# **Seoul Bike Sharing Demand Prediction**

**(Raavi sahu)**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract:**

Bike Sharing System is an emerging mode of 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 use of such a system can be helpful for the users to plan their travels and also for the entrepreneurs to set up the system properly. This paper presents different ways to predict the number of bikes that can be rented in such a system, for case study purposes we have used a public data set. The predictions are made for every hour of a day.

***Keywords: Exploratory Data Analysis,Train-Test split, Machine learningmodel,(LR,LS,RR,DT,RF,GB,XGB)***

**1.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

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: locations, hour, weather, 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 a lot 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.

The giving problem statement makes generate the intuition for the research :

**2. Introduction**

Bike-sharing systems allow users to take one-way bike trips over short distances. Generally, these systems are operated via automated 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 a reduction in the use of motor vehicles which leads to a reduction in emission of pollutants in 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 a gap 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 bike stations. The growing concern led the bike operators to consider the matter seriously, and

## **3. 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; 3.1 provides information on the data transformation techniques utilized in related works, 3.2 illustrates the details of widely used machine learning models for bike-share prediction.

## **3.1 Data information& Data Transformation**

The nature of the bike share data limits the option of methods, which can be utilized for analysis. Most of the bike share data consist of 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 were no 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(%)',**

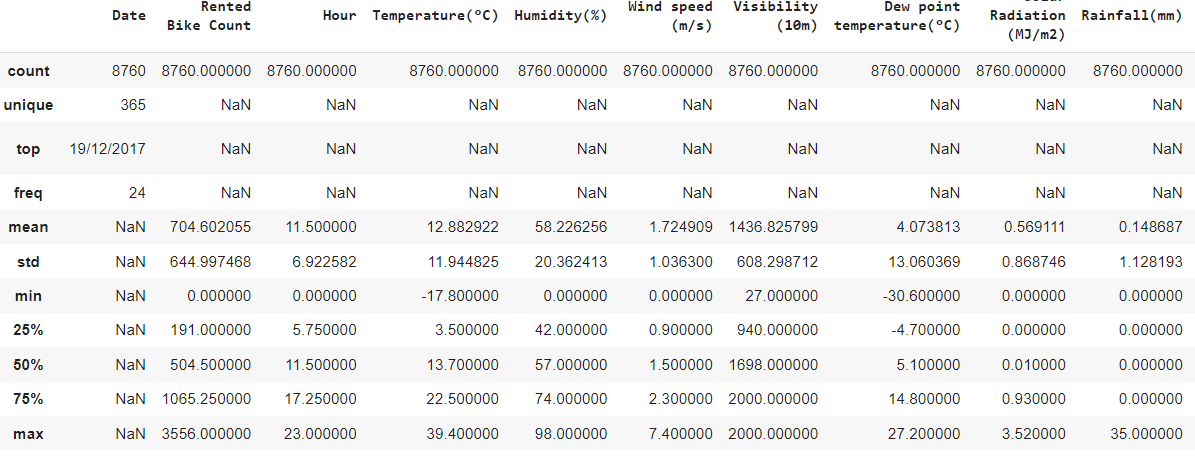
**'Wind speed (m/s)','Visibility (10m)','Dew point temperature(°C)','Solar Radiation (MJ/m2)','Rainfall(mm)','Snowfall (cm)','Seasons', 'Holiday', 'Functioning Day**

# **3.2 Machine learning Models:**

A bike-share system data majorly constitutes time-dependent features. These features fluctuate randomly making it impossible to build a predictive model using static stochastic time series techniques. We start fitting our feature or data from Linear Regression Model and then step-wise move forward to Lasso and regression to 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 80% so we used next Model that is XGBoost and fit the data to this model and achieve performance more than 82% on the training data

**4. Dealing with Outliners:**

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



**5. 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 the Linear Regression model**
* **Developing and optimizing the Lasso Regression model**
* **Developing and optimizing Ridge Regression model**
* **Developing and optimizing ElasticNet Regression model**
* **Developing and optimizing Decision Tree**
* **Developing and optimizing Random forest**
* **Developing and optimizing Gradient Boosting**
* **Developing and optimizing Xtream Gradient Boosting**

