Capstone Project

Supervised ML - Regression

Seoul Bike Sharing Demand Prediction

Subham Choudhary

Contents

Introduction

Bike sharing system has recently received increasing attention around the world. Bike-sharing customers prefer to quickly find a bike whenever they need one. Thus, bike provider companies need to allocate bikes efficiently according to the demand.

There are many underlying factors — for example, time of the day, day of the week, events, weather, seasons, etc. — play an important role in determining the patterns of bikes renting demand in a city.

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.

The objective of the project is to predict the hourly number of bikes rented for the stable supply of rented bikes.

Data Overview

- We are provided with Seoul Bike Sharing Demand Prediction Dataset.

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

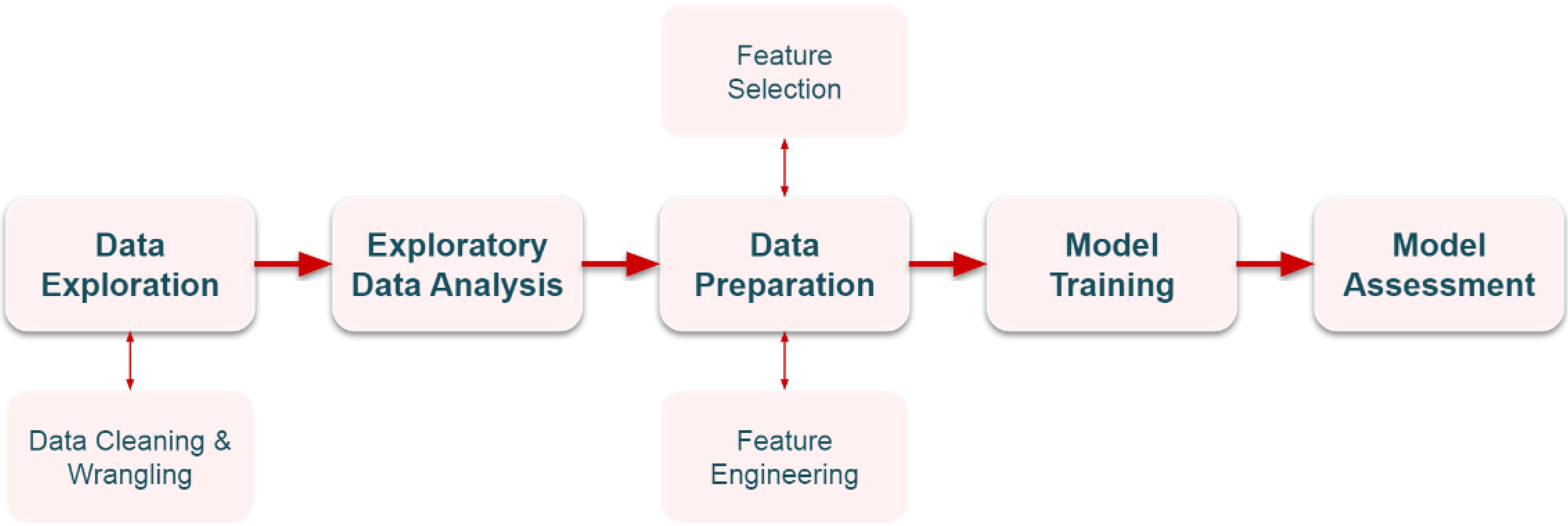
- The dataset contains 8760 non-null observations and 14 columns with a mix of numerical and categorical variables.

Data Overview (continued)

Attributes 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
- Seasons - Winter, Spring, Summer, Autumn
- Holiday - Holiday/No holiday
- Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

