Seoul Bike Sharing Demand Prediction

(Vineeta Singh, Tushar Gupta)Data science trainees,
Alma Better , Bangalore

Abstract:

Bike Sharing System is an emerging modeof transport in the world and most of the developing countries are on the path of following the western model of Bike Sharing Systems. In India, some entrepreneurs have tried to set up a bike-share system and have failed in the past as they have failed to use data analytics properly. There is a possibility that bike stations can be full or empty when a traveler comes to the station. Thus to predict the useof such a system can be helpful for the usersto plan their travels and also for the entrepreneurs to set up the system properly. This paper presents different ways to predictthe number of bikes that can be rented in such a system, for case study purposes we have used a public data set. The predictionsare made for every hour of a day.

Keywords: Exploratory Data Analysis,Train-Test split,
Machin learning model,(LR,LS,RR,ER,Poly.Features,DT,RF,GB,XGB)

• Problem Statement

Currently, Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make therental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the citywith a stable supply of rental bikes becomesa major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental

The model which returned the **highest Quality Listing** within a certain radius based on the

- What can we learn from booking of bike for a different date, day, and year
- What can we learn from predictions?(ex: hour, weather,holiday etc)
- Which days are the busiest and why?
- Is there any noticeable difference in booking bikes on different functioning days, sessions, holidays,and what could be the reason for it?
- we will conduct a demand as per the Season and working day of the week.
- If mined properly, Data can tell us alot about the customer mindset, their expectations, and how well those were met
- The weather condition also having major importance for predicting the demand
- Some of Day’s have the most listing or demanding throughout the year.

• Introduction

Bike-sharing systems allow users to takeone-way bike trips over short distances. Generally, these systems are operated viaautomated kiosks to save manpower and reduce waiting time for the users. Bike Sharing System ensures that pollution is
reduced as with the use of bicycles there is areduction in the use of motor vehicles which leads to a reduction in emission of pollutantsin the air. This practice of Bike Sharing Systems is common in Western Countries
while the same is not seen yet in countries like India. In India, most of the bike-sharing systems could not achieve their maximum potential as data analysis was not used properly. The advantages of this system are that we can have public bike stations
without any human involvement. However, the popularity of the bike-share system increased drastically which led to creating agap between the supply and demands of bikes and docks at bike stations. And the most common issues faced by the users are the lack of bikes and docks available at bikestations. The growing concern led the bike operators to consider the matter seriously,
and

• Related Work

Since the last decade, a lot of work has been presented on the bike-sharing systems but very few actually aim to quantitatively predict the demand at a bike station. Initial studies involved the application of optimization algorithms which were proven to be ineffective for the situation However, the application of machine learning models for bike-share networks provided significant results which are briefly described in the sub-sections. The following subsections are structured as follows;

- provides information on the data transformation techniques utilized in related works,
- Illustrates the details of widely used machine learning models for bike-shareprediction.

• Data information & DataTransformation

The nature of the bike share data limits the option of methods, which can be utilized foranalysis. Most of the bike share data consistof bike trip records and station location records, which usually do not include bikes and docks demand attributes. Hence, most
studies usually focus on analyzing the demographics of the data and how it affects the system

We had to perform a few imputations and transformations on our dataset for us
to

create the desired visualizations. There wereno major inconsistencies or mismatches in the data. We rename some columns and Extract useful information from the date
column. Our data set have the value:-

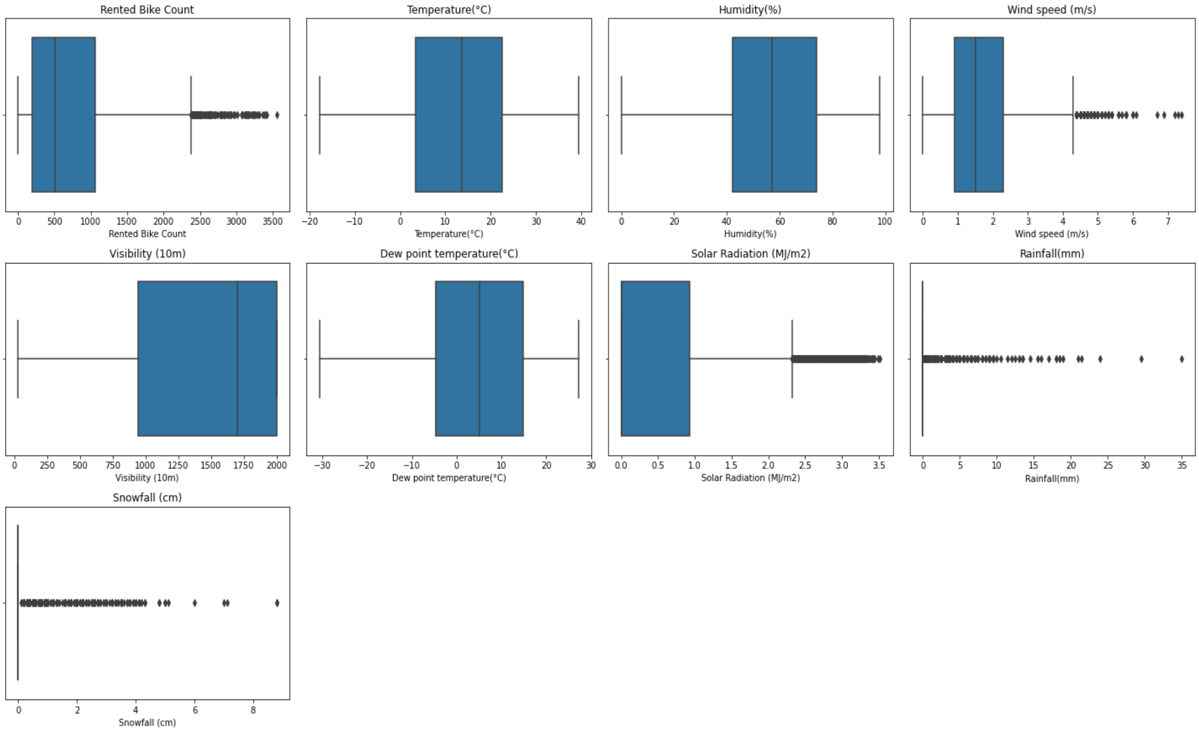
Date', 'Rented Bike Count', 'Hour','Temperature(°C)','Humidity(%)', 'Wind speed (m/s)','Visibility (10m)','Dewpoint temperature(°C)','Solar Radiation (MJ/m2)','Rainfall(mm)','Snowfall(cm)','Seasons', 'Holiday', 'FunctioningDay

• Machine learning Models:

A bike-share system data majorly constitute stime-dependent features. These features fluctuate randomly making it impossible to build a predictive model using static stochastic time series techniques. We startfitting our feature or data from Linear Regression Model and then step-wise move forward to Lasso, Ridge and Elastic regression to make more improvement of the linear model. we also try to fit data on the decision tree and visualize the tree . Random Forest also gives a better result then move forward for the Gradient boosting and we find that model performance get increases but score still below 88% so we used next Model thatis XGBoost and fit the data to this model and achieve performance more than 98% on the training data

• Dealing with Outliners:

We see no outlier in the data set so no worryfor dealing with outlier. we just make our focus on data extraction and correlation



	Date	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Dew point temperature(°C)	Radiation (MJ/m2)	Rainfall(mm)
count	8760	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000
unique	365	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	19/12/2017	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	24	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	704.602055	11.500000	12.882922	58.226256	1.724909	1436.825799	4.073813	0.569111	0.148687
std	NaN	644.997468	6.922582	11.944825	20.362413	1.036300	608.298712	13.060369	0.868746	1.128193
min	NaN	0.000000	0.000000	-17.800000	0.000000	0.000000	27.000000	-30.600000	0.000000	0.000000
25%	NaN	191.000000	5.750000	3.500000	42.000000	0.900000	940.000000	-4.700000	0.000000	0.000000
50%	NaN	504.500000	11.500000	13.700000	57.000000	1.500000	1698.000000	5.100000	0.010000	0.000000
75%	NaN	1065.250000	17.250000	22.500000	74.000000	2.300000	2000.000000	14.800000	0.930000	0.000000
max	NaN	3556.000000	23.000000	39.400000	98.000000	7.400000	2000.000000	27.200000	3.520000	35.000000

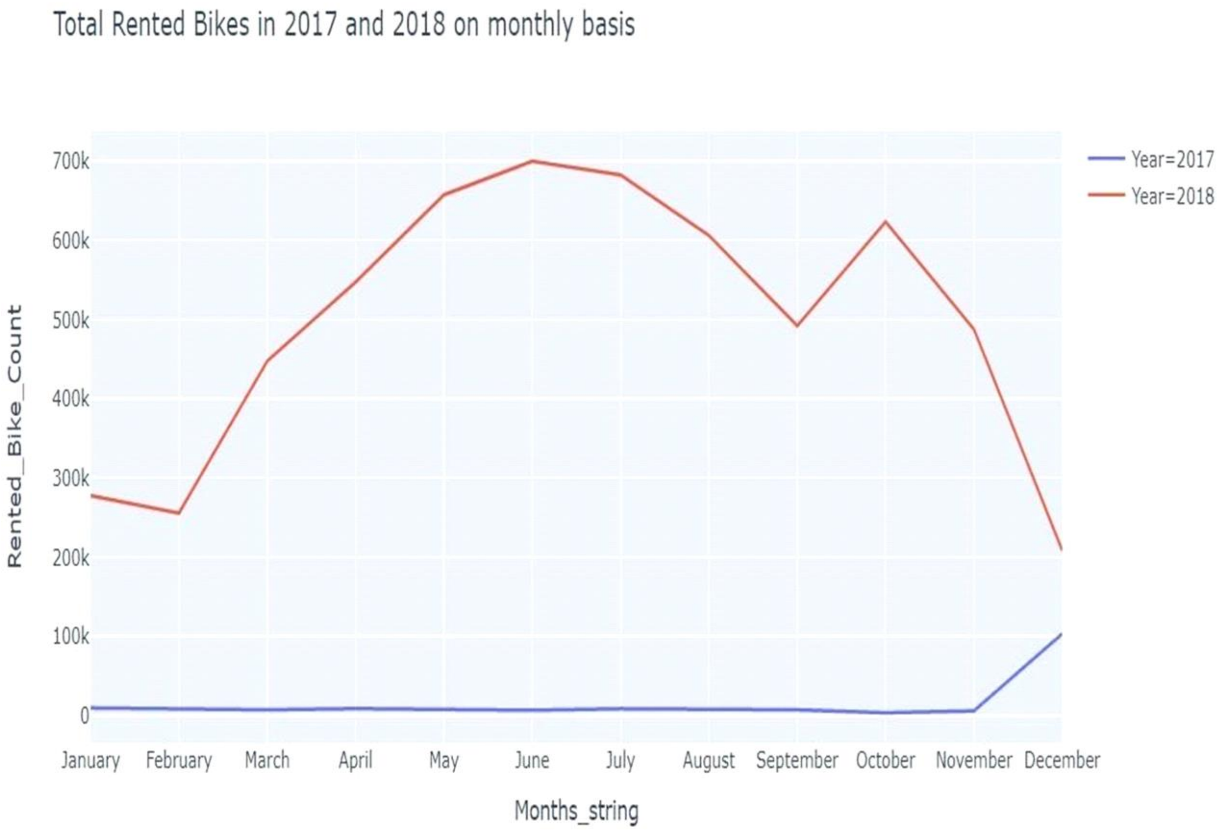
- # Methodology:

The existing methodologies for predictions are regression, decision trees, random forest,Gradient Boosting, XGBoost etc. This research work allows to have insight of the performance of various prediction algorithms and walk through the whole process of prediction.

- **Data pre-processing and transformation**
- **Developing and optimizing theLinear Regression model**
- **Developing and optimizing theLasso Regression model**
- **Developing and optimizing RidgeRegression model**
- **Developing and optimizing ElasticNet Regression model**
- **Developing and optimizing Polynomial Regression model**
- **Developing and optimizingDecision Tree**
- **Developing and optimizingRandom forest**
- **Developing and optimizing Gradient Boosting**
- **Developing and optimizing XtreamGradient Boosting**

