**Seoul Bike Sharing Demand Prediction**

**(Akshay Kumar Siani)**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

### Bike-Sharing, as a new green public transportation mode, has been developed in several western cities and most of the developing countries are on the path of following the western model of Bike Sharing Systems. There is a possibility that bike stations can be full or empty when a traveller comes to the station. Thus, to predict the use of such a system can be helpful for the users to plan their travels and for the entrepreneurs to set up the system properly. In this project, we analyse the data of Seoul bike sharing system deeply and make the prediction of bike count required at each hour for the stable supply of rental bikes.

.

***Keywords: Exploratory Data Analysis,Train-Test split, Machine learningmodel,***

**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 upon the following

* Various insights from the booking of bike for a different date, day, and year
* Variation of demand of bike on each day of the week. Does the demand fluctuate or remain same throughout the week?
* Is there any noticeable difference in booking bikes on different functioning days, seasons, holidays, and what could be the reason for it?

* Variation of the demand as per the Season and working day of the week.
* Variation of demand on the basis of weather parameters like humidity, temperature, rainfall etc

**2. Introduction**

Bike-sharing systems allow users to take one-way bike trips over short distances. 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. Data Handling and Machine learning models**

The application of machine learning models for bike-share networks can provide 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 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. 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 extract useful information from the date column. Our data set have the following features:

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 to Linear Regression Model and then moves forward to Lasso and Ridge regression for more effectiveness of the linear model. We also fit data on the decision tree and Random Forest and after that we move towards XGBoost and fit the data to this model and achieve performance more than 82% on the training data

**4. Methodology:**

Supervised learning algorithms like linear regression, decision trees, random forest, XGBoost has been used for predicting the demand of bike per hour. The step-by-step procedure of this project are as follows:

* **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 Decision Tree**
* **Developing and optimizing Random Forest**
* **Developing and optimizing XGBoost**

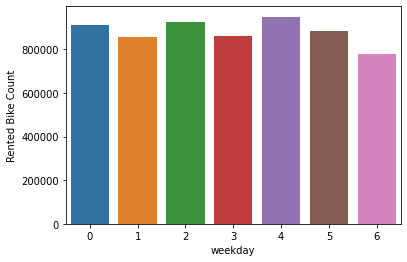
**4.1 Data pre-processing and transformation**

In pre-processing, we extract the information from the date string for finding the booking done for bike weekdays wise, month wise and year wise. Data given in our dataset is for 365 days. We look the effect of factors like function days, holidays, weekdays, various months of the year etc on the demand of bike. The structure of the dataset is shown in the figure.



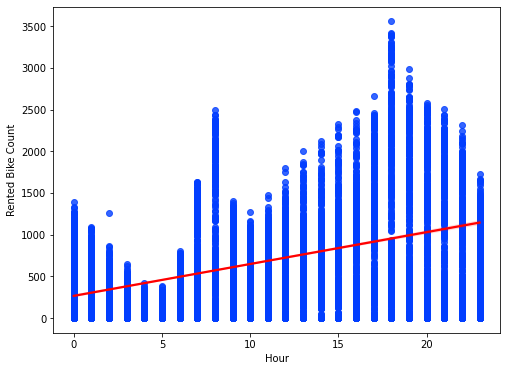
From the above graph, it is clear that in summer months demand for bike on rent is at peak throughout the year and in winters demand is very less.

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



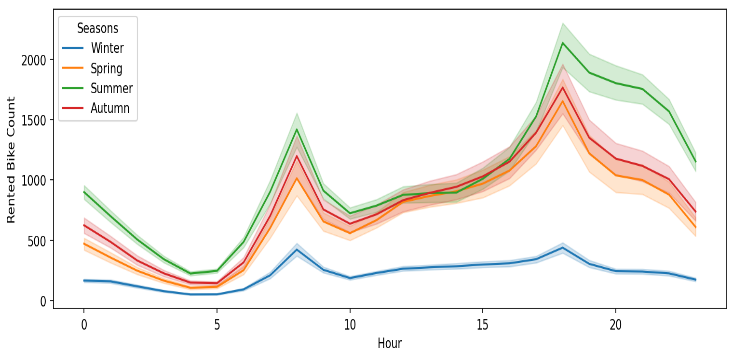
From the above graph, we can observe that demand of bike doesn’t vary much throughout the week. However, demand is comparatively low at weekends as compared to other days of the week.

