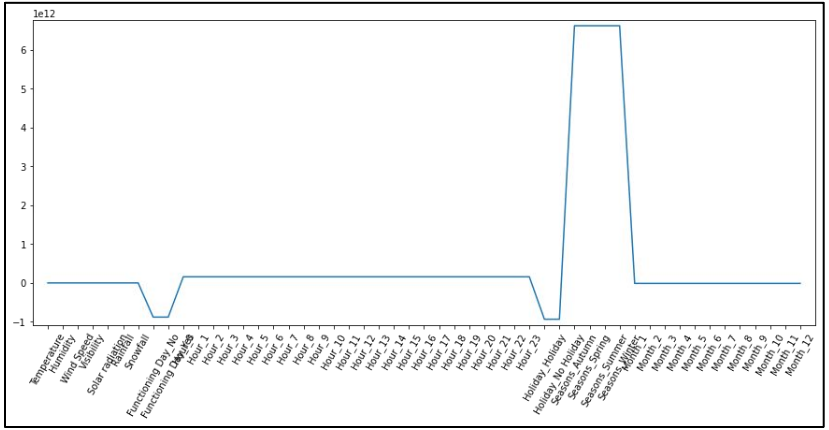


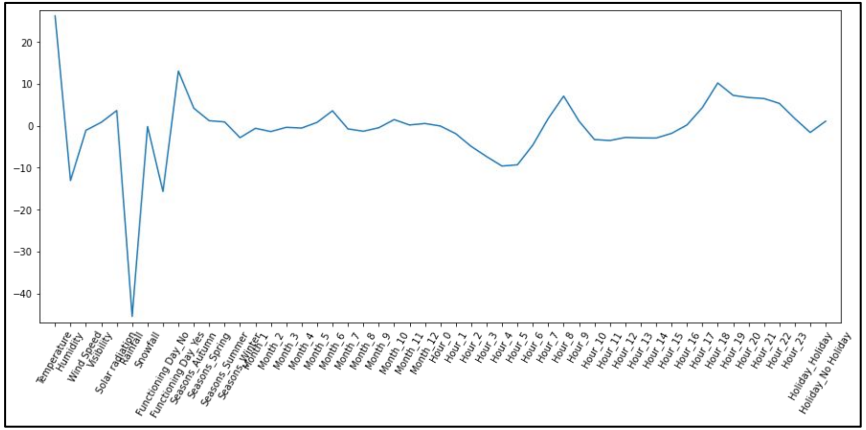




Results

Linear Regression v/s Regularized Regression- Coefficients plot





- High estimated coefficients for few features leading to incomparable coefficients.

- Functioning hours and peak hours have high positive coefficients.
- Humidity, Rainfall and No functioning hours are negatively related to bike count.

Results- Decision Tree Regression (GridsearchCV)

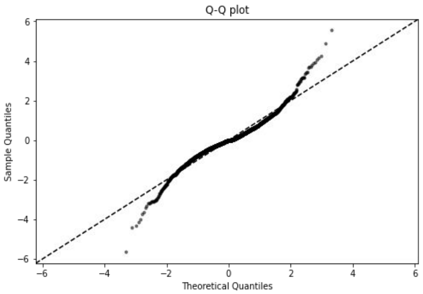
DecisionTreeRegressor(max_depth=14, min_samples_split=14)

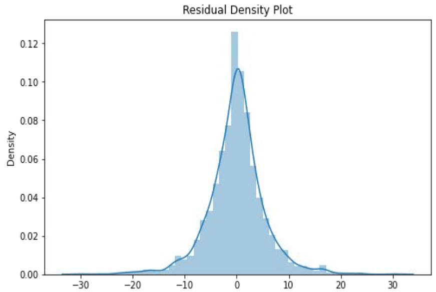
Train

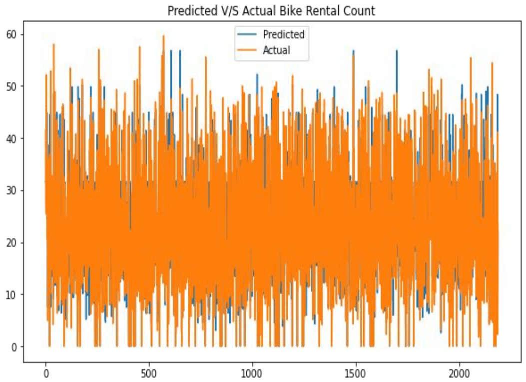
MSE	RMSE	R2_score	Adjusted R2_score
17.8624	4.2264	0.8857	0.8857