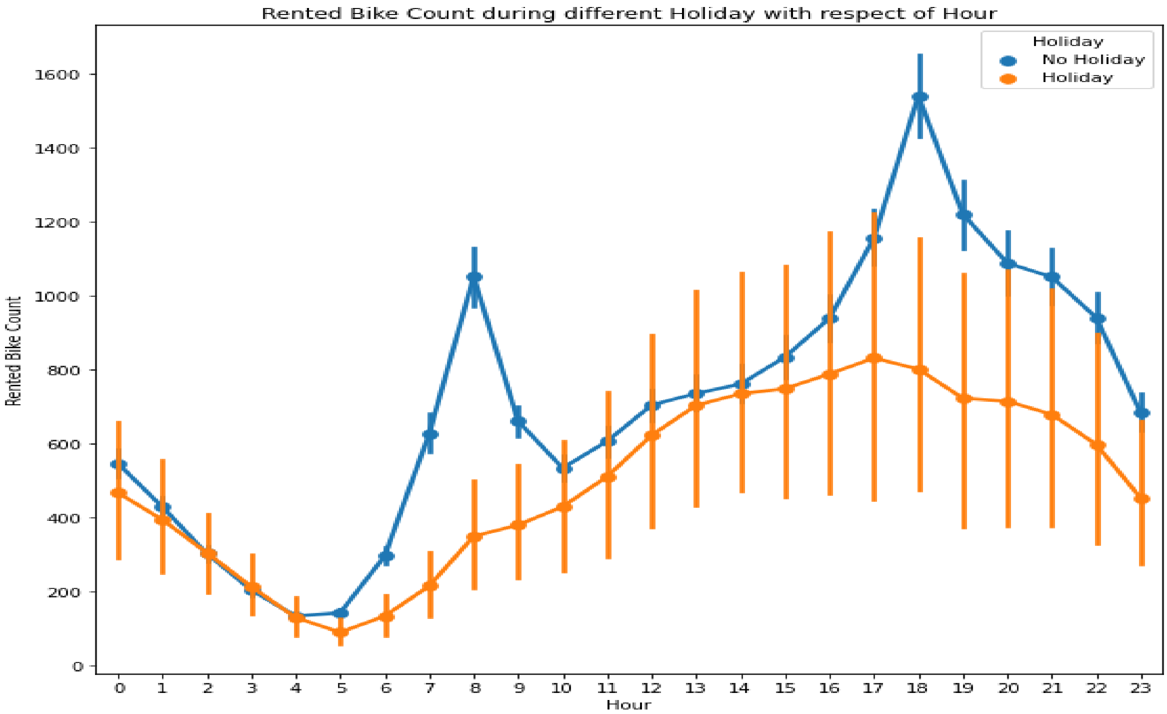
- **Data pre-processing and transformation**

In pre-processing, we extract the information from the date string for finding the booking prediction for the day year and week. The unwanted features and missing values were dropped from the newlyformed data set. The dataset created was essential for developing graph-structured data, which is a necessity for the proposed graph convolution models. The structure ofthe dataset is shown in the figure