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



From the graph, it can be seen that the people prefer to borrow bike during 8am and 6pm.

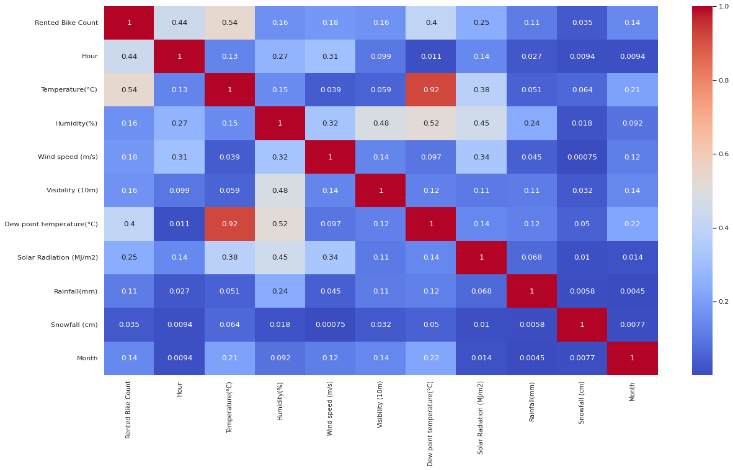
**Bike rented on various seasonal at  various hours of a day:**



.

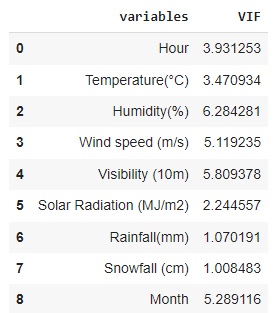
People prefer borrowing bikes more in a particular season. Rented bike count is highest in summer and least in winter. Peak demand of bike is at 8am and 6pm i.e. during office opening and closing times irrespective of the seasons.

**Correlation between various parameters:**

****

Weather parameters like temperture, dew point temperture and humidity are highly correlated with each other.

To deal with the multicollinearity, we find VIF values of independent variable



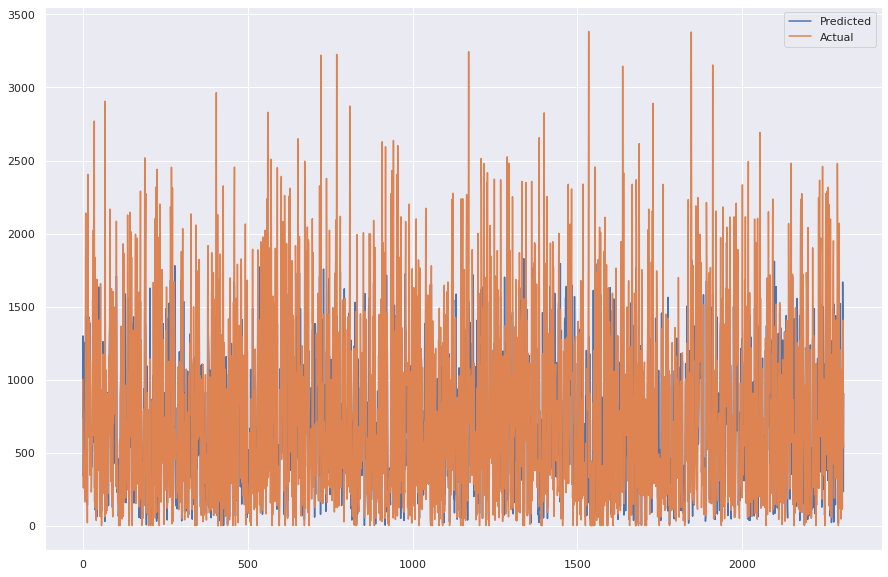
On removing dew point feature from the dataset, the VIF value of independent variables are under 6.

**Feature Engineering**

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 the new features help in improving the model accuracy. We did the data encoding for converting categorical variable into numerical form. We convert categorical variables ‘seasons’, ’holiday’ and ‘functioning day’ into numerical variable.

