**Capstone Project**

**Bike Share Demand Prediction**

**Rohit Madame**

**Abstract:**

A bike sharing is a system where users can rent a bike to their destination. For this, they may be charge an amount. This is a computer-based process that allows users to pick up the bike from one place and park in their other parking stand. The bike can be Rent on an hourly or weekly basis.

**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 bikes.

This dataset contains the data of rented bike count and include information such as

* Date : year-month-day
* Rented Bike count - Count of bikes rented at each hour
* Hour - Hour of he 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)

**Introduction:**

What is the busiest time to rent a bike? Do customers prefer to rent a bike on holiday or on no holiday? What are the chances of getting a bike for rent on weekends or weekdays? Which month has highest rented bike count? Which is the season do customer like to rent the bike?

**Steps Involved:**

* Data Preparation
* Data Cleaning and Feature Engineering
* EDA
* Applying algorithms
* Comparing Different Models
* Conclusion

**Data Preparation:**

Importing all the necessary libraries and Seoul Bike Data .This dataset is from 2017 to 2018 and dataset contains 8760 rows and 14 columns.

**Data Cleaning and Feature Engineering:**

This dataset doesn’t contain null values or duplicate values. But the column name has symbol in it. So we have renamed the column name. I have changed Date column to datetime64[ns]. I have created Month and Day column with the help of date column. Also I used Day column to create Weekend data column and then dropped Day column.

**EDA:**

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it. First I have plotted for the target value i.e Rented Bike Count and have outliers in it. So we have to remove it as it may impact our performance.

For Weekdays the value of the Rented Bike Count has highest than for the weekends. The number of rented bikes count is highest in summer. Customers like to rent more bikes in summer and the rent count is less in winter as compared to other seasons. The peak time for bike rentals in the morning is from 7 am to 9 am whereas in the evening the peak time is from 5 pm to 9 pm. So we can think that the count is higher as customers go to school, office or company in the morning. When it’s no holiday the bike count is higher as compared to the holiday. On no functioning day, the bike count is zero so we can say that it is an imbalanced dataset.

**VIF (Variation inflation factor):**

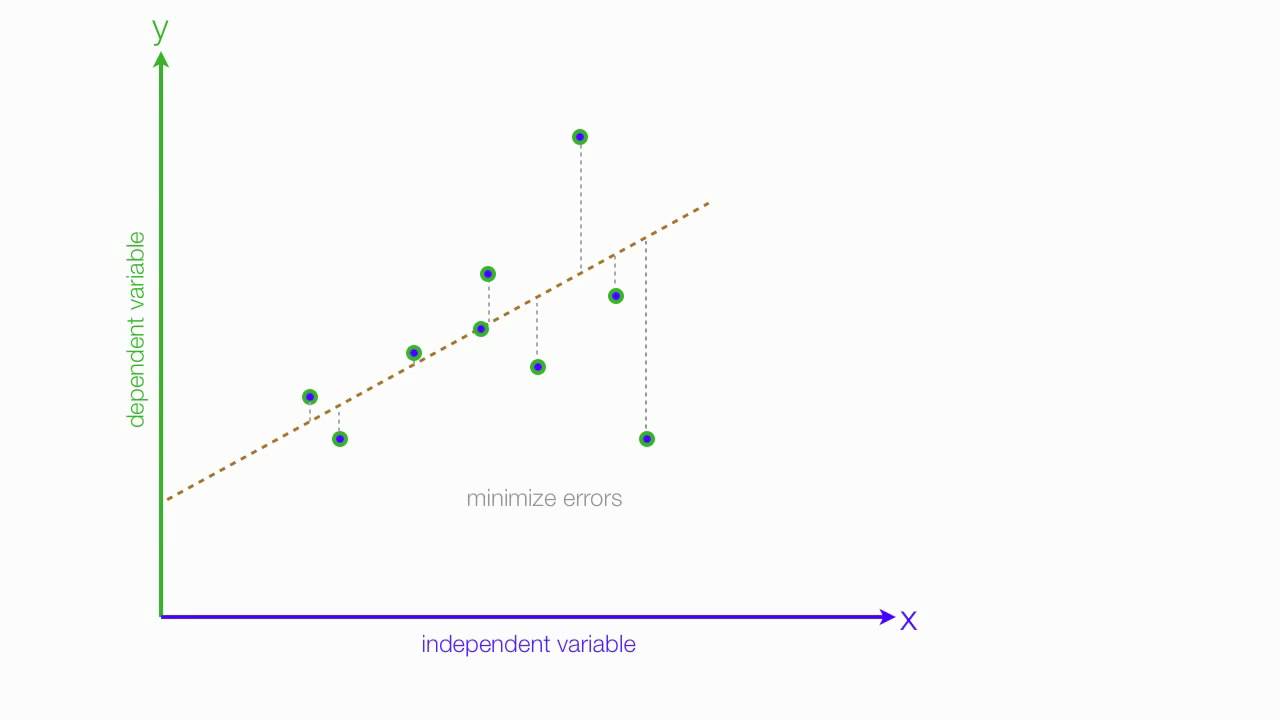
For multiple linear regression we can measure the amount of multicollinearity by variance inflation factor. We have applied VIF to our data and found that some columns have high VIF. So we have removed Humidity, Visibility and minimize the VIF.