Steps Involved



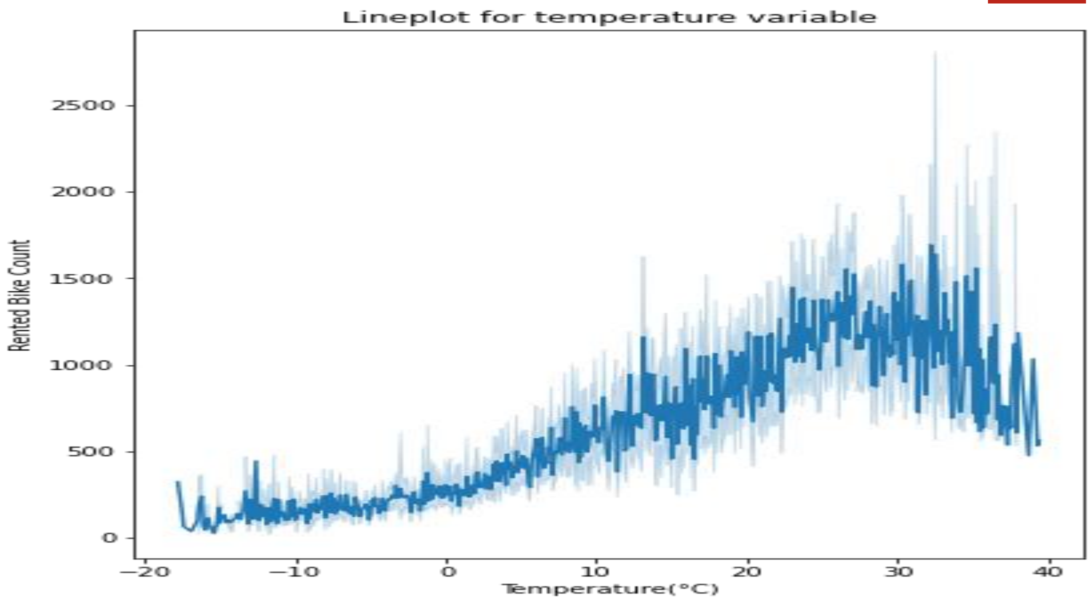
Exploratory Data Analysis

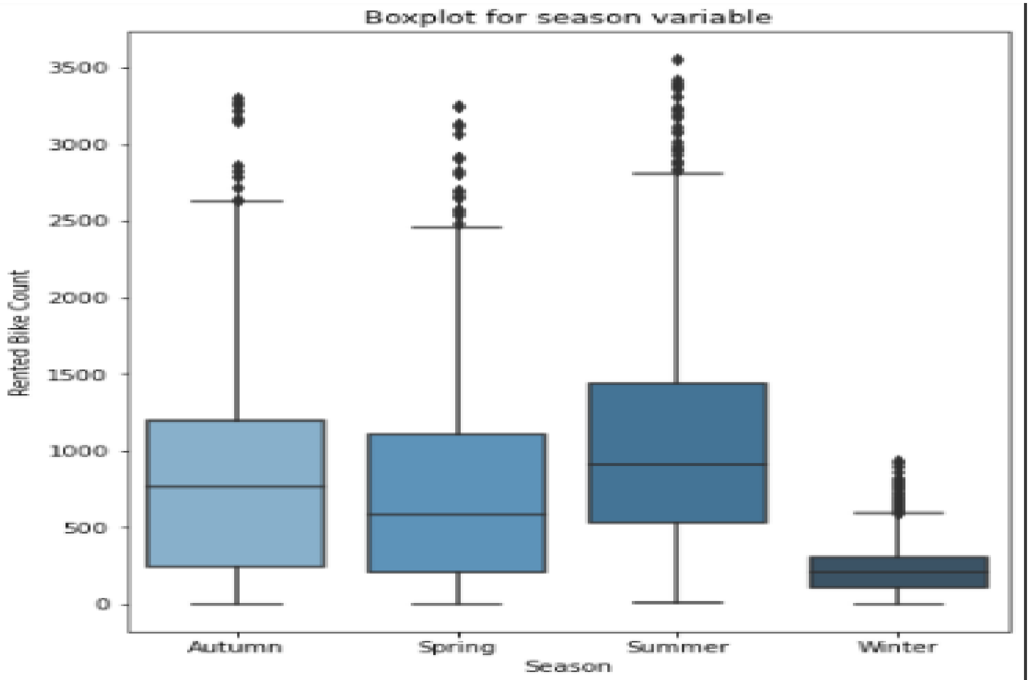
- Checking for null values
- Converting all columns to lowercase
- Data type of date is converted to Datetime and hour, seasons, holiday and functioning day converted to category data type
- No missing values
- No null values

Data Visualization

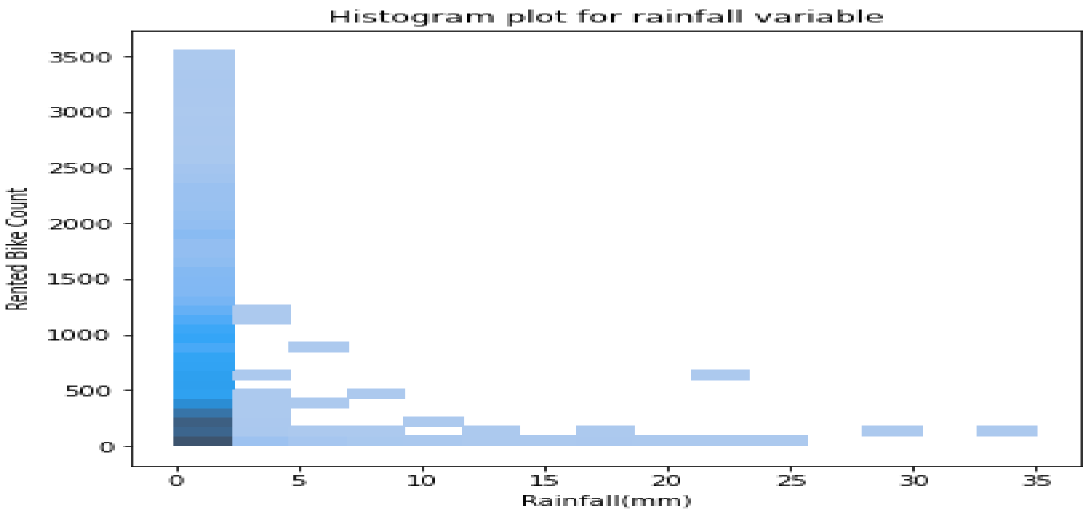
- When temperature is sub zero or below zero, the demand in rented bikes is minimal but as the temperature increases, the demand of rented bikes increases.

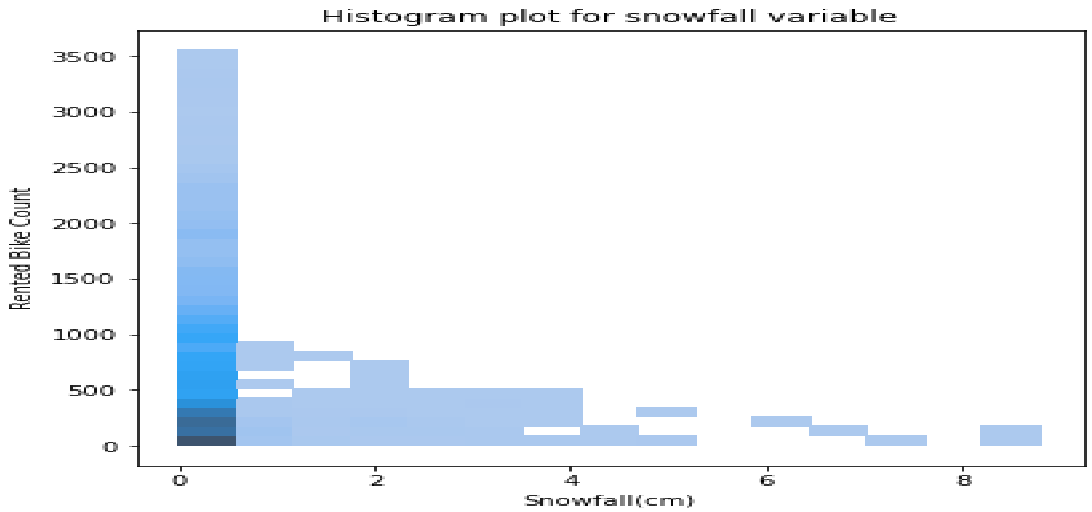
- The following trend can also be seen in the box plot of season variable where the number of rented bikes is minimal during Winter and maximum during Summer.