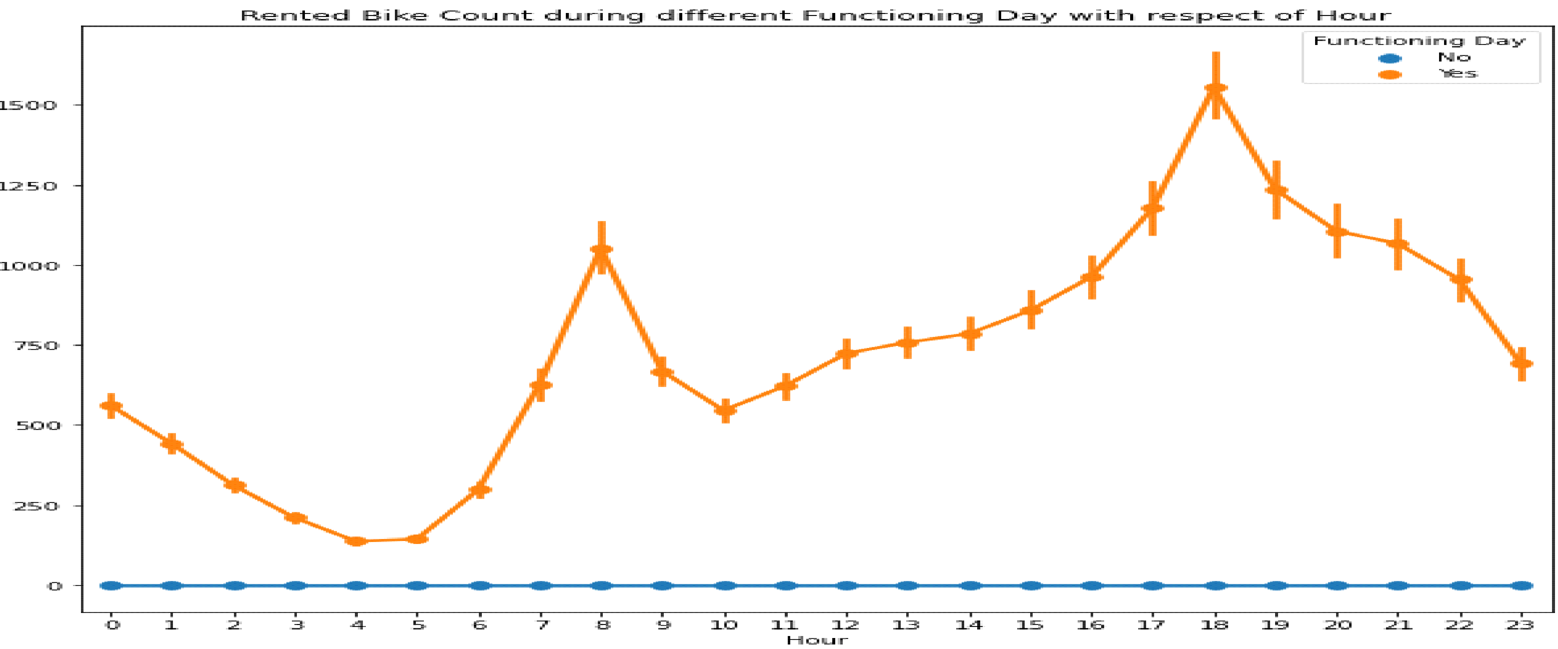


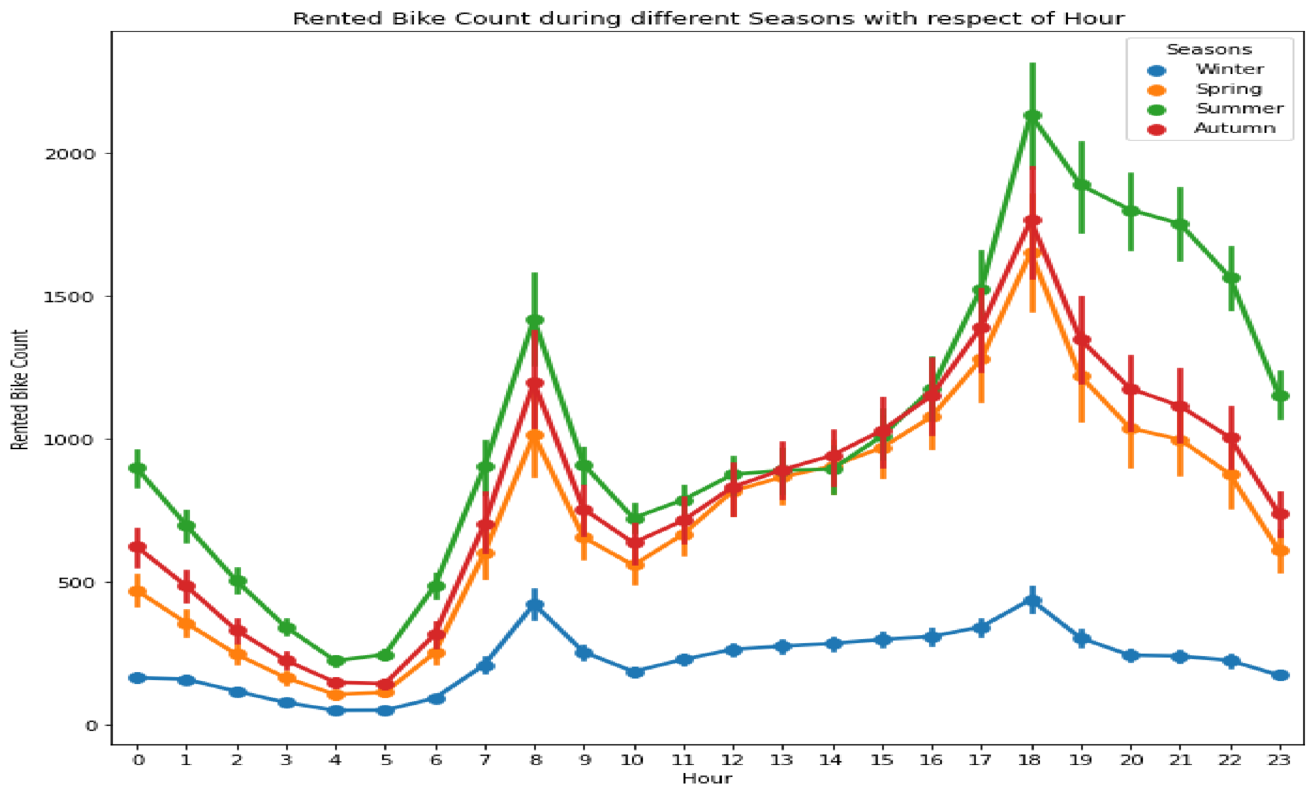
This is a basic graph that shows that in theyear of 2018 demand increasing rapidly, summer months are more demanding throughout the year and winter days are less demanding.

Which days in a week are more rented bike count?

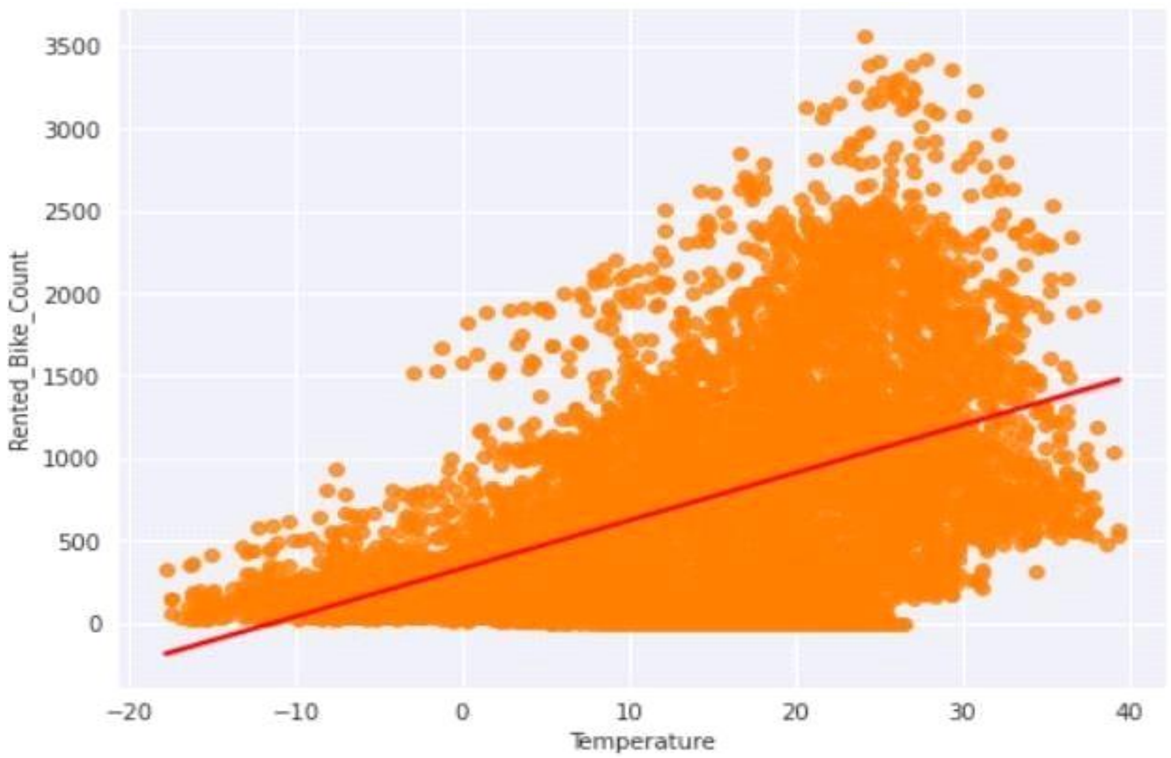


In the Holiday column, The demand is low during holidays, but in no holidays the demand is high, it may be because people use bikes to go to their work.





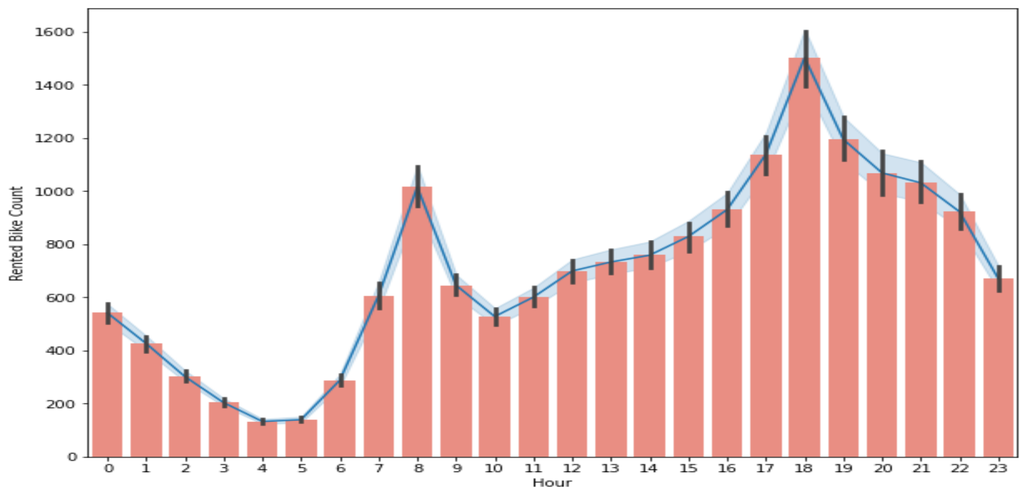
Effect of temperature on bike demand count:



Graphs give the reflection about the demand when the temperature of the weather gets increases people demanding more for booking bikes.

