**Bike Sharing Demand Prediction**

# Abstract:

In last decade Bike sharing program has become more popular. It has its own advantage such as it is beneficial for our health as well as for environment. It also reduces congestion and improving air quality. But there are challenges of providing adequate number of bikes at the peak hour and also to provide parking place at the time of drop off. To maintain the quality of the service well maintained bikes should be available with lesser waiting period. So, 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 bikes. This study will be focused on the demand prediction to run the bike sharing program successfully.

# Keyword

Machine learning, Data mining, Bike sharing demand prediction

# Problem Statement

Data provided by Seoul Bike Data consists of weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. Our target variable is bike count and we will be analysing the bike demand based on the above features.

For the programme to be success we need to :

Maximize: The availability of bikes to the customer.

Minimize: Minimise the time of waiting to get a bike on rent.

**The main goal of the project is to** Finding factors and cause those influence shortages of bike and time delay of availing bike on rent. Using the data provided, this paper aims to analyse the data to determine what variables are correlated with bike demand prediction. Hourly count of bike for rent will also be predicted.

# Data Description:

In the given set of data, we have 14 variables and 8760 rows. The dataset contains

* weather information
* Temperature
* Humidity
* Windspeed
* Visibility
* Dewpoint
* Solar radiation
* Snowfall
* Rainfall
* The number of bikes rented per hour
* Date information.

# Attribute Description

* **Date**: The date of the day, from 01/12/2017 to 30/11/2018, formatting is DD/MM/YYYY but we need to convert into “date-time format” for further analysis and model implementation.
* **Rented Bike Count**: Number of rented bikes per hour which is a dependent variable and prediction will be done for the same.
* **Hour:** The hour of the day, starting from 0-23 it's in a digital time format
* **Temperature (°C):**  Temperature of the weather in Celsius and it varies from -17°C to 39.4°C.
* **Humidity (%)**: Availability of Humidity in the air during the booking and ranges from 0 to 98%.
* **Wind speed (m/s)**: Speed of the wind while booking and ranges from 0 to 7.4m/s.
* **Visibility (10m)**: Visibility to the eyes during driving in “m” and ranges from 27m to 2000m.
* **Dew point temperature (°C)**: Temperature At the beginning of the day and it ranges from -30.6°C to 27.2°C.
* **Solar Radiation (MJ/m2)**: Solar radiation during ride booking which varies from 0 to 3.5 MJ/m^2.
* **Rainfall (mm)**: The amount of rainfall during bike booking which ranges from 0 to 35mm.
* **Snowfall (cm)**: Amount of snowing during the booking in cm and it ranges from 0 to 8.8 cm.
* **Seasons**: Here the whole year is categorized in 4 distinct seasons I.e. summer, autumn, spring and winter.
* **Holiday:** If the day falls in holiday or not.
* **Functioning Day**: If the day is a Functioning Day or not and it contains object data type yes and no.

# Introduction

The concept of bike sharing was first introduced in 1965 in Netherland but in last decade it has become more popular. The first bike sharing projects were initiated by various sources, such as local community organizations, charitable projects intended for the disadvantaged, as way to promote bicycles as a non-polluting form of transport, or as business enterprises to rent out bicycles.

Most large-scale urban bike sharing programmes have numerous bike check-out stations, and operate much like [public transit](https://en.wikipedia.org/wiki/Public_transport) systems, catering to tourists and visitors as well as local residents. It provides free or affordable access to [bicycles](https://en.wikipedia.org/wiki/Bicycle) for short-distance trips in an [urban area](https://en.wikipedia.org/wiki/Urban_area) as an alternative to private [vehicles](https://en.wikipedia.org/wiki/Vehicle), thereby reducing [congestion](https://en.wikipedia.org/wiki/Traffic_congestion), [noise](https://en.wikipedia.org/wiki/Noise) and [air pollution](https://en.wikipedia.org/wiki/Air_pollution). This system has its own pros and cons such as:

Pros:

* Improved air quality
* Convenience
* Better bike Laws
* Healthier people

Cons:

* First-time bikers
* No helmets
* Shortage of parking area

For the success of this programmes, we need to overlook the facts and features which can be worked and can predict the short-run bike-sharing demand, it will help operating agencies rebalance bike-sharing systems in a timely and efficient way.