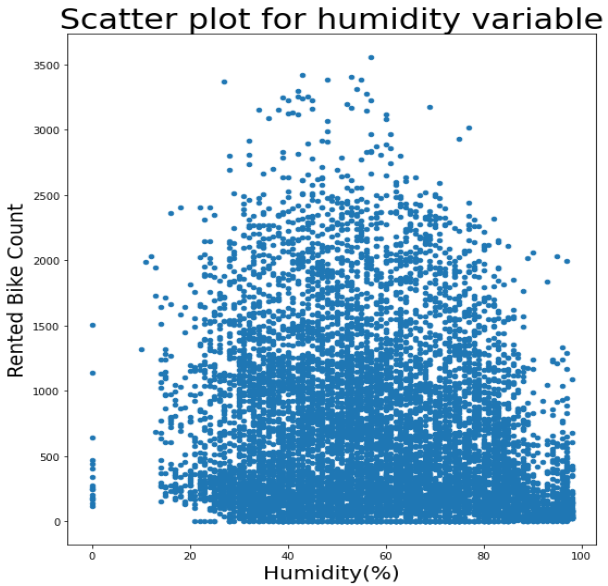


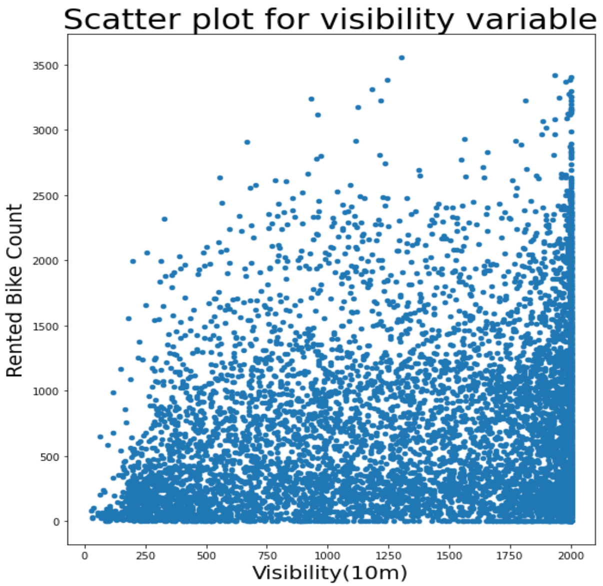


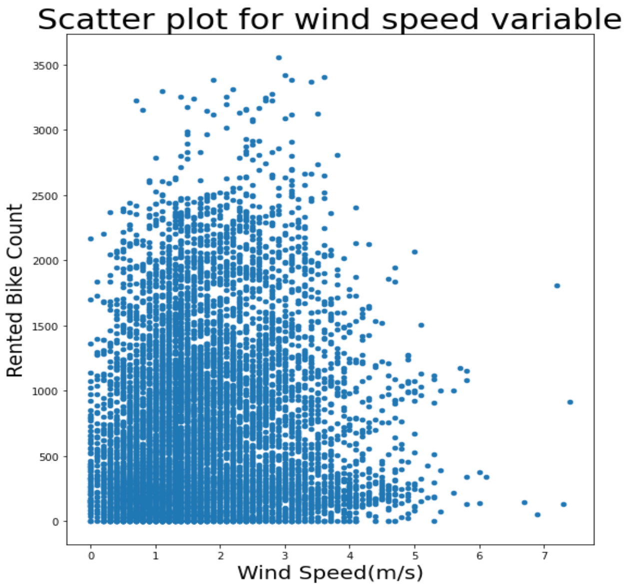
Plots – Numerical variable vs Rented bikes

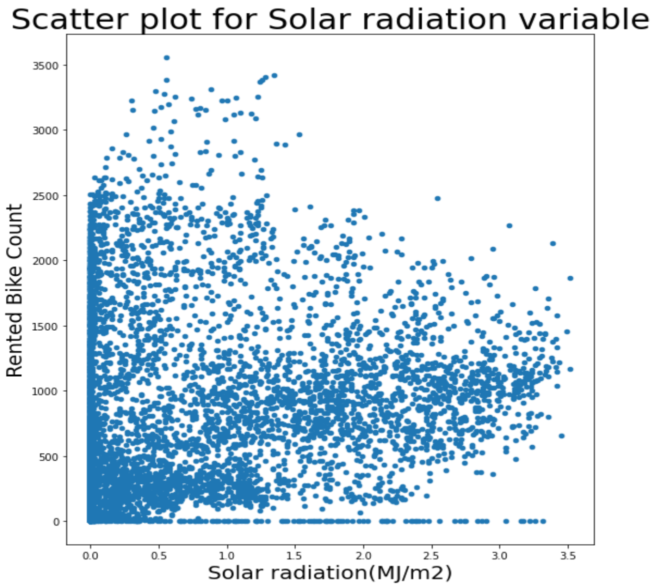












Explanation

- The histogram plot shows that when the rainfall and snowfall is minimum, the number of rented bikes is maximum, but as the amount of rainfall and snowfall increases, the number of rented bikes decreases. So, there is an inverse relationship between Rented bikes count and rainfall or snowfall.

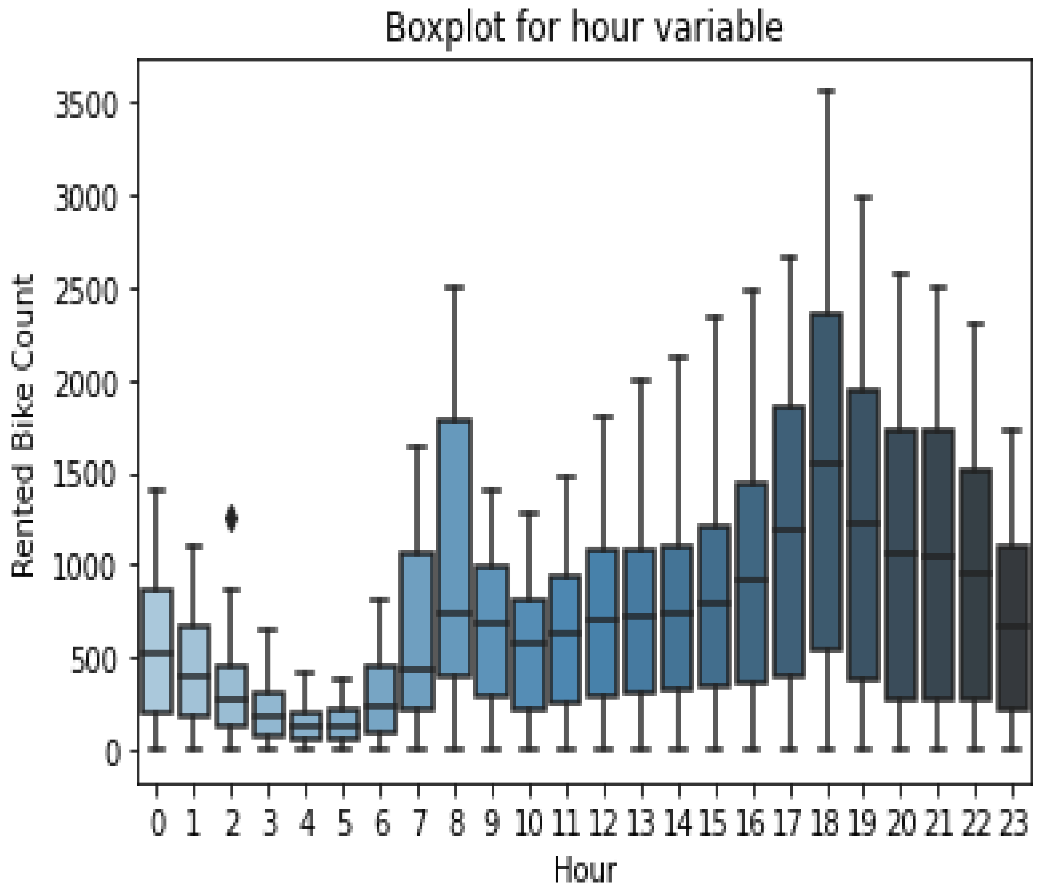
- The scatter plot of various numerical variables are skewed except humidity. This shows the demand for bikes is maximum when humidity is 20-40%.

- Visibility variable is left skewed, which means demand increases when visibility is higher.

- Demand is higher when windspeed and radiation is low.

Boxplot – Hour vs Rented bike count

The number of rented bikes has a sharp increases during 8am when most people goes to school or offices, and again there is a sharp increase during 18pm, that is when most people return from their work. The demand is fairly high for the most part of afternoon and evening.