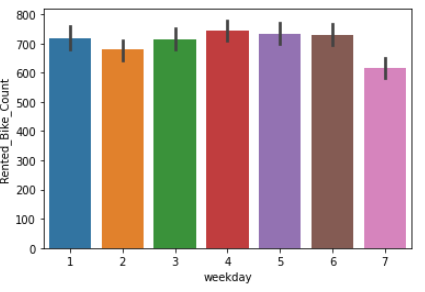
**5.1Data pre-processing and transformation**

In pre-processing, we extract the information from the date string for finding the booking prediction for the day year and season. The records present in trip start time (bike trip dataset) were iterated and added to a respective interval in the time interval dataset to produce bike demand feature. Similarly, process was repeated with trip end time to give dock demand feature. The unwanted features and missing values were dropped from the newly formed data set. Moreover, the bike trip data was further processed to generate. The dataset created was essential for developing graph-structured data, which is a necessity for the proposed graph convolution models. The structure of the dataset is shown in the figure.

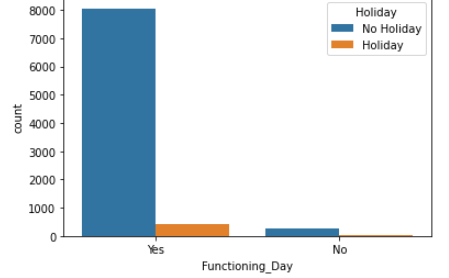


This is a basic graph that shows that in the year of 2018 demand increasing rapidly, summer months are more demanding throughout the year and winter days are less demanding.

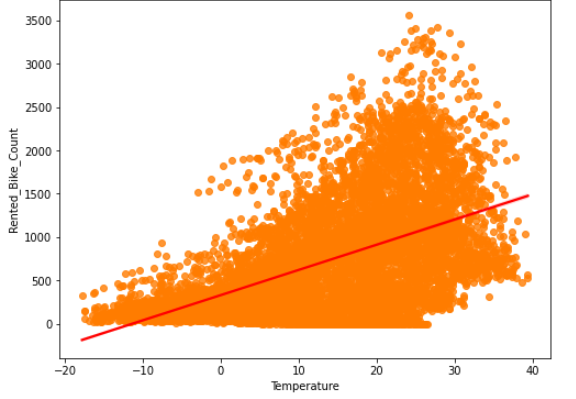
**Which days in a week are more rented bike count?**



In This graph, we can observe that weekends are less demanding and working days are more demanding throughout the day.



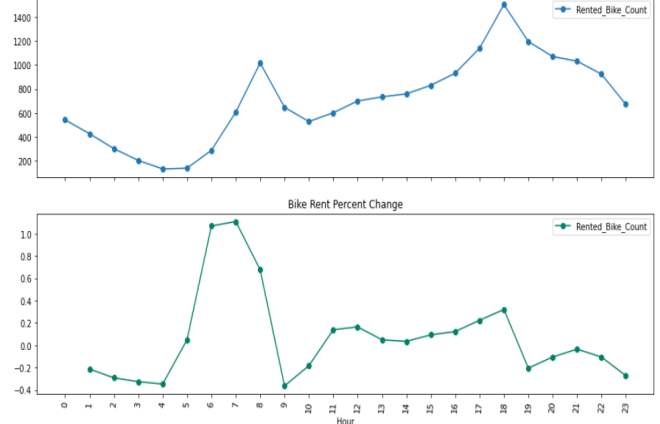
**Effect of temperature on bike demand count:**