**ALGORITHMS:**

**Linear Regression:**

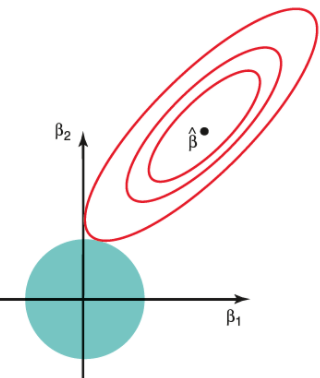
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

The equation for linear regression is y = B0 + B1\*x



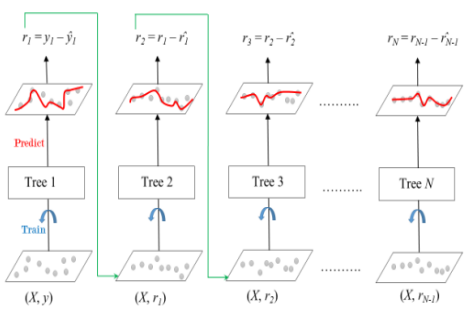
**Ridge Regression:**

Ridge [regression](https://www.mygreatlearning.com/blog/what-is-regression/) is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.



**Gradient Boosting :**

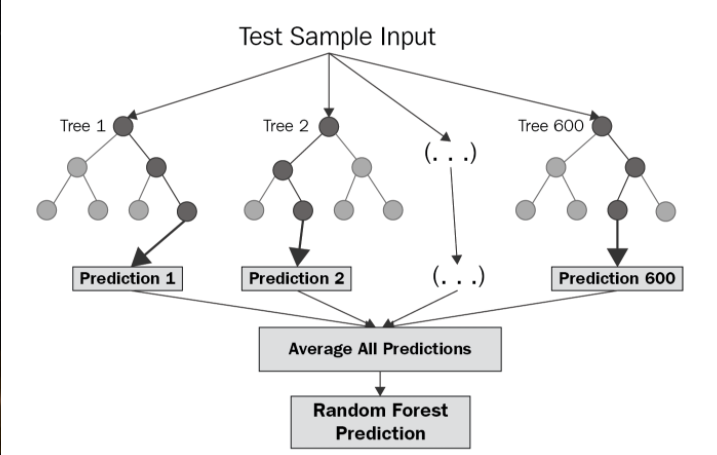
**Gradient Boosting** is a popular boosting algorithm. In gradient boosting, **each predictor corrects its predecessor’s error.** There is a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees).



The ensemble consists of N trees. Tree1 is trained using the feature matrix X and the labels y. The predictions labelled y1(hat) are used to determine the training set residual errors r1. Tree2 is then trained using the feature matrix X and the residual errors r1 of Tree1 as labels. The predicted results r1(hat) are then used to determine the residual r2. The process is repeated until all the N trees forming the ensemble are trained. Each tree predicts a label and final prediction is given by the formula**, y(pred) = y1 + (eta \* r1) + (eta \* r2) + ....... + (eta \* rN)**

**Random Forest Regressor:**

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**.



The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.   
Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

**Conclusion:**

* Demand for bike rent count is high from May to October and highest in the month of June.
* On weekdays demand for bike is high as compare to weekends.
* Summer is the best season to rent most bike.
* Peak hours in morning is from 7am to 9am and in evening it is from 5pm to 9pm.
* No holiday is preferred for highest bike count.

**References:**

* GeeksforGeek
* Almabetter
* stackoverflow