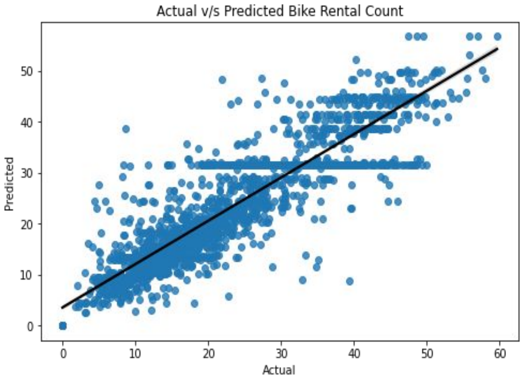
Test

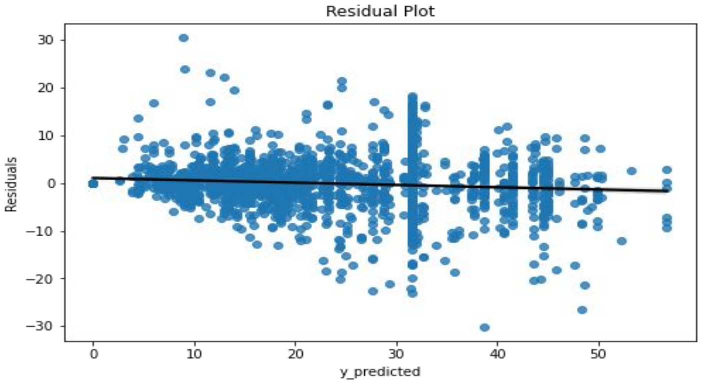
MSE	RMSE	R2_score	Adjusted R2_score
29.1661	5.4006	0.8073	0.8074











Results- Random Forest Regression (GridsearchCV)

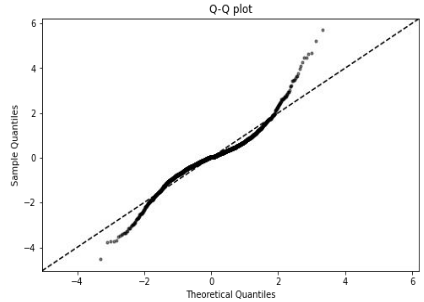
RandomForestRegressor(max_depth=19, min_samples_split=3, n_estimators=500)

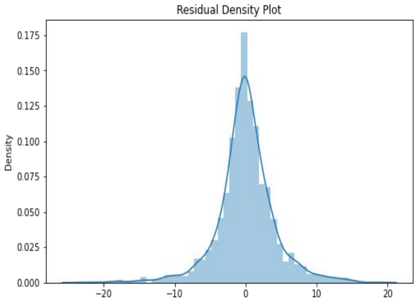
Train

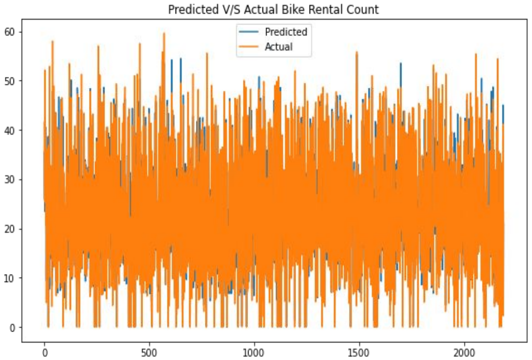
MSE	RMSE	R2_score	Adjusted R2_score
6.304	2.5108	0.9596	0.9597