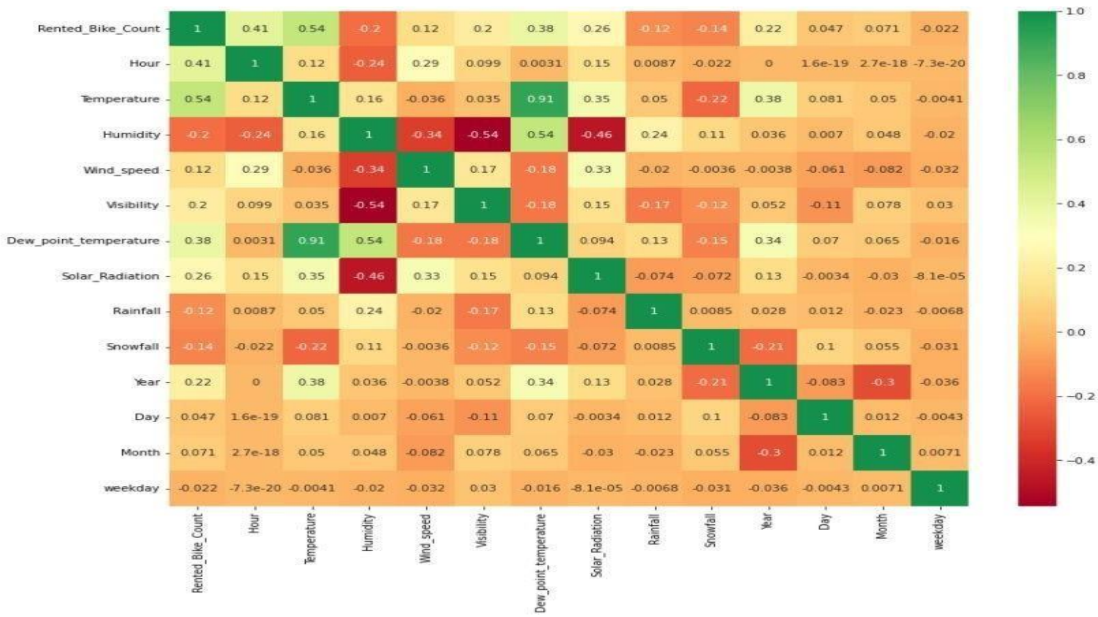
Bike Rent per hour:

Every product-based company has atendency to increase the price of the product as the demand increases, we observe some pattern for the bike rented count for a set duration of time.



Correlation between independent features:

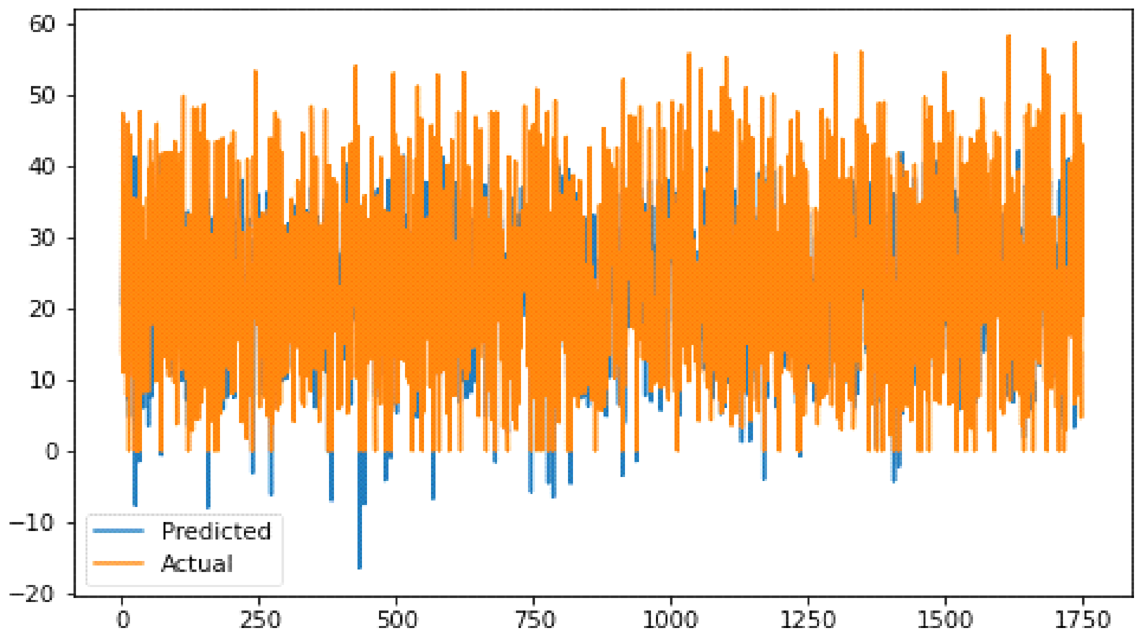
We can see the lots of weather parameters like temperature and Dew point temperature etc are correlated to each so in the next step we drop some of the features



Data set after the feature Engineering and dummy variable:

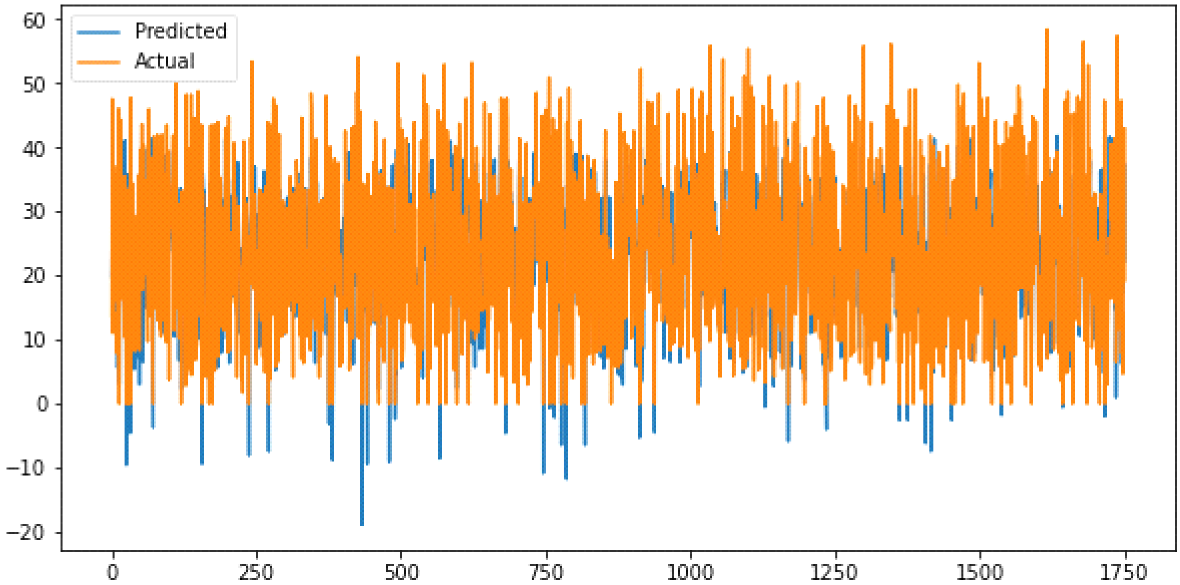
It is a process in which analysts use domain knowledge about the data and create new features in the data set in a way such that thenew features help in improving the model accuracy. There is no definite path for feature engineering, but it depends on the skills of the analyst and the type of data. Feature engineering needs to be done on both training and testing data and is a very important part of building a good prediction model. We used One Hot Encoding to produce binary integers of 0 and 1 to encodeour categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to the numerical format. Create one hot coding for a different seasons.