# Steps Involved for EDA

EDA stands for exploratory data analysis. To understand any problem, we need to explore the data related to it to understand dependency and the corelation among all the features. Univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers. It involves data sanity check, missing values and null values, creating hypothesis and understanding the patterns among them.

* **Data Sanity checking** - Data sanity checks is performed to identify if there are any missing values & if any data transformation is required before analysing the data.
* **Null values Treatment -** If there is any possibility of null value, it might tend to disturb our accuracy, hence we need to drop them at the beginning of our project in order to get a better result.

Our dataset does not contain any null values, we can say our data set is cleaned data set

* **Changing the Datatype** - In the given data set day is written in DD-MM-YYYY pattern. Program is considering this data as “str” type but for further analysis we may need to extract some more information using this feature. So, we will convert this column in “date time format.
* **Deduplication- “**Duplication" just means that you have repeated data in your dataset. This could be due to things like data entry errors or data collection methods. By removing duplication in our data set, time can be saved by not sending identical communications multiple times to the same person. The number of duplicate values in the data set is 0.
* **Dropping Unwanted variables**
* **Exploratory Data Analysis**- Various EDA techniques used for this capstone project are – Univariate Analysis, Bivariate Analysis & Multivariate Analysis to know more about target variable’s dependency
* **Univariate Analysis**- Used Data summary, boxplot & count plot for Univariate analysis -

Data Summary**:**

****

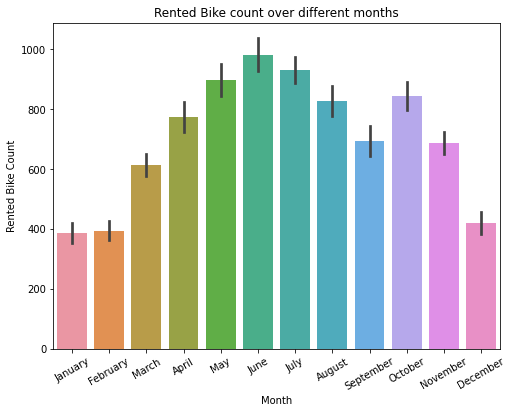
Fig 1

The table gives the basic information about the given data set.

* Temperature at that place varies from -17.8 to max 39.4**°C, with average temp 12.88.**
* Minimum humidity is 0 whereas maximum have been observed 98% with an average of 58.22%.
* Average wind speed is 1.72 m/s, with the range of 0 – 7.4 m/s.
* Visibility (10m) is varying from 27 to 2000 with considerable standard deviation of 608.299.
* Dew point temperature is varying from -30.60**°C** to 27.2**°C with an average of 4.07 °C.**
* Minimum solar radiation has been found 0, with std deviation in solar radiation 0.868 (MJ/m2), highest radiation has been recorded 3.5 (MJ/m2).
* Maximum rainfall is 35mm, average rain fall observed is 0.148 mm.
* Average snowfall is observed 0.075 cm with maximum 8.8 cm.

Further bike demand can be categorised as:

* Monthly Demand:

Fig 2

The above plot reflects that highest rented bike demand was in the month of July whereas January and February had same and least demand in that year.

* Season wise Demand:

When we try to analyse based on season, we got the below plot-

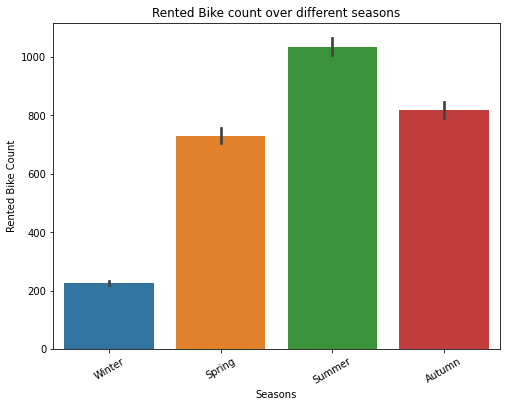


Fig 3

From this plot we can say most popular season for bike rental is Summer whereas winter season has least demand.

* Holiday: As per the name only, holiday (office /school/colleges) definitely reduces the bike demand and it is clear through the below plot. On holiday demand was lesser than the no holiday that is quiet obvious .