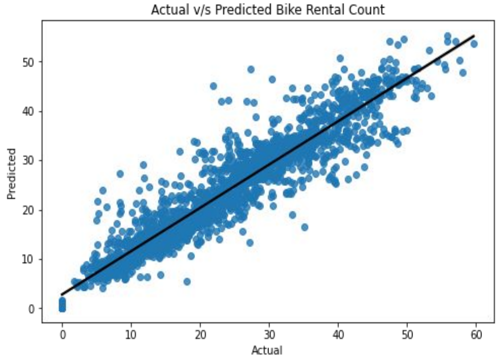
Test

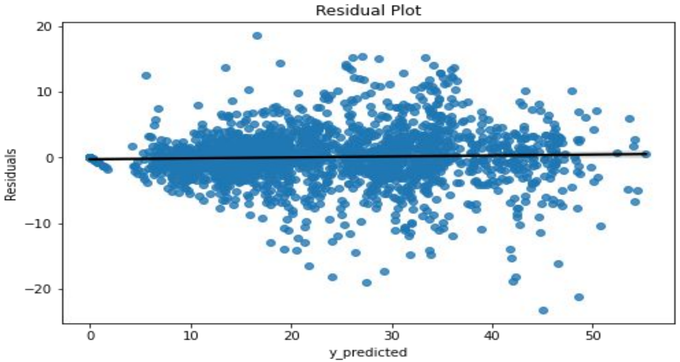
MSE	RMSE	R2_score	Adjusted R2_score
16.7718	4.0953	0.8892	0.8892











Results- Gradient Boosting Regression (GridsearchCV)

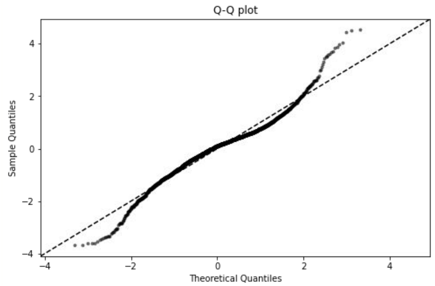
GradientBoostingRegressor(learning_rate=0.02, max_depth=8, n_estimators=150)

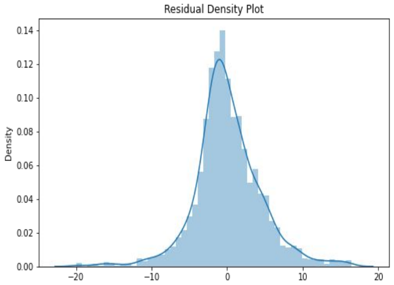
Train

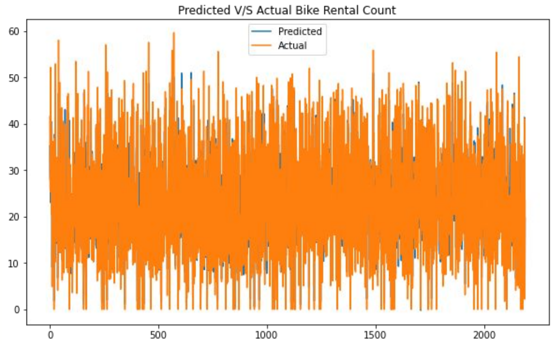
MSE	RMSE	R2_score	Adjusted R2_score
12.9139	3.5936	0.9173	0.9174

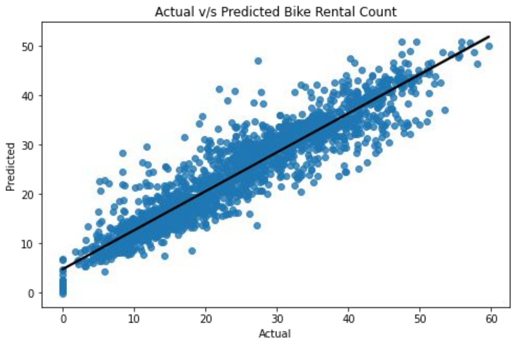
Test

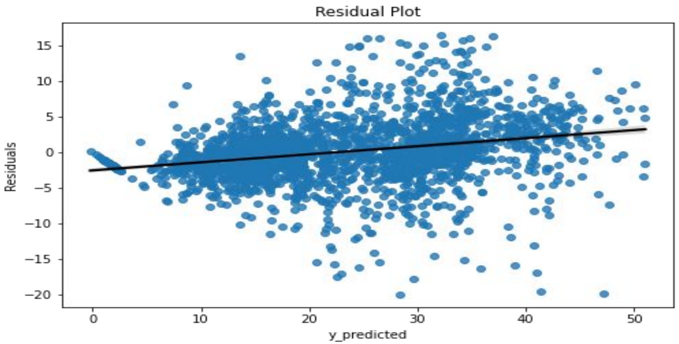
MSE	RMSE	R2_score	Adjusted R2_score
19.642	4.4319	0.8702	0.8703