Wind_speed	Visibility	Solar_Radiation	Rainfall	Snowfall	Year	Day	Month	weekday	Temperature_and_DP_Temp	Seasons_Spring	Seasons_Summer	Seasons_Winter
2.2	2000	0.0	0.0	0.0	2017	12	1	4	-22.8	0	0	1
0.8	2000	0.0	0.0	0.0	2017	12	1	4	-23.1	0	0	1
1.0	2000	0.0	0.0	0.0	2017	12	1	4	-23.7	0	0	1
0.9	2000	0.0	0.0	0.0	2017	12	1	4	-23.8	0	0	1
2.3	2000	0.0	0.0	0.0	2017	12	1	4	-24.6	0	0	1



- **Developing and Optimizing Linear Regression model:**

Linear regression model gives up to 77.71% metrics score on the train as well on test data. linear regression model work with lotsof assumptions



- **Developing and optimizingLasso Regression model**

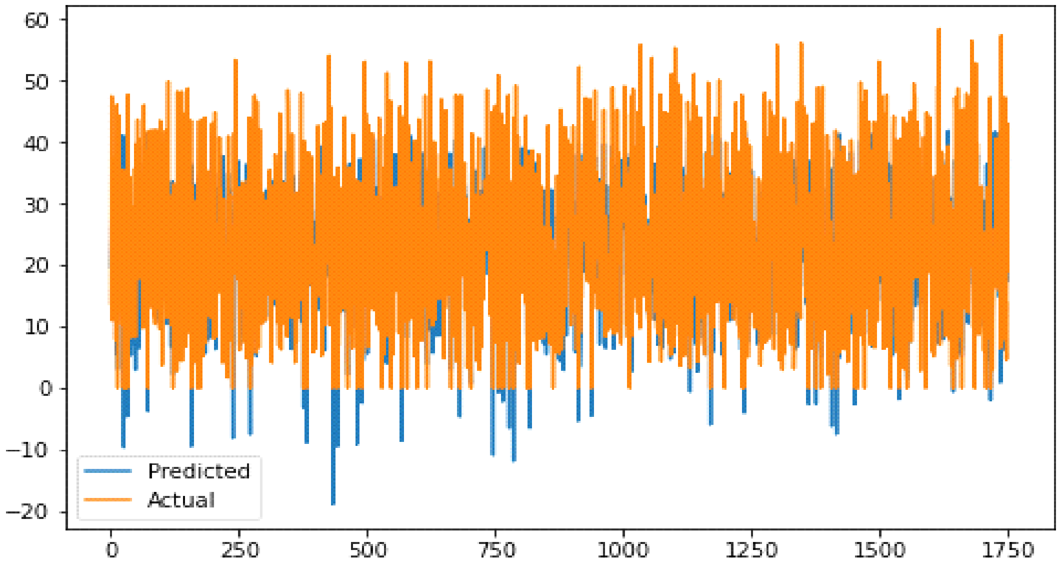
Lasso is variable panelized regression method it deletes the less performing feature .lasso regression gives fewer metrics score than the normal linear regression both on the train and test data approx- 77.40%

