**Summary**

We were asked to forecast bike rental demand of Bike sharing based on historical usage patterns in relation to weather, time, and other data. The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, and date information.

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 the bike count required at each hour for the stable supply of rental bikes. Using these Bike Sharing systems, people rent a bike from one location and return it to a different or same place on a need basis. People can rent a bike through membership or on a demand basis.

**Steps to solve this problem statement:**

* Understanding the dataset, doing some basic inspection on the raw data to check the number of columns, understanding the distribution of data, and checking statistics of the data in each variable. Checking for missing values, Visualizing the distributions and boxplots of each variable to handle the outliers, and Cleaning the data.
* Feature engineering, creating some new features, dropping unnecessary features, and encoding the data into numeric form. Tried making the dependent variable normally distributed by some transformations.
* Finally scaling the data and experimenting with different algorithms. First, we started with simple models like Linear Regressor and Decision Trees then to enhance the accuracy we tried some complex algorithms like Tree ensemble.
* Since there was not much linear relation between the independent and dependent variables, the linear regressor model did not perform well, so we moved to Tree based algorithms, and performance was drastically improved. We kept on improving the model performance by using some boosting and ensemble algorithms and tuning the hyperparameters. The best performance was given by the XG Boost model.
* We observed different evaluation metrics with the best set of hyperparameters for the experimented models to overcome underfitting or overfitting and also had a rough idea of feature importance for each model.

## Result:

* RMSE Comparisons:
  + Linear Regressor RMSE: 417.90
  + Decision Tree Regressor RMSE: 276.89
  + Random Forest Regressor RMSE: 202.66
  + Gradient Boosting Regressor RMSE: 197.85
  + XG Boost Regressor RMSE: 185.15
* Temperature is the most influencing feature and humidity is in second place for Linear Regressor.
* Hour and seasons are the most important feature for the Decision Tree, Random Forest, and Gradient Boosting Regressor.
* Functioning day is the most important feature and Winter is the second most for Extreme Boost Regressor.
* R2 scores of the linear, lasso, and ridge regression are similar.
* R2 score of decision tree is 0.81 while r2 sore of random forest is 0.90.
* XG Boost is acting differently from all the regressors as it is considering whether it is winter or not. And is it a working day or not? Though winter is also a function of temperature only it seems this trick of XG Boost is giving better results.
* XG Boost Regressor has the Least Root Mean Squared Error of 185 and an r2 score of 0.91. So, it can be considered the best model for the given problem.