Results- Light GBM Regression (GridsearchCV)

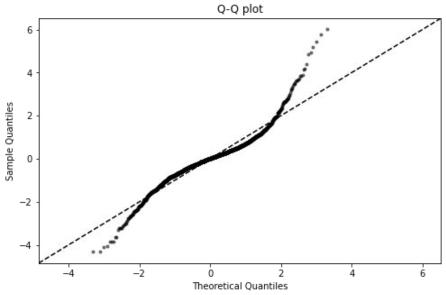
LGBMRegressor(max_depth=15, n_estimators=250)

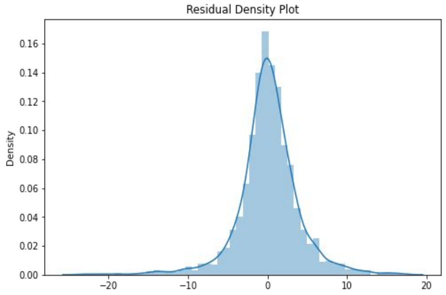
Train

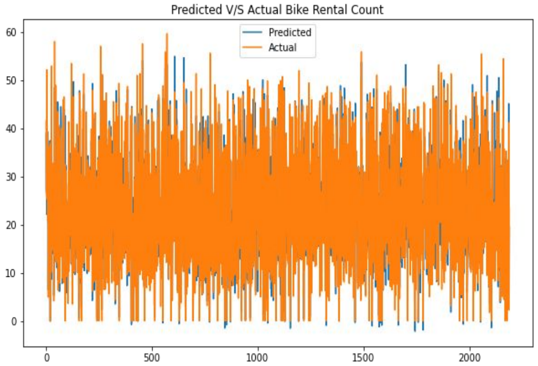
MSE	RMSE	R2_score	Adjusted R2_score
6.049	2.4595	0.9613	0.9613

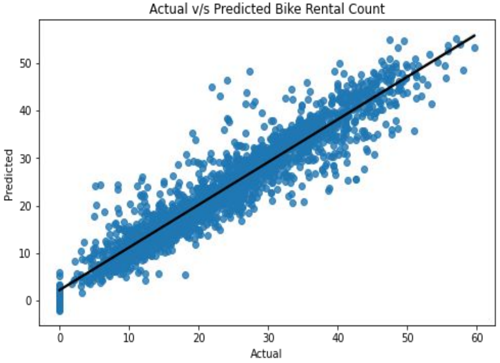
Test

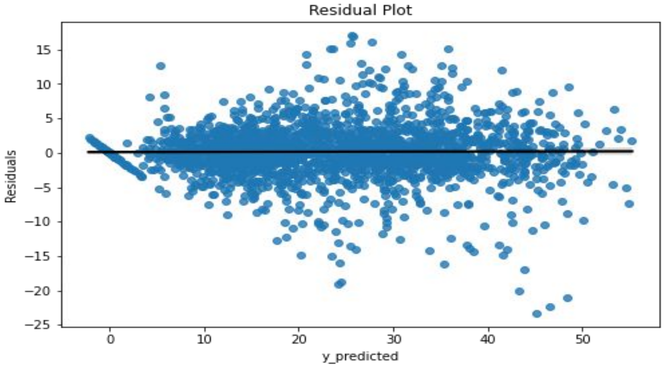
MSE	RMSE	R2_score	Adjusted R2_score
15.0204	3.8756	0.9008	0.9008