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

**Bike Rent percentage change per hour:**

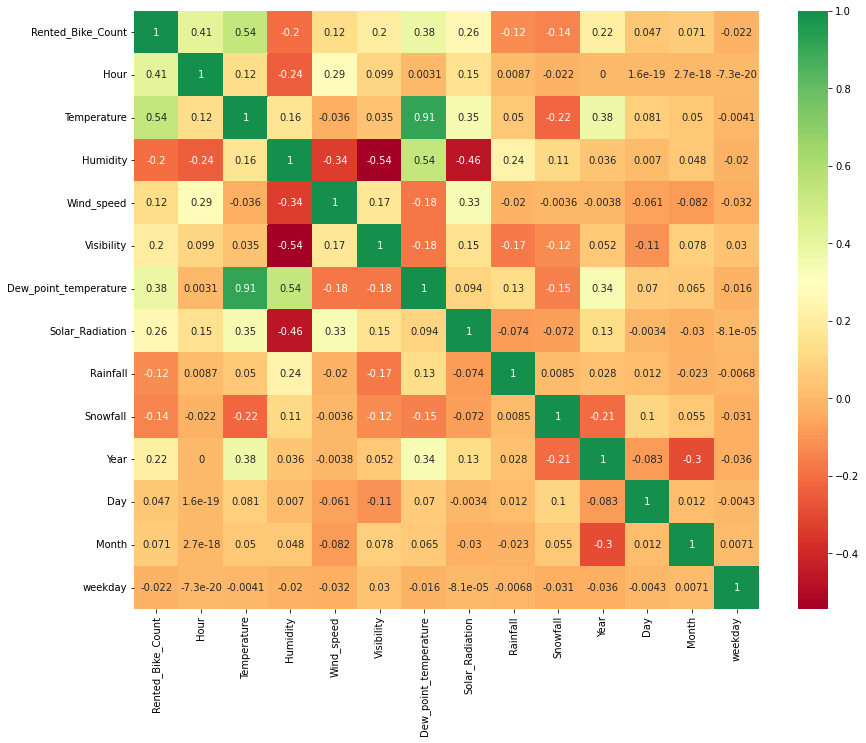
Every product-based company has a tendency to increase the price of the product as the demand increases we observe some pattern for the bike rented count for a set duration of time.



.

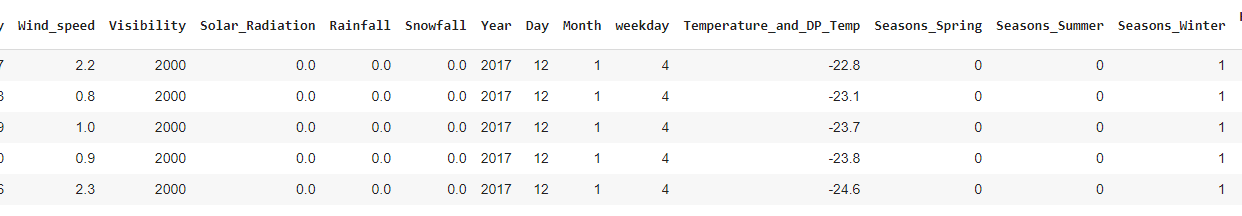
**Correlation between weather parameters:**

We can see the lots of weather parameter like temperature humidity etc are correlated to each so in the next step we neglect the some of the feature and use the single one.



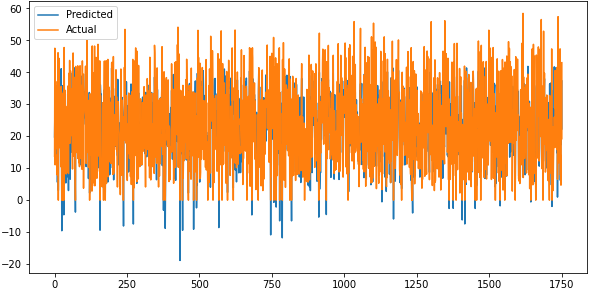
**Data set after the feature Engineering and dummy var:**

It is a process in which analysts use domain knowledge about the data and to create new features in the data set in a way such that the new 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 encode our 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.