-

- **Developing and optimizingRidge Regression model**

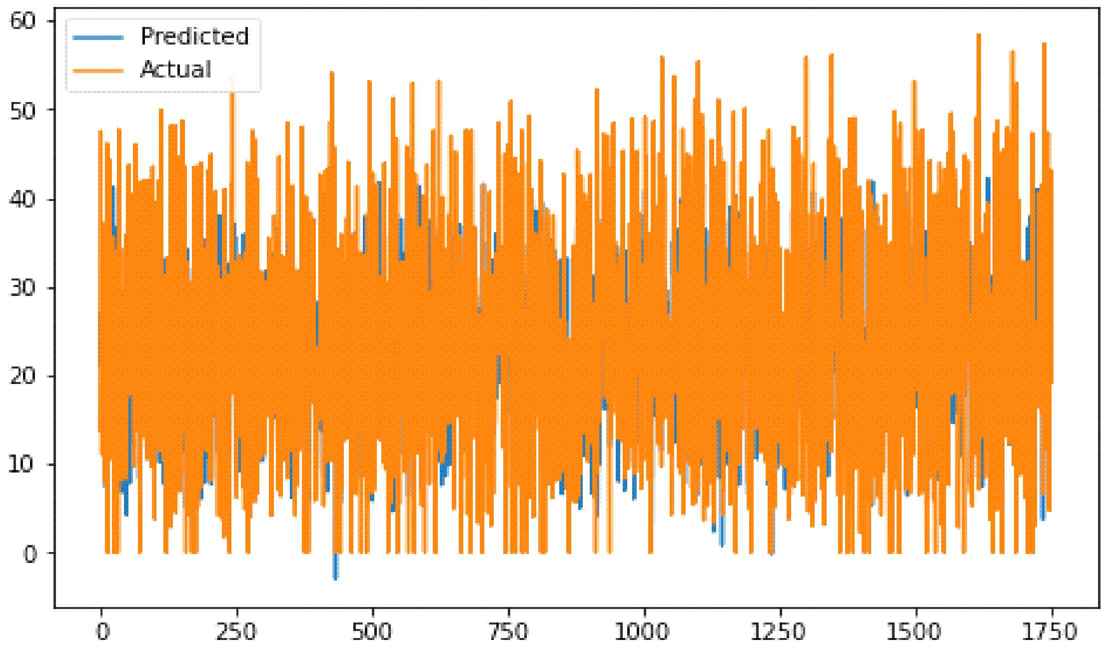
Ridge regression making the features

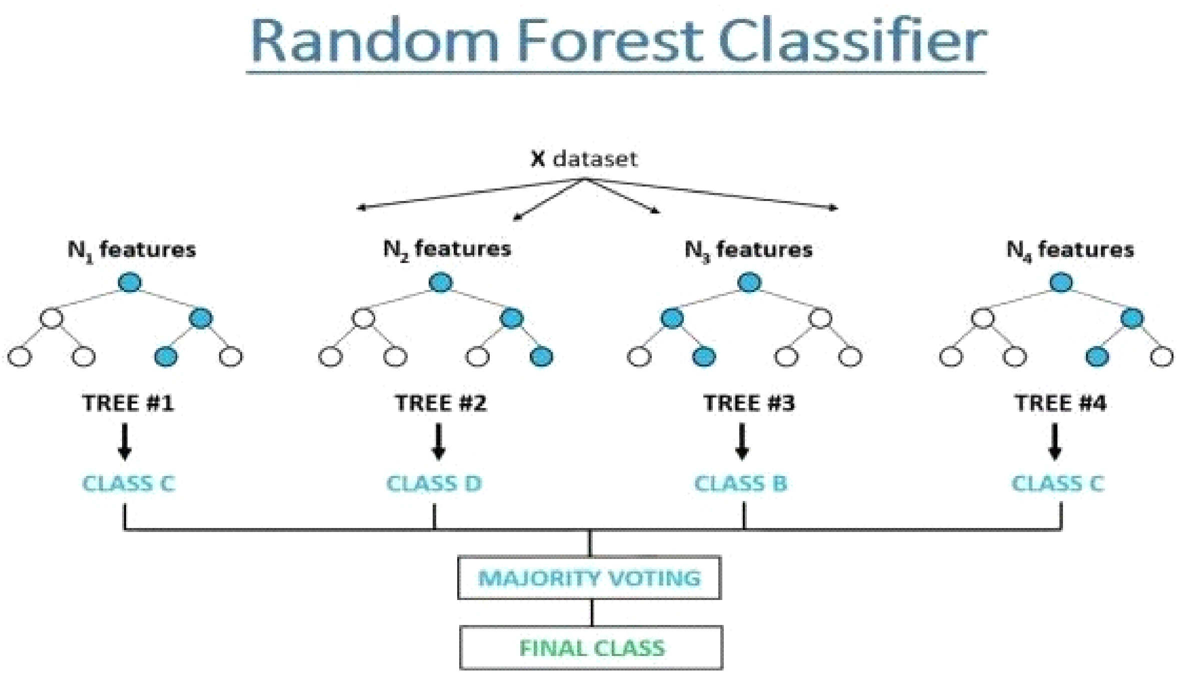
coefficient optimization. Its metrics show some improved result comparison to lassoregression



- **Developing and Optimizing ElasticNet Regression model**

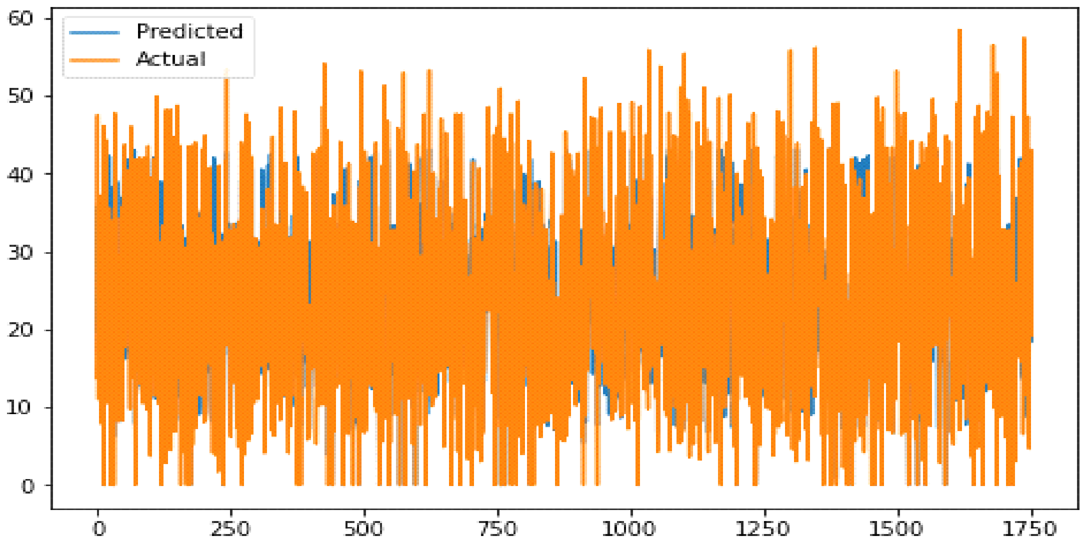
Elastic net is avg of lasso and ridge regression so its metrics score is not looking good less the lasso and ridge regression. it has a 77.71% score on train data and a 77.39% score with test data





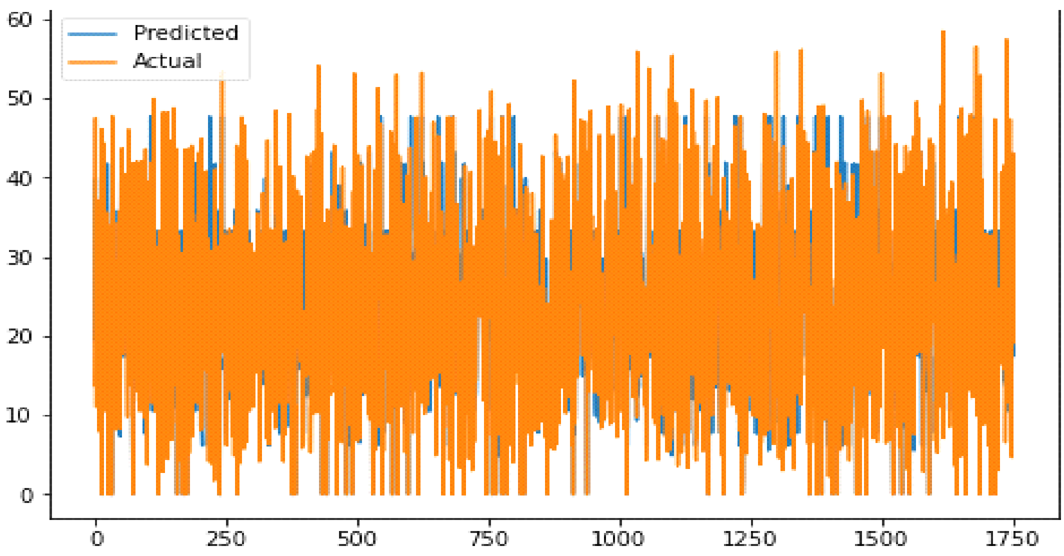
• Developing and optimizing Random ForestTree

Random Forrest have the most peak metricsscore 99% for train and 93.21% for test data set but when we did some cross-validation so this metrics come with a train score of 77.71% and test score 77.40%,which must satisfactory.