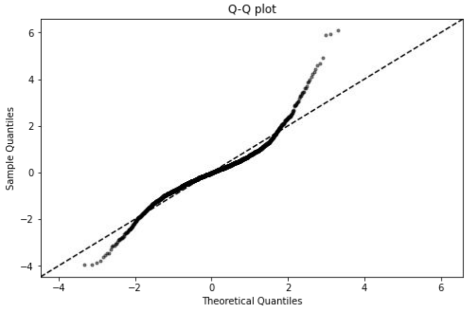
Results- CatBoost Regression

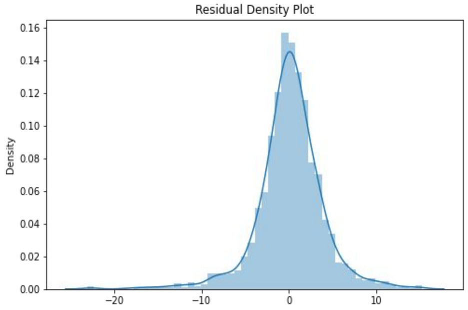
Train

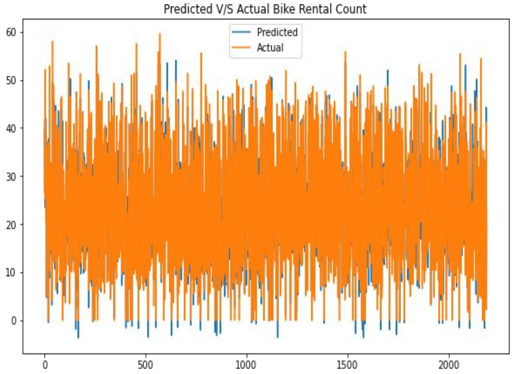
	MSE	RMSE	R2_score	Adjusted R2_score
	7.4535	2.7301	0.9523	0.9523

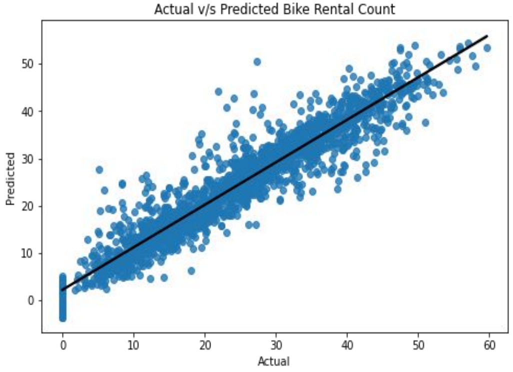
Test

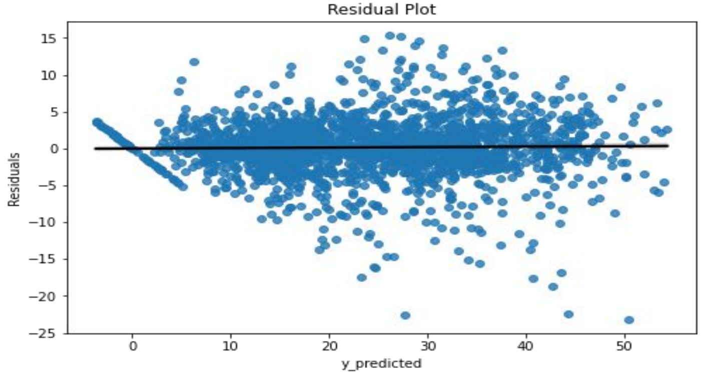
	MSE	RMSE	R2_score	Adjusted R2_score
	14.7605	3.8419	0.9025	0.9025