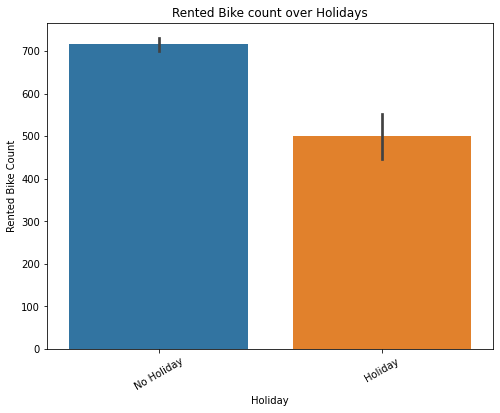
****

Fig 4

* Day wise Analysis:

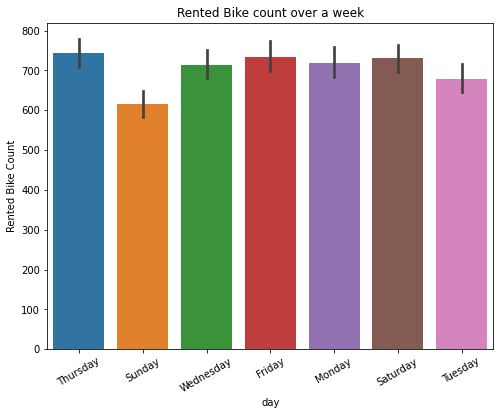


Fig 5

From the above plot we can conclude that Sunday being a holiday has the least demand whereas other days have slight variation in demand with small margin but thursday shows highest demand for rented bike.

* Hourly Demand:

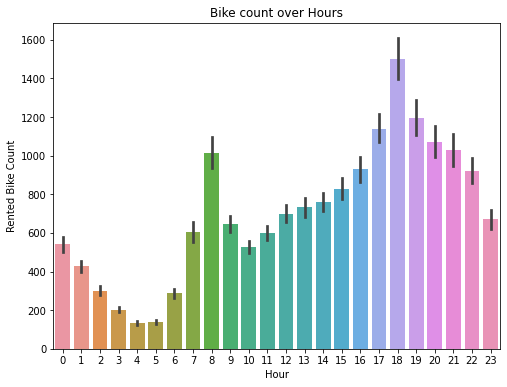


Fig 6

When we tried to analyse the demand over a 24 hour time period of the day , a beautiful plot we can see above. The time count has been started from midnight 12 as “0” and has been shown till night.

From above plot we can say morning hour 4 -5 has the least demand as less user may go at that hour where as evening 6 pm has the maximum demand. Similarly, we can see spike in demand around morning 8 after seeing a continuous decreasing pattern in demand from late night hour.

Skewness:

One more very important feature while data analysis is “Skewness”. In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined. We will use distplot to check skewness.

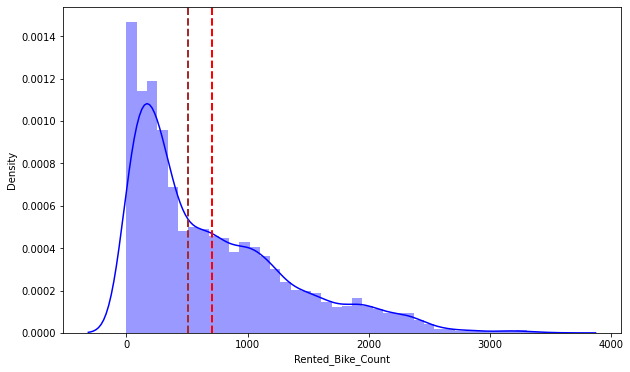


Fig 7

When we see distribution plot for rented bike count, we found that it was right skewed from the box plot it is clearly visible that it has higher number of outliers.

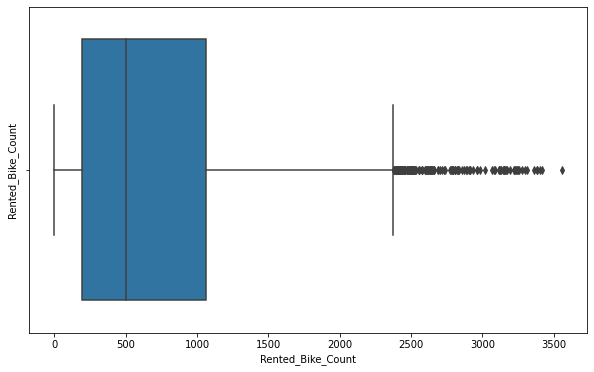


Fig 8

The outliers will in turn have an effect on different accuracy measures of a model and can further lead to errors in estimations as well. Outlier impact prediction and prediction results including model estimates. Hence, we have tried to improve the skewness by applying square root function which also resulted in outlier correction.