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

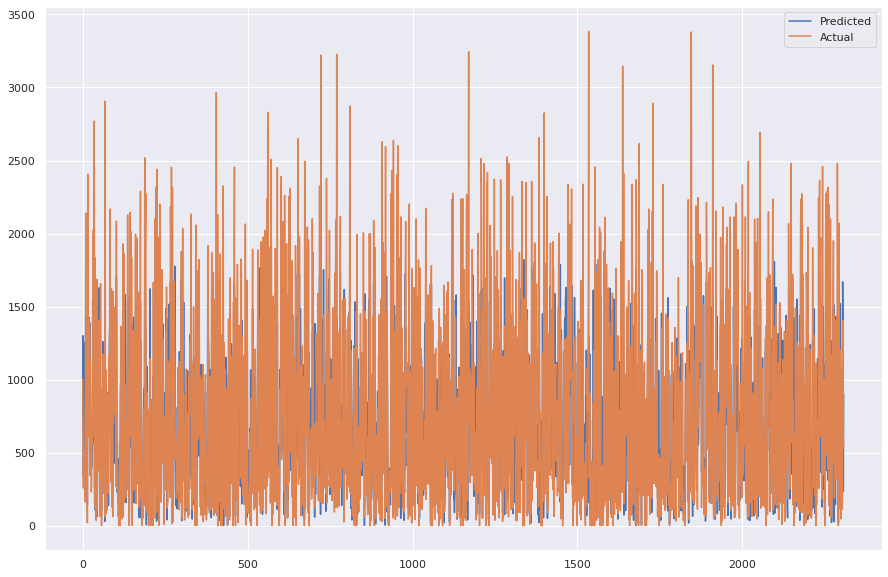
Linear regression model gives r square score of 0.582375. Linear regression model work with lots of assumptions.



From above graph, it is clear that linear regression model doesn’t predict the values effectively.

**4.3 Developing and optimizing Lasso Regression model**

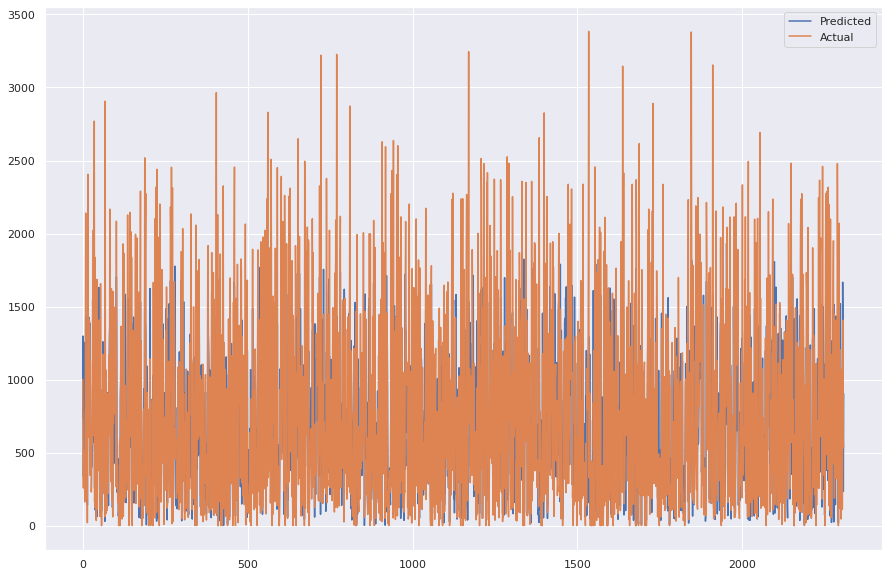
Lasso is variable panelised regression method. Lasso regression gives r square score of 0.581669.



From above graph, it is clear that lasso regression model doesn’t predict the values effectively.

**4.4 Developing and optimizing Ridge Regression model**

Ridge regression making the features coefficient optimization. Ridge regression model gives r square score of 0.581966.

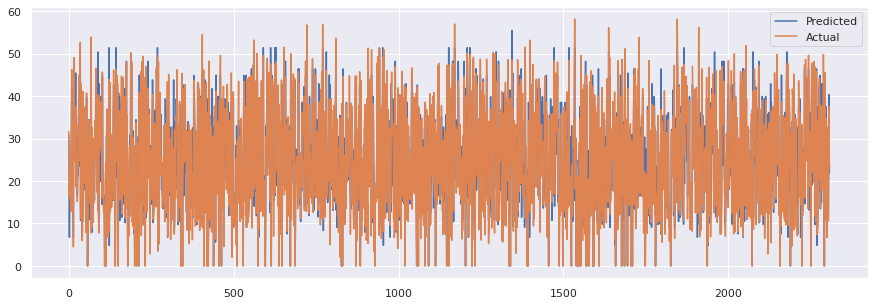


From above graph, it is clear that ridge regression model doesn’t predict the values effectively.