Data Preparation

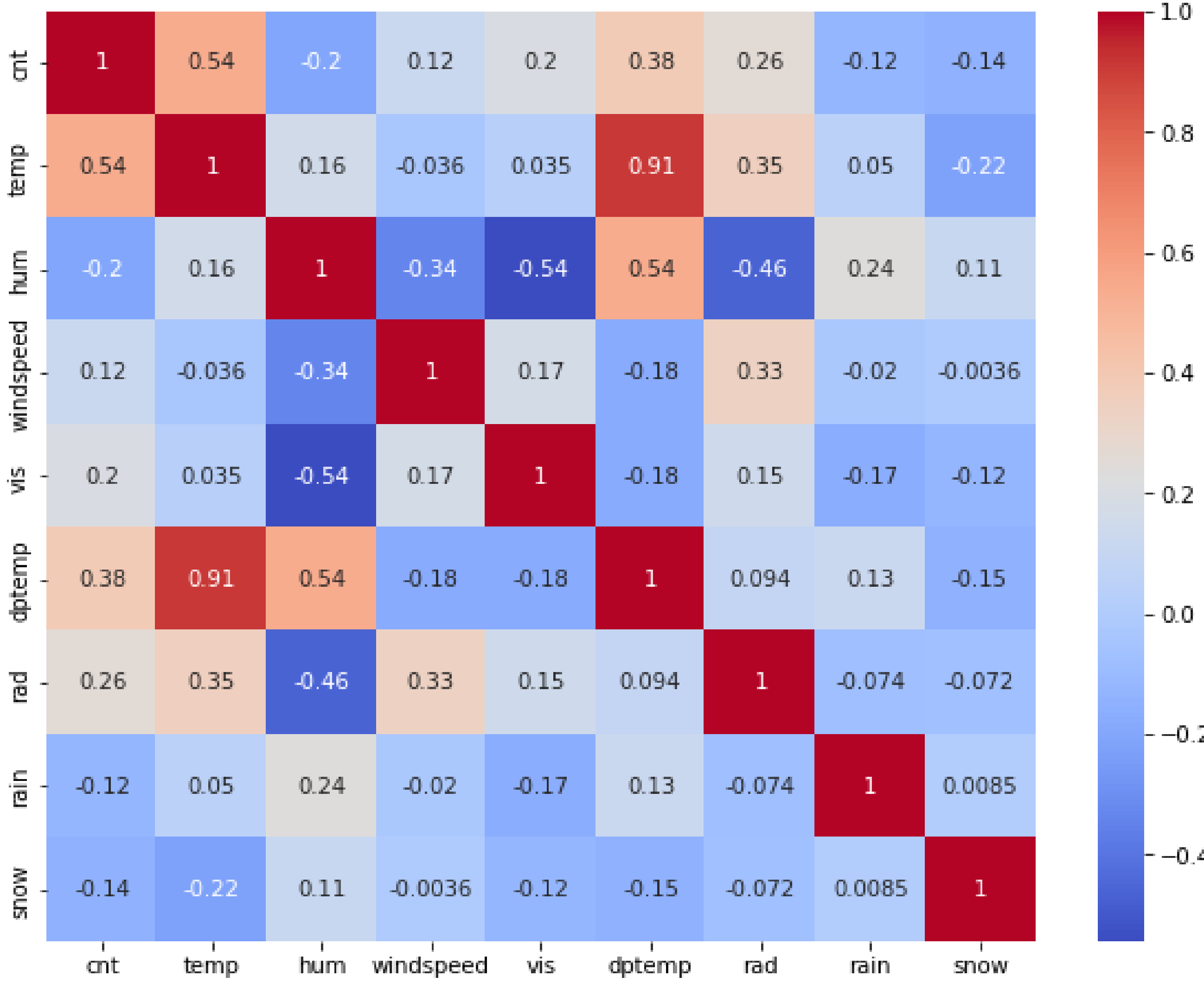
Feature selection

- Feature selection with Pearson Correlation Heatmap.
- Detecting multi-collinearity.
- Dropping highly correlated variables.

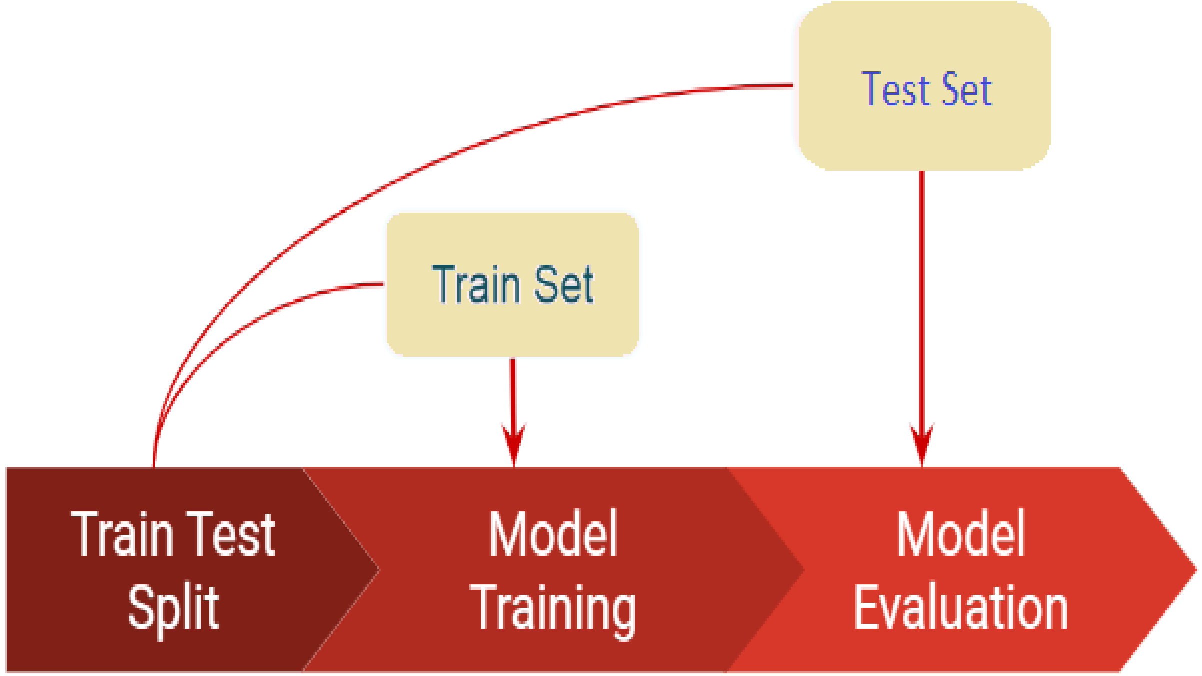
Feature engineering

- Seasons are assigned with numeric values.
- Three new variables are defined, isHighSeason, isRushHourMorning and isRushHourEvening, according to the feature explanation.
- Few other categorical variables are used to create dummy variables

Correlation Heatmap



Predictive Modelling



Regression models used:

- Linear Regression
- Lasso Regression
- Ridge Regression
- KNN (K Nearest Neighbor)
- SVM (Support Vector Machine)
- XGBoost