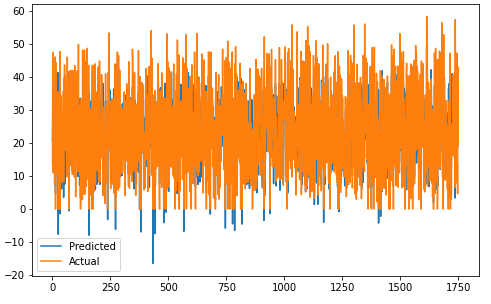
**5.2Developing and optimizing Linear Regression model:**

Linear regression model gives up to 65.6% metrics score on the train as well on test data. linear regression model work with lots of assumptions



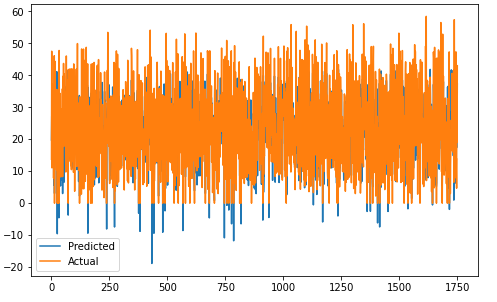
**5.3 Developing and optimizing Lasso Regression model**

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 64%



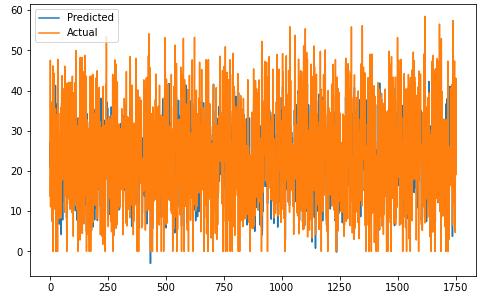
**5.4 Developing and optimizing Ridge Regression model**

Ridge regression making the features coefficient optimization. Its metrics show some improved result comparison to lasso regression



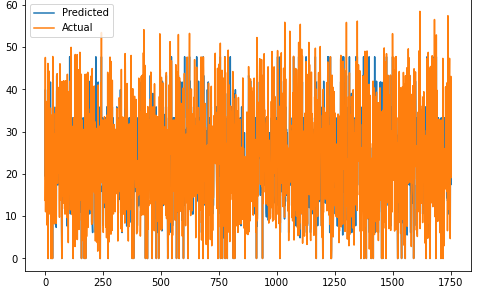
**5.5 Developing and optimizing ElasticNet Regression model**

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 57% score on train data and a 58% score with test data



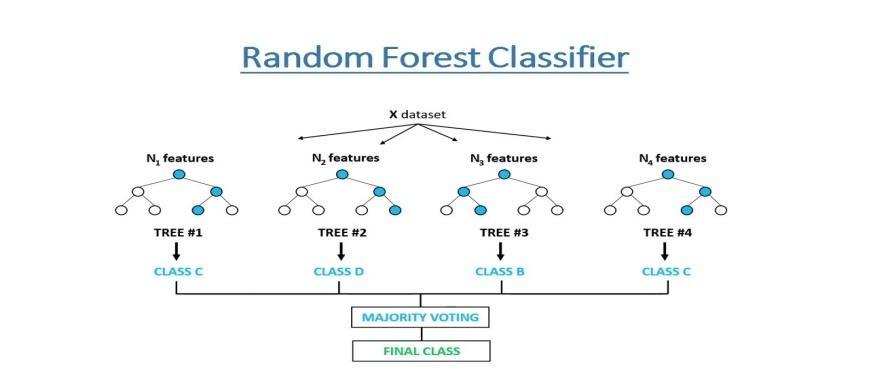
**5.5 Developing and optimizing Decision Tree**

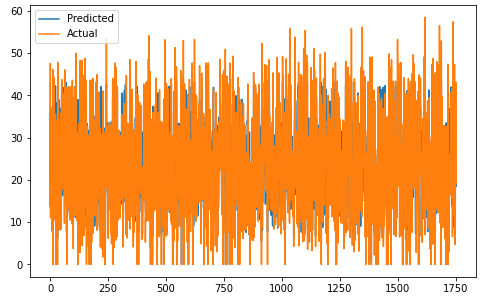
Decision tree showing better metrics score then the ridge lasso. On the train data set it has an 83% score and on the test data set it has an 81% score.