**4.5 Developing and optimizing Decision Tree**

Decision tree showing better metrics score then the linear, ridge and lasso regression. Decision tree regression model gives

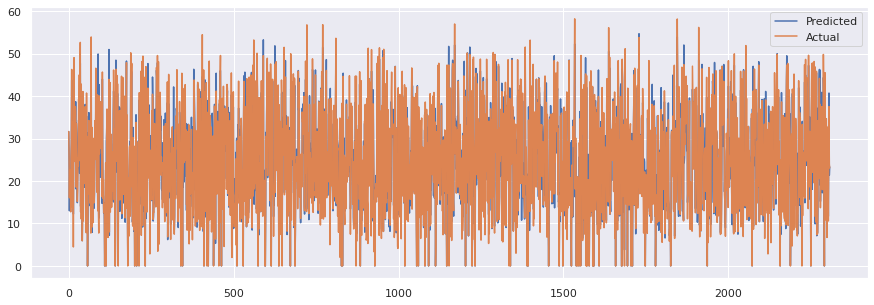
r square score of 0.823632.



From above graph, it is clear that decision tree regression model can predict the values effectively as compared to linear, ridge and lasso regression.

**4.6 Developing and optimizing Random Forest Tree**

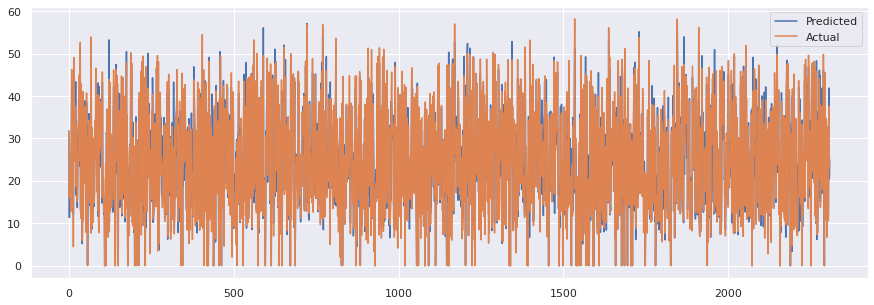
Random forest showing better metrics score then the linear, ridge, lasso and decision tree regression. Random forest regression model gives r square score of 0.866629.



From above graph, it is clear that Random forest regression model can predict the values effectively as compared to linear, ridge, lasso and decision tree regression.

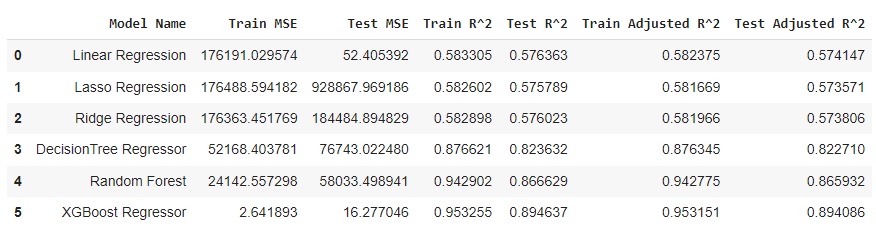
**4.7 Developing and optimizing Xtream Gradient Boosting**

XGBoost showing better metrics score then the linear, ridge, lasso, decision tree and random forest regression. Random forest regression model gives r square score of 0.894637.



From above graph, it is clear that XGBoost regression model can predict the values more effectively as compared to linear, ridge, lasso, decision tree and random forest regression.

**5. Comparison between various models on the basis of evaluation metrices**



**6. Conclusion**

This project proposed the use machine learning techniques to identify the demands in a bike-sharing system. Five supervised machine learning algorithms are applied on the bike share dataset for predicting the count of bikes that will be rented per hour.

We got some good results and accuracy with random forest and Xgboost. The performance of various models has been compared using evaluation matrices like mean square error (MSE), r square and adjusted r square. XGBoost with r square score of 0.894637 performs more effectively than all other regression models used in this project.

**7.REFERENCES**

[1] J. Yoon, Fabio Pinelli, “City ride: a predictive bike sharing journey advisor”2012 IEEE 13th International Conference on Mobile Data Management.[2] I.Frade, A.Ribbero, “Bicycle Sharing Systems Demand” unpublished.

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

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