Predictive modelling include –

- Building and training the models.
- Tuning the hyperparameters to get better results
- Model evaluation and selection

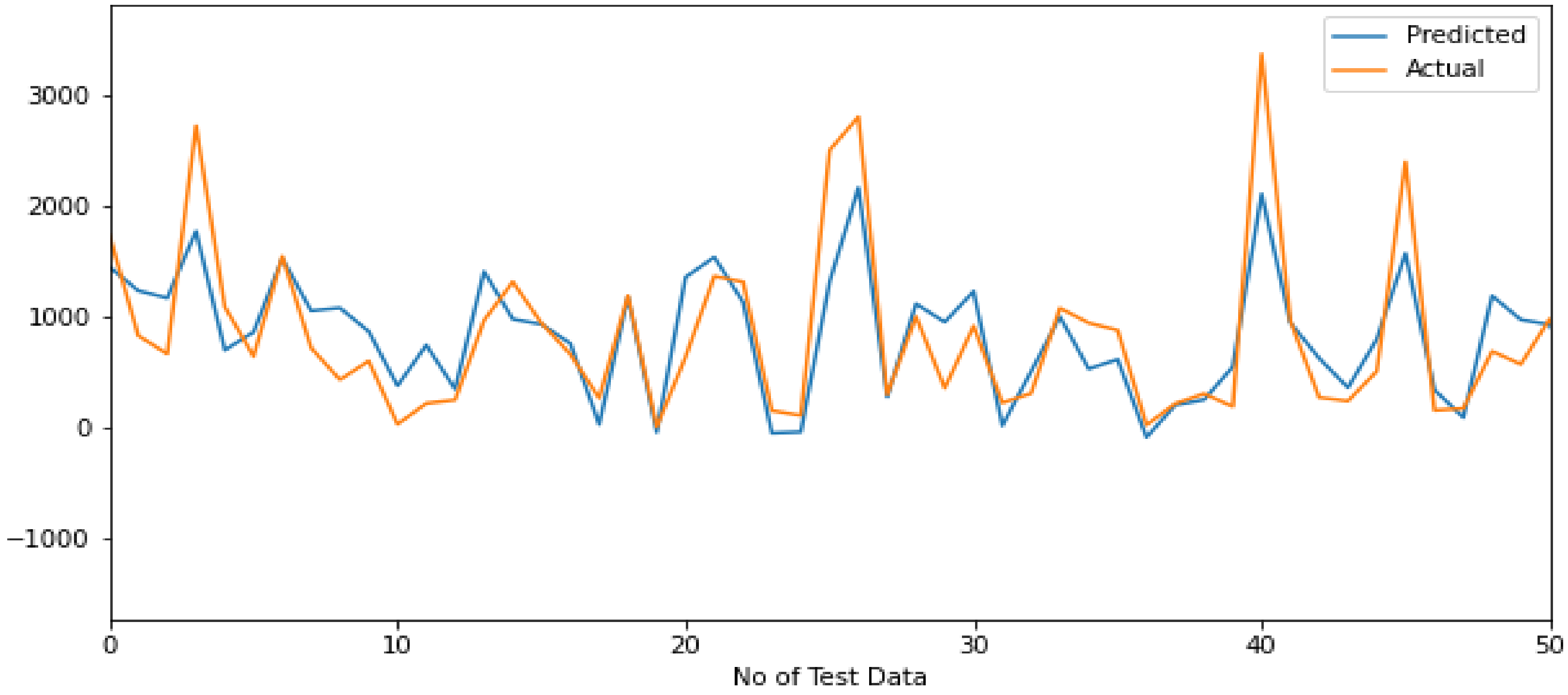
Linear Regression

Metrics Evaluation

```
MAE : 285.4584486260877
MSE : 141391.15457345723
RMSE : 376.0201518182998
MAPE : 1374.8866592584004
R2 : 0.6606445905977423
```

Linear regression gives r^2 score = 66%

Graph of Actual and Predicted values



Lasso Regression

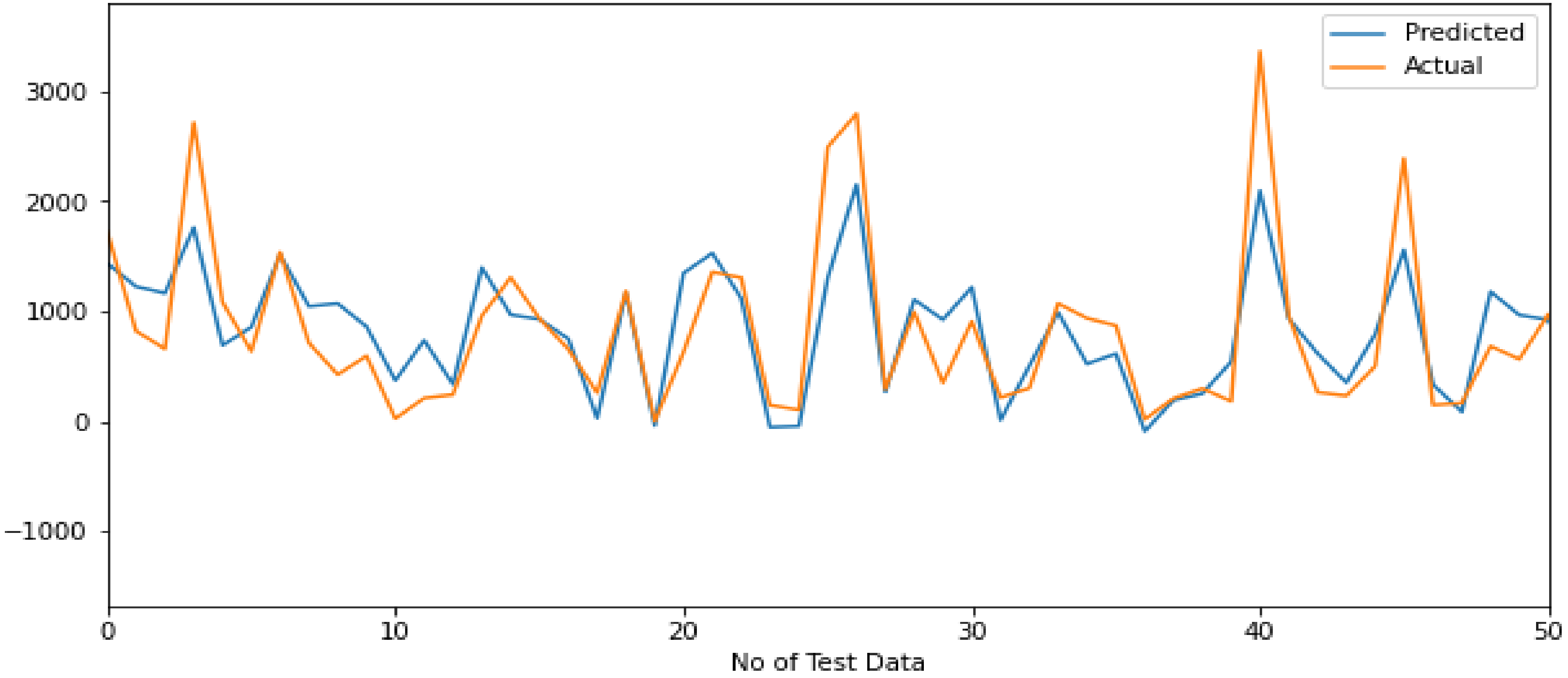
```
parameters = {'alpha': [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 5, 10, 20]}

The best fit alpha value is found out to be : {'alpha': 0.1}
```