**5.6 Developing and optimizing Random Forest Tree**

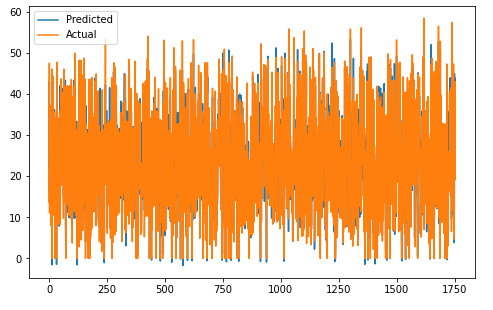
Random Forrest have the most peak metrics score 98% for train and 90% for test data set but when we did some cross-validation so this metrics come with a train score of 83% and test score 82%,which must satisfactory.





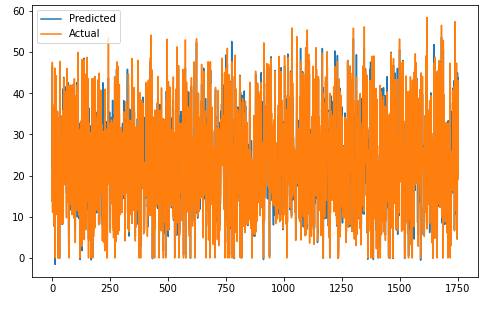
**5.6 Developing and optimizing Gradient Boosting**

it gives best metrics score for training data set approx 98% and for the test data set approx 94% and after the cross validation the metrics score would be 94% for training data and 91% which is best from the random forrest.



**5.6 Developing and optimizing Xtream Gradient Boosting (XGB)**

XGBoost Given one of the best results for the training as well as for test data set after cross validation or training our data set. Metrics score would be 97% for train data and 91% for the test data, which is best out of the rest one.



**6. Conclusion**

This study proposed the use machine learning techniques to identify the demands in a bike-sharing system. The seven 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 forrest, Gradient boosting, and Xgboost by using Cross validation. The accuracy and performance has been compared between the models using Root Mean Squared Logarithmic Error (RMSLE). If these systems include the use of analytics the probability of building a successful system will increase.

**7.REFERENCES**

[1] K. Gebhert, R. Roland,“Impact of Weather on Capital Bike Share Trips” presented at the 92ndAnnual Meeting of the Transportation Research Board 2013.

[2] J. Yoon, Fabio Pinelli, “City ride: a predictive bike sharing journey advisor”2012 IEEE 13th International Conference on Mobile Data Management.

[3] R. Giot, R.Cherrier“Predicting Bike Share Demand upto One hour ahead” 2013 IEEE 9th International Confrence on Data Management, France.

[4] I.Frade, A.Ribbero, “Bicycle Sharing Systems Demand” unpublished.

[5] T. Rui, Lin Li Hua“Quantitative Research on Vehicle Exhausts Pollution in the City” Published in 2012.

[6] Y. Zhang, Z Huang “Performance evaluation of bike sharing system in Wuchang area of Wuhan, China”, 6th China Planning Conference (IACP), 2012

[7] Kaggle: Bike Share Demand ,”https://www.kaggle.com/c/bike-sharingdemand”

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

[9] M. Meriem ,J. Usmail “A comparative study of predictive algorithms for time series forecasting ”, 3rd International Conference on Information Science and Technology, 2014.

[10] C. Ramirez, N. Naggapan, “Studying the impact of evolution in R libraries on software engineering research “ , 1st International Workshop on Software Analytics, 2015.

[11] D.J. Newman A. Asuncion. UCI machine learning repository, 2007.

[12] Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”,second edition Morgan Kaufmann publisher.