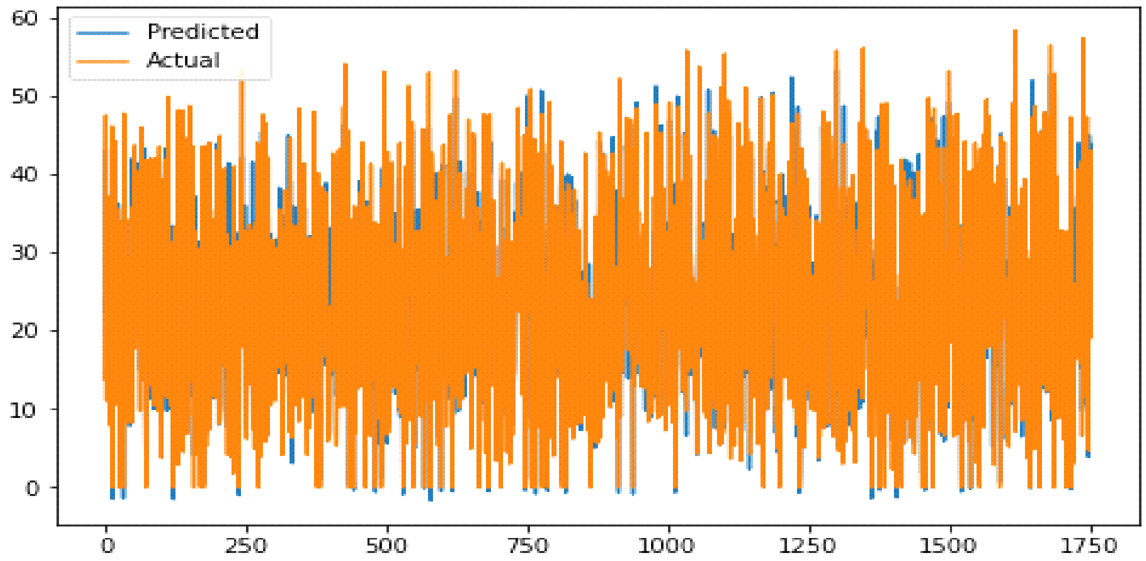
• Developing and optimizingDecision Tree

Decision tree showing better metrics scorethen the ridge lasso. On the train data set ithas an 66.43% score and on the test data set ithas an 64.50% score.



Developing and optimizingGradient Boosting

It gives best metrics score for training dataset approx 88% and for the test data set

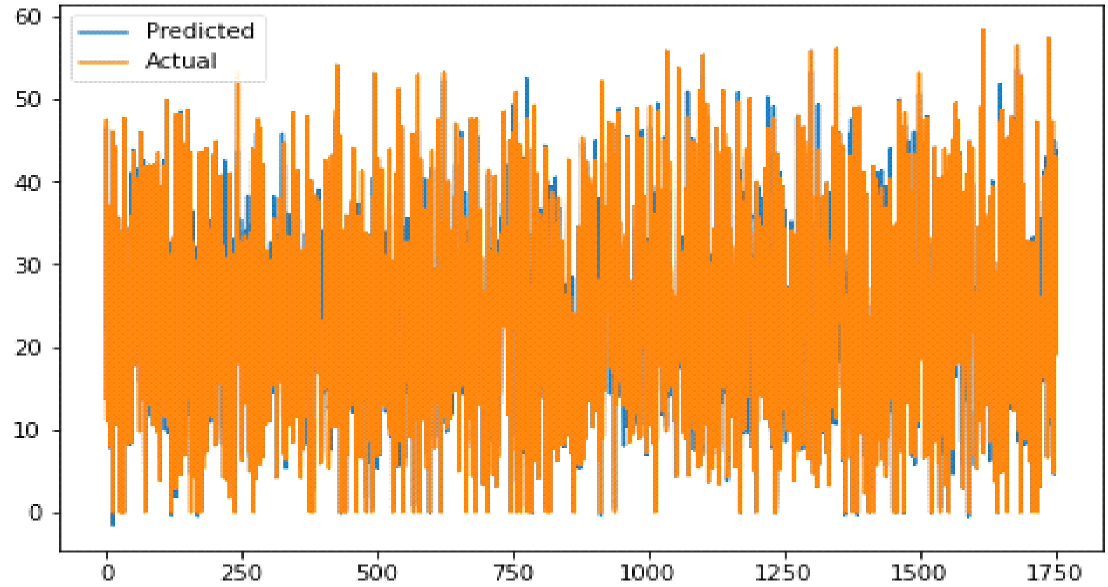


approx 94% and after the cross validation the metrics score would be 77.71% for training data and 99% which is best from the randomforrest.

Developing and optimizing Xtream Gradient Boosting (XGB)

XGBoost Given one of the best results for the training as well as for test data set aftercross validation or training our data set.

Metrics score would be 98% for train data and 93% for the test data, which is best outof the rest one.



• Conclusion

This study proposed the use machine learning techniques to identify the demandsin a bike-sharing system. The Nine algorithms are applied on the bike share dataset for predicting the count of bicycles that will be rented per hour

We got some good results and accuracy with random forest, Gradient boosting, and Xgboost by using Cross validation. Theaccuracy and performance has been compared between the models using Root Mean Squared Logarithmic Error (RMSLE) and R squared .If these systems include the use of analytics the probability of building a successful system will increase.

REFERENCES

- K. Gebhert, R. Roland,"Impact of Weatheron Capital Bike Share Trips" presented at the 92ndAnnual Meeting of the Transportation Research Board 2013.

- J. Yoon, Fabio Pinelli, "City ride: a predictive bike sharing journey advisor"2012 IEEE 13th International Conference on MobileData Management.

- R. Giot, R.Cherrier"Predicting Bike ShareDemand upto One hour ahead" 2013 IEEE 9thInternational Confrence on Data Management,France.

- I.Frade, A.Ribbero, "Bicycle SharingSystems Demand" unpublished.
- T. Rui, Lin Li Hua"Quantitative Research onVehicle Exhausts Pollution in the City" Published in 2012.
- Y. Zhang, Z Huang "Performance evaluationof bike sharing system in Wuchang area of Wuhan, China", 6th China Planning Conference(IACP), 2012

- Kaggle: Bike Share Demand ,"https://www.kaggle.com/c/bike-sharingdemand"

- Using Gradient Boosted Trees to Predict Bike Share Demand, "http://blog.dato.com/using-gradient-boosted -trees-to-predict-bikesharing-demand"