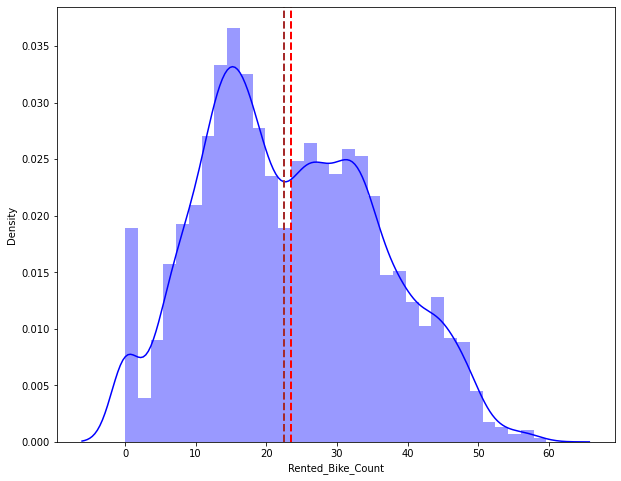


Fig 9

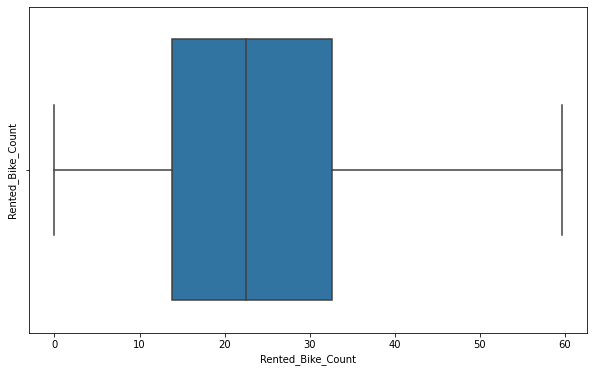


Fig 10

We check correlation between variables using Correlation heatmap, it is graphical representation of correlation matrix representing correlation between different variables.

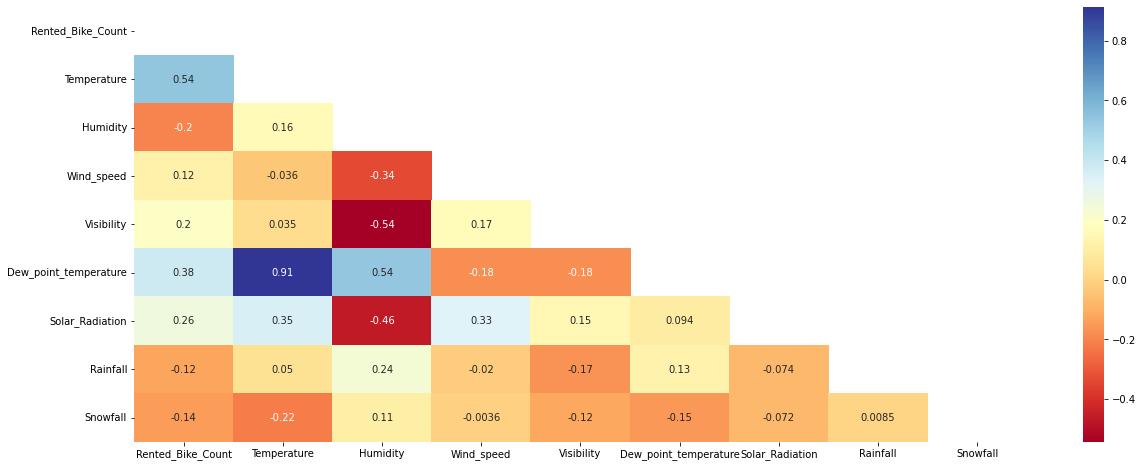


Fig 11

We can observe on the heatmap and can summarize about the co-relation as follows:

* The temperature
* The dew point temperature
* The solar radiation

Whereas most negatively correlated variables are:

* Humidity
* Rainfall
* snowfall

From the above correlation heatmap, we see that there is a positive correlation between columns 'Temperature' and 'Dew point temperature' i.e. 0.91 so even if we drop this column then it will not affect the outcome of our analysis. And they have the same variations. So, we can drop the column 'Dew point temperature(°C)'.

