**Bike Sharing Demand Prediction**

**Sharath Diwakar**

**Data Science Enthusiast**

**AlmaBetter, Bangalore**

# Abstract:

Rental Bike Sharing is the process by which bicycles are procured on several basis- hourly, weekly, membership-wise, etc. This phenomenon has seen its stock rise to considerable levels due to a global effort towards reducing the carbon footprint, leading to climate change, unprecedented natural disasters, ozone layer depletion, and other environmental anomalies.

In our project, we chose to analyze a dataset pertaining to Rental Bike Demand from South Korean city of Seoul, comprising of climatic variables like Temperature, Humidity, Rainfall, Snowfall, Dew Point Temperature, and others. For the available raw data, firstly, a through pre-processing was done after which a Here, hourly rental bike count is the regress and. To an extent, our linear model was able to explain the factors orchestrating the hourly demand of rental bikes.

Keywords: - ***Data Mining, Linear Regression, Correlation Analysis, Bike Sharing Demand Prediction, Carbon Footprint.***

# 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.

# Data Summary:

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, and date information.

Attribute Information:

* Date: year-month-day
* Rented Bike count - Count of bikes rented at each hour
* Hour - Hour of the day
* Temperature-Temperature in Celsius
* Humidity - %
* Windspeed - m/s
* Visibility - 10m
* Dew point temperature - Celsius
* Solar radiation - MJ/m2
* Rainfall - water in mm
* Snowfall - Thickness cm
* Seasons - Winter, Spring, Summer, Autumn
* Holiday - Holiday/No holiday
* Functional Day - NoFunc (Non-Functional Hours), Fun (Functional hours)

# Steps involved:

## Exploratory Data Analysis

After loading the dataset, we compared our target variable that is the Rented Bike count Type with other independent variables. This process helped us figure out various aspects and relationships among the dependent and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the dependent variable.

## Null values Treatment

Our dataset didn’t have any null values to be treated.

## Encoding of categorical columns

We used One Hot Encoding (converting to dummy variables) 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.

## Standardization of features

Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

## Fitting different models

For modeling, we tried various classification algorithms like:

* Linear regression with regularization (Lasso & Ridge)
* Decision Tree
* Random Forest regression
* Gradient Boosting

## Tuning the hyperparameters for better accuracy

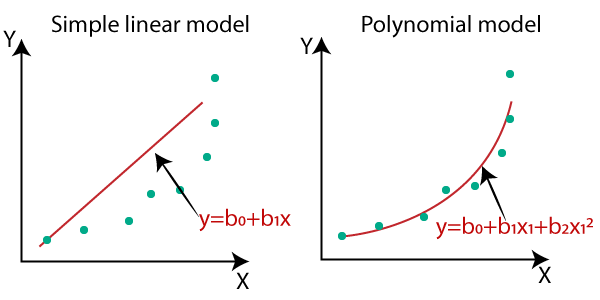
Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in the case of tree-based models like Random Forest Classifier and Gradient Boosting.

# Algorithms:

## Linear Regression:

Linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

The optimization algorithm used is Gradient Descent. We also performed different types of regularization techniques to prevent overfitting in the model.

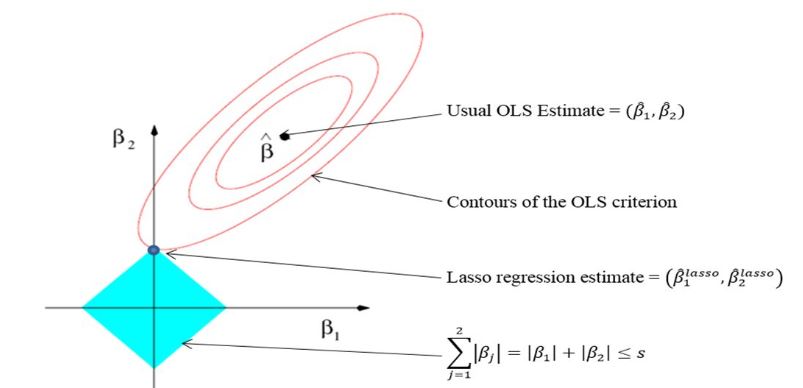


1. **Lasso Regression**

**Lasso regression** is a type of **linear regression** that

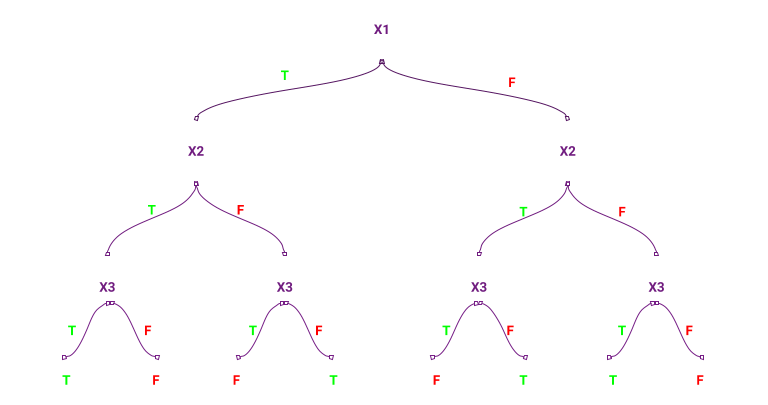
uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

The acronym “LASSO” stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator.