Metrics Evaluation

MAE : 285.40458870571035
MSE : 141387.94131978304
RMSE : 376.0158790793057
MAPE : 1363.6863251411025
R2 : 0.6606523027846851

R^2 score for Lasso Regression = 66%

Graph of Actual and Predicted values



Ridge Regression

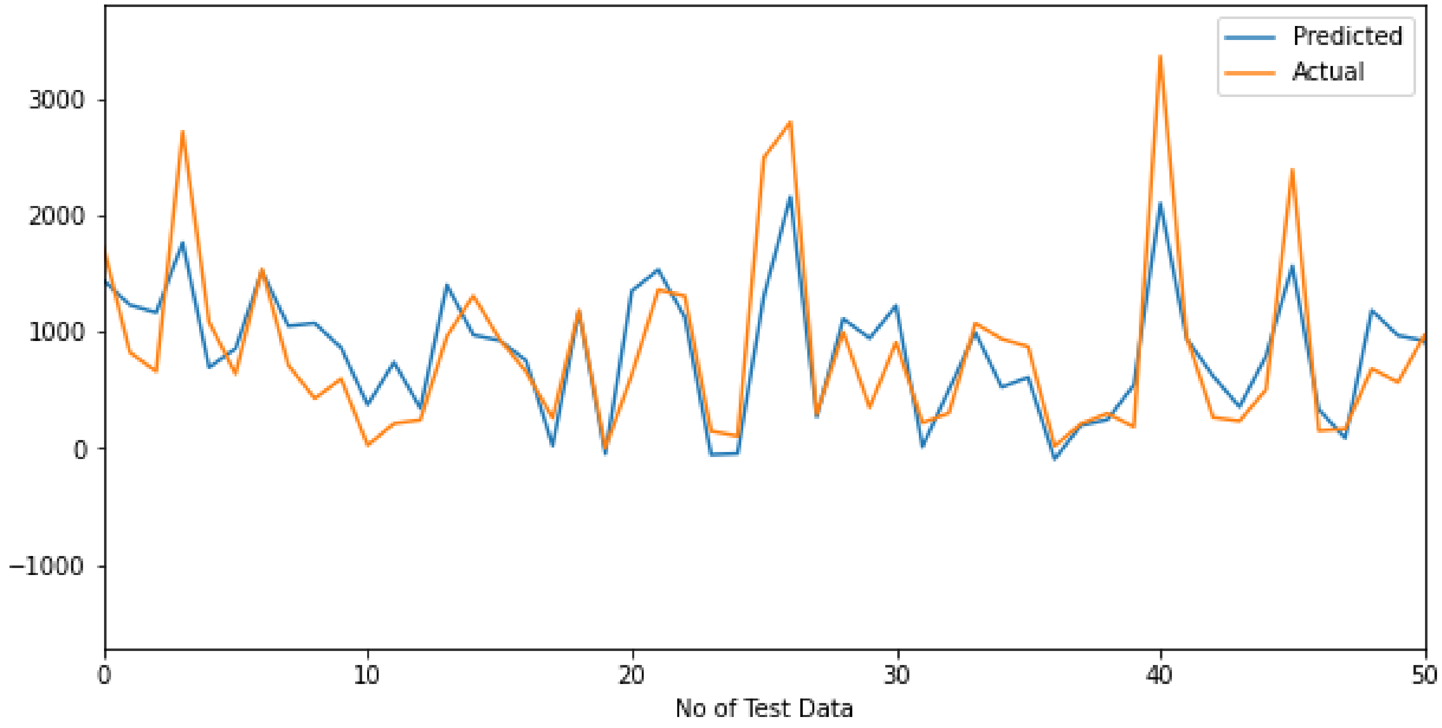
```
parameters = {'alpha': [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 5, 10, 20]}

The best fit alpha value is found out to be : {'alpha': 0.1}
```

Metrics Evaluation

```
MAE : 285.44750969732405
MSE : 141368.15590889205
RMSE : 375.9895688830902
MAPE : 1373.346929361722
r^2 score : 0.4971362789701247
```

R^2 score for ridge regression = 49%

Graph of Actual and Predicted values



K Nearest Neighbor Regression

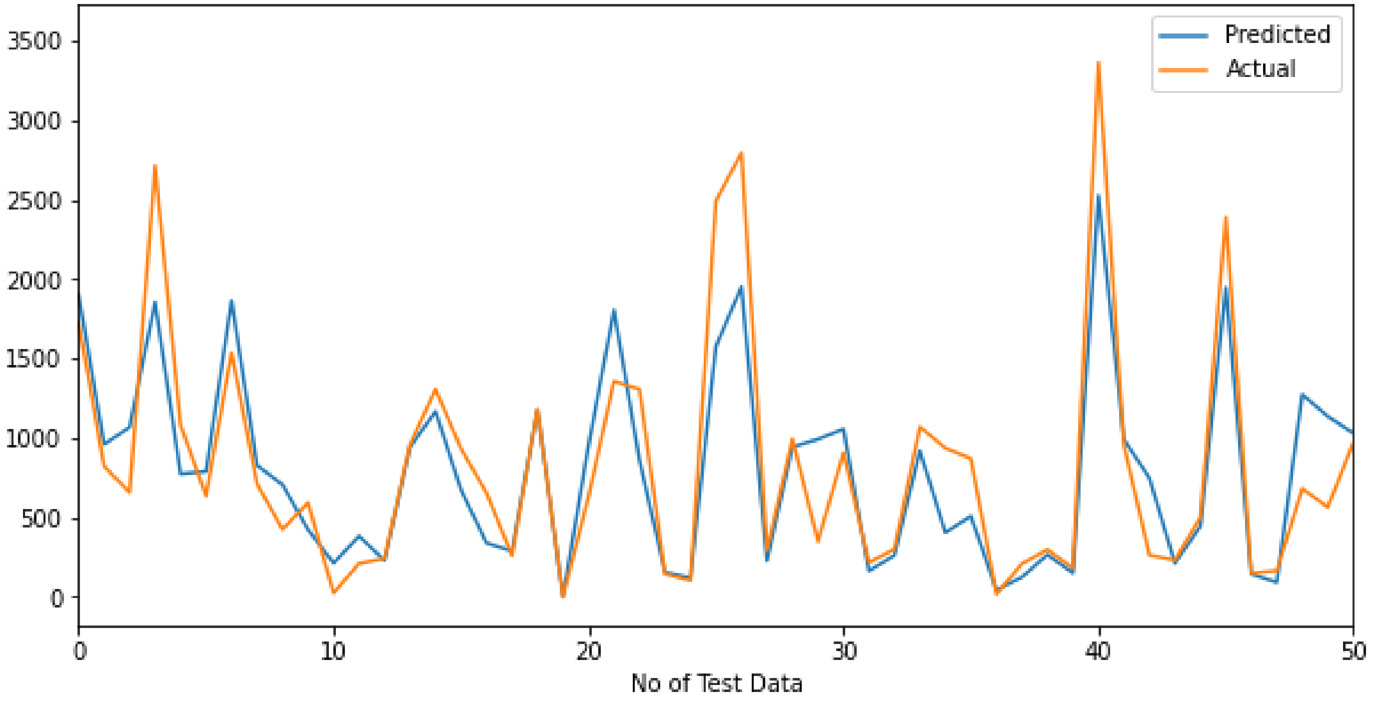
```
params = {'n_neighbors':[2,3,4,5,6,7,8,9]}
```