**Create the dummy variables:**

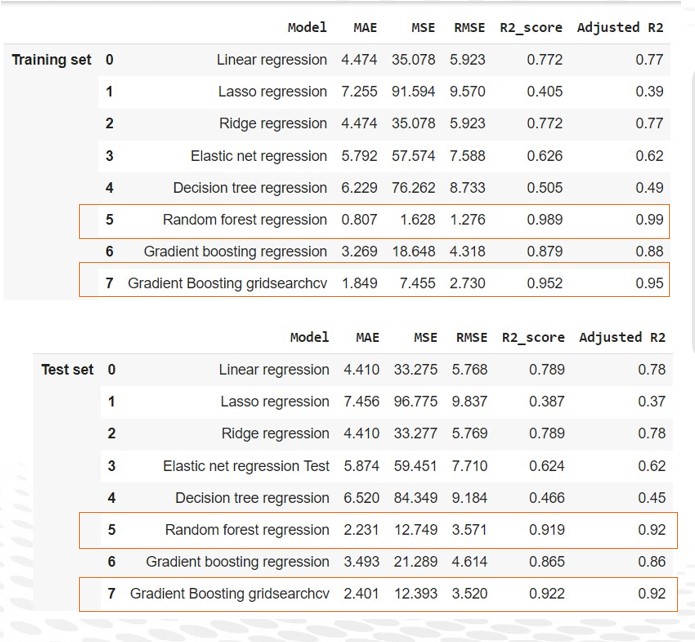
A dataset may contain various type of values, sometimes it consists of categorical values. So, in-order to use those categorical value for programming efficiently we create dummy variables.

**One Hot Encoding:**

A one hot encoding allows the representation of categorical data to be more expressive. Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers and required for both input and output variables that are categorical**.**

**Model Training & Testing:**

Before, fitting any model it is a thumb rule to split the dataset into a training and test set. This means some proportions of the data will go into training the model and some portion will be used to evaluate how our model is performing on any unseen data. The proportions may vary from 60:40, 70:30, 75:25 depending on the person but mostly used is 80:20 for training and testing respectively. In this step we will split our data into training and testing set using scikit learn library.