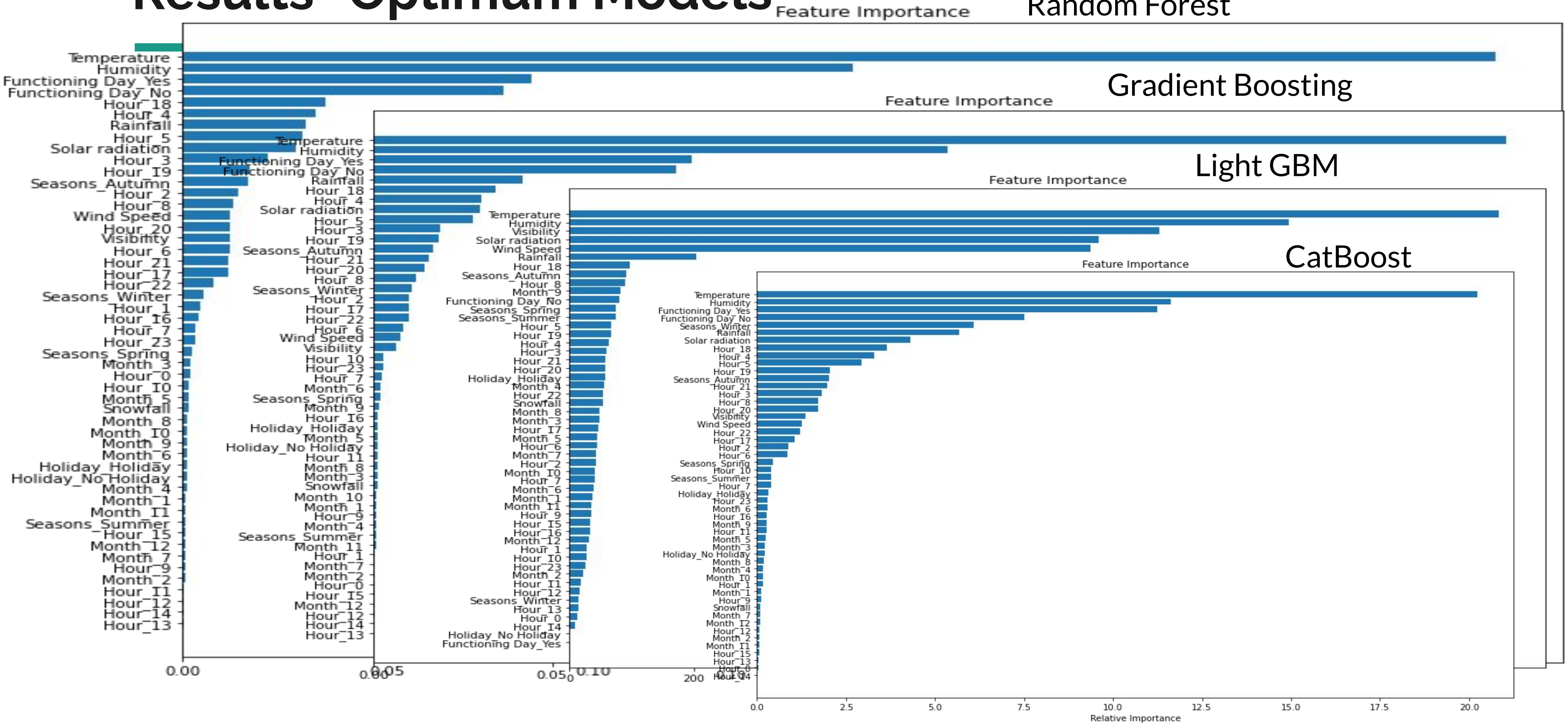




Results- Summary

Model	MSE-train	MSE-test	Adjusted R2_score-train	Adjusted R2_score-test
Linear Regression	37.0831	39.2343	0.7627	0.7409
Lasso Regression GridSearchCV	37.0875	39.0978	0.7626	0.7418
Ridge Regression GridSearchCV	37.0830	39.2264	0.7627	0.7409
Elastic-Net GridSearchCV	37.0854	39.0690	0.7627	0.7420
Decision Tree Regression	30.7392	36.9465	0.8033	0.7560
Decision Tree GridSearchCV	17.8624	29.0001	0.8857	0.8085
Random Forest Regression	2.3587	16.2013	0.9849	0.8930
Random Forest GridSearchCV	6.3040	16.7718	0.9597	0.8892
Gradient Boosting Regression	21.1626	22.2409	0.8646	0.8531
Gradient Boosting GridSearchCV	12.9139	19.6146	0.9174	0.8705
Light Gradient Boosting GridSearchCV	6.0490	15.0204	0.9613	0.9008
CatBoost Regression	7.4535	14.7605	0.9523	0.9025

Results- Optimum Models



Conclusion

- Bike rental count is seen to follow a certain trend to grow high during peak hours of favorable days, seasons, functioning hours and high temperatures.
- Temperature and Humidity remain the most important features in ensemble predictive models, followed by Functioning Day, Rainfall, Solar radiation, Hour_18.
- The predictive analysis could be enhanced if provided with location of bike stands in the city, and traffic details in real time.

Thank you