Metrics Evaluation

```
MAE : 205.59574771689498
MSE : 108227.07088327626
RMSE : 328.9788304485203
MAPE : 450.72825292559156
r^2 score : 0.6923857866475096
```

R^2 score for knn = 69%

Graph of Actual and Predicted values



Support Vector Machine

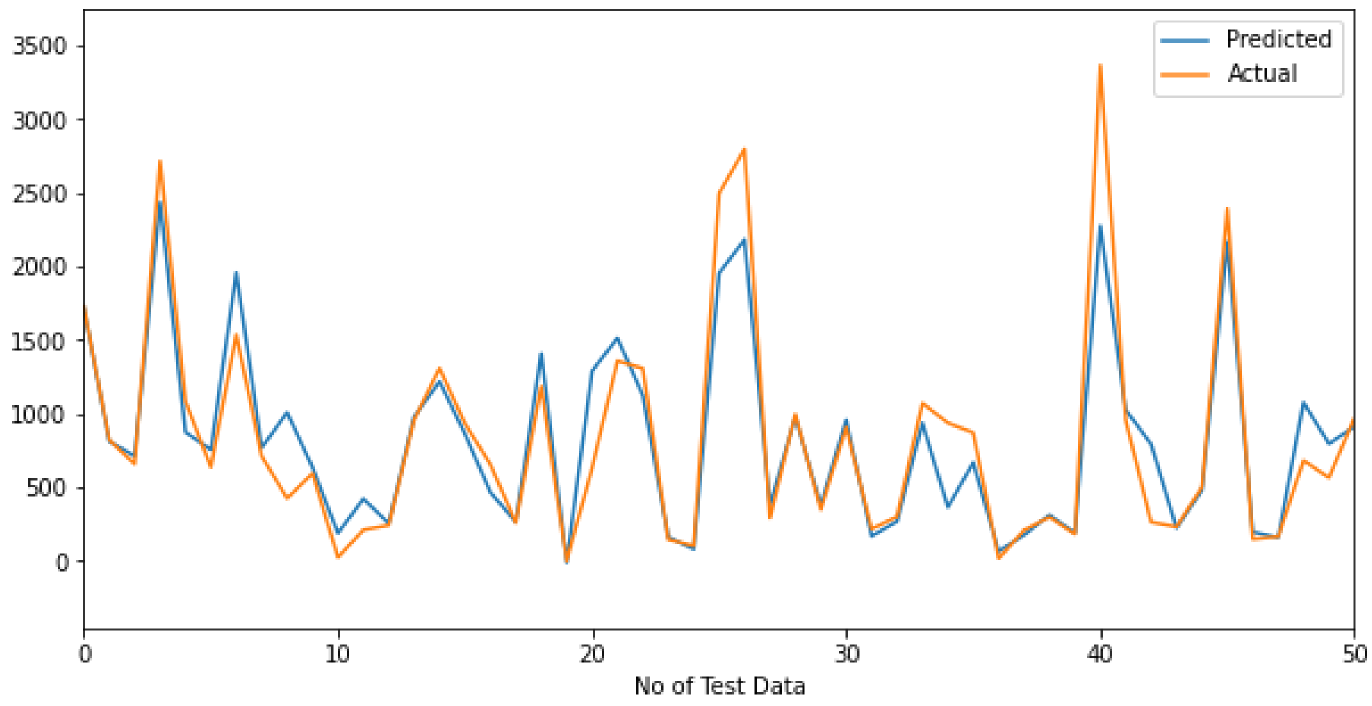
```
param_grid = {'C': [0.1, 1, 10, 100, 1000],
              'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
              'kernel': ['rbf']}
```

Metrics Evaluation

```
MAE : 163.40531722342553
MSE : 73896.93641829322
RMSE : 271.8399095392235
MAPE : 200.5080876226743
r^2 score : 0.7954733972626448
```

R^2 score for SVM = 79%

Graph of Actual and Predicted values



XGBoost Regression

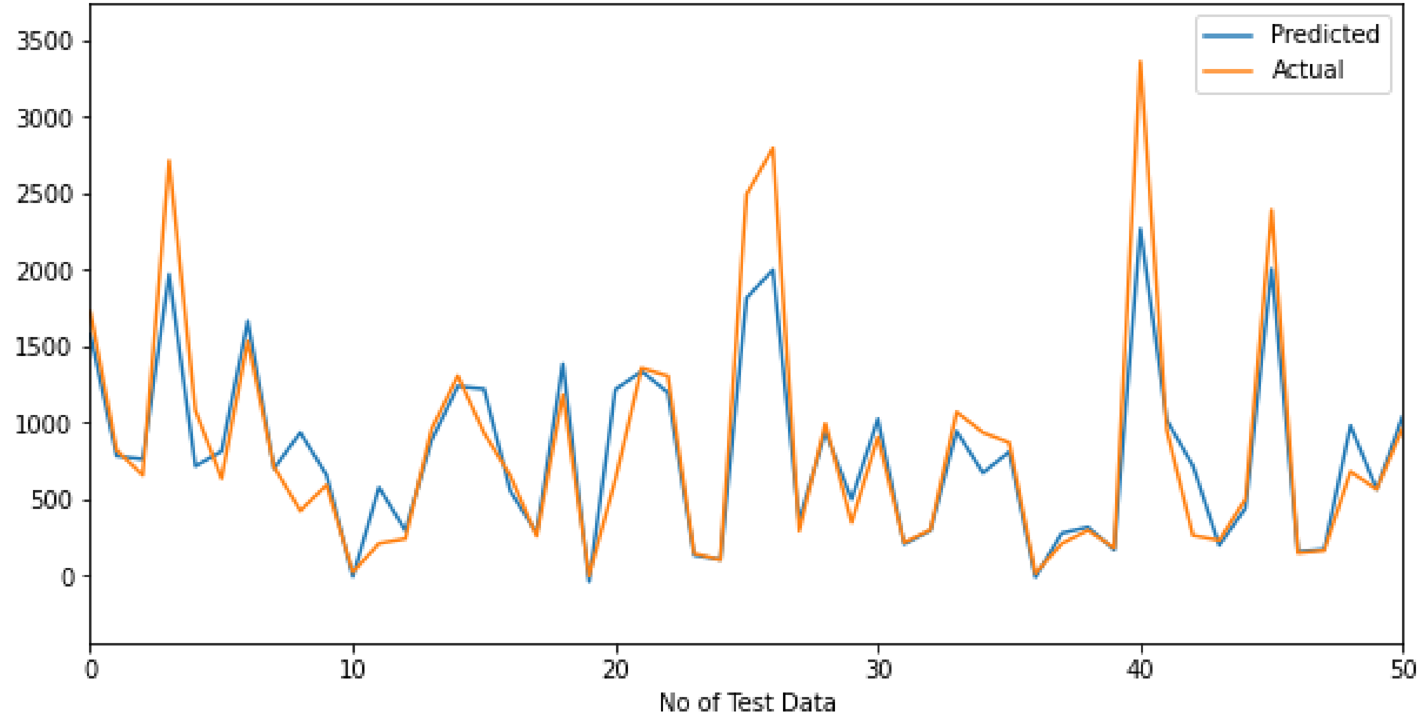
```
param_grid={"max_depth": (6,7), "learning_rate": (0.06, 0.08),
            "n_estimators": (400, 600), "subsample":(0.7,0.8),
            "colsample_bytree":(0.4,0.5), "gamma" : (1.4,1.5)}
```

Metrics Evaluation

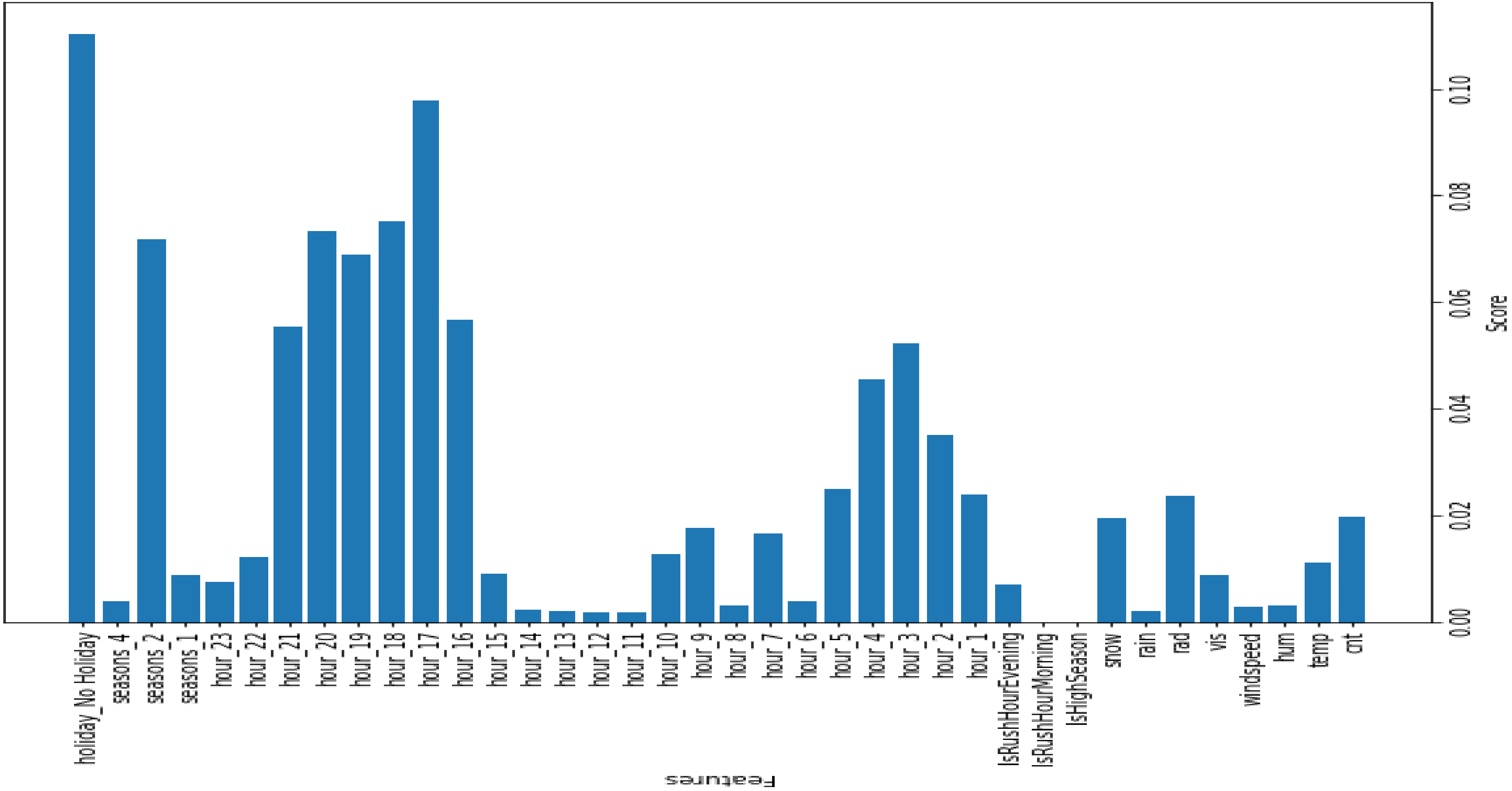
```
MAE : 153.69188623048672
MSE : 58127.081189772434
RMSE : 241.09558517271202
MAPE : 468.74455746685777
r^2 score : 0.8315591191704701
```

R^2 score for XGBoost = 83%

Graph of Actual and Predicted values



Feature importance



Metrics Comparison

Model_Name	MAE	MSE	RMSE	MAPE	r2_score
Ridge regression	285.45	141368.16	375.99	1373.35	49.71
Linear regression	285.46	141391.15	376.02	1374.89	66.06
Lasso regression	285.40	141387.94	376.02	1363.69	66.07
Knn regression	205.60	108227.07	328.98	450.73	69.24
SVM	163.41	73896.94	271.84	200.51	79.55
XGBoost regression	153.69	58127.08	241.10	468.74	83.16

Conclusion

- ✓ XGBoost Regression has given the best r2_score of 83.16% and with the least MAE, MSE, RMSE, MAPE scores.

- ✓ Linear Regression doesn’t provide good fit of the data, so did Lasso. Ridge regression performance was even worse as it shrunk the parameters to reduce complexity.

- ✓ Knn has performed better than Linear and Lasso regression.

- ✓ SVM performs relatively well than KNN, as it can easily handle multiple continuous and categorical variables.

- ✓ Features like days of no holidays, evening hours(17, 18, 19 and 20) and Autumn season plays important role in predicting the number of rented bikes.

Thank You