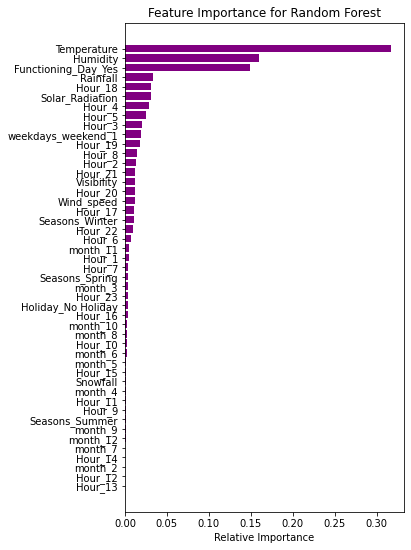
****

Fig 12

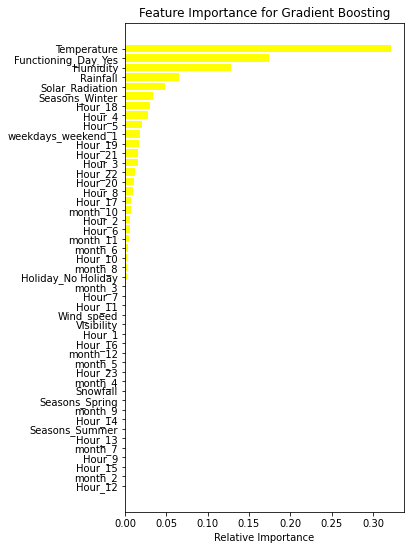
****

Fig 13

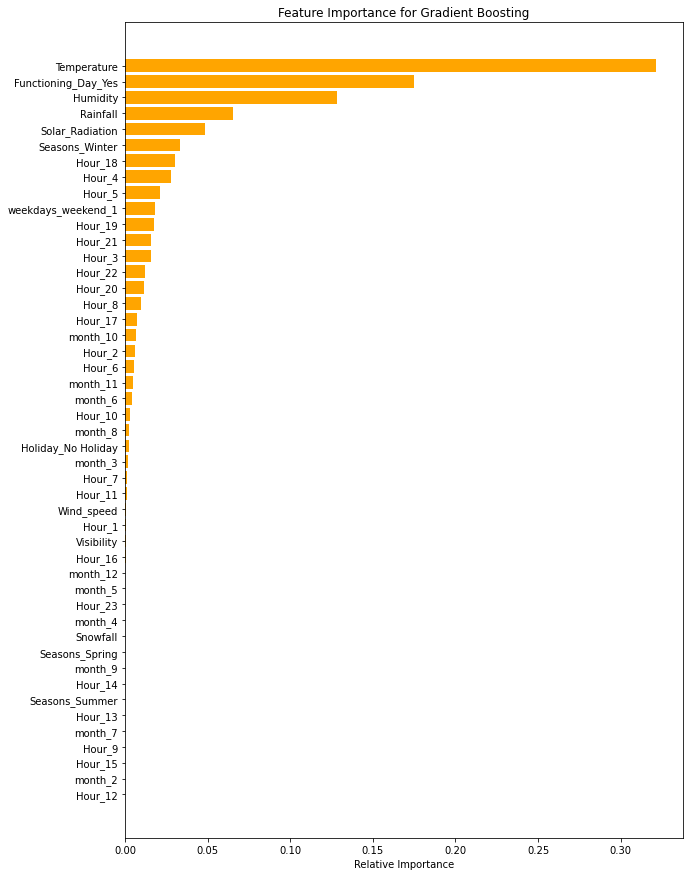
****

Fig 14

We have plotted graph for random forest, gradient boosting and gradient boosting hyperparameter tuned. From the above plots we can conclude that 4 out of Top 6 features in each of the best performing model is weather related (Temperature, Humidity, Rainfall & Solar radiation). Further to enhance the model, weather forecast data can also be considered as one of the inputs. Whereas all the above three models clearly indicates that Temperature is the most affecting feature in case of demand prediction. And similar observation we have seen in our EDA as well. It proves the accuracy of the model.

**Conclusion:**

During our analysis, we did EDA and found the factors affecting the bike demand. We have also plotted heatmap and studied the correlation between the variables with our target variable that is “rented bike count”.

To predict bike demand we have splitted test set and train set of the given data set and implemented various model for comparative study to find the best fit model. Here are some insights:

1. Gradient Boosting with hyper tuning or Random Forest model suits to be best model for predicting the Bike demand based on given dataset. Recommended to review the model every 3-4 months for accuracy & with new data
2. 4 out of Top 6 features in each of the best performing model is weather related (Temperature, Humidity, Rainfall & Solar radiation). Thus, weather forecast plays very important role in predicting the bike demand. Further to enhance the model, weather forecast data can also be considered as one of the inputs
3. Peak demand is observed at 18:00 hrs especially during functioning days and non-holidays. This indicates mostly demand is for users returning from office/school etc. Hence, availability of bikes can be planned near the office zones. Such zones can be earmarked and more bikes should be made available during these peak hours in these areas.
4. High demand is in summer while demand in winter is very less. Annual maintenance of bikes can be planned during winter.
5. Low demand during 02:00 hrs to 05:00 hrs, daily maintenance can be planned during these hours
6. Rainfall, Snowfall & Humidity are only 3 variables which negatively impacts the bike demand.
7. Further insights can be developed by analysing location specific data so that availability of bikes can be maximized & waiting times can be minimized.
8. Incentivize users through discount offers, if they drop the bike near peak zones before high demand hours (18:00 hrs), thus enabling user driven bike availability.
9. Additional discounts can be offered to users, if they pick the bike before and after peak hours. This will ensure spreading the demand to wider horizon (say instead of 18:00 hrs, demand can be spread in range of 16:00 hrs to 20:00 hrs)

However, this model might not perform well forever as the target variable is dependent upon temperature, wind speed, rainfall, solar radiation and these might not be consistent forever. Machine learning is an exponentially evolving field and we need to keep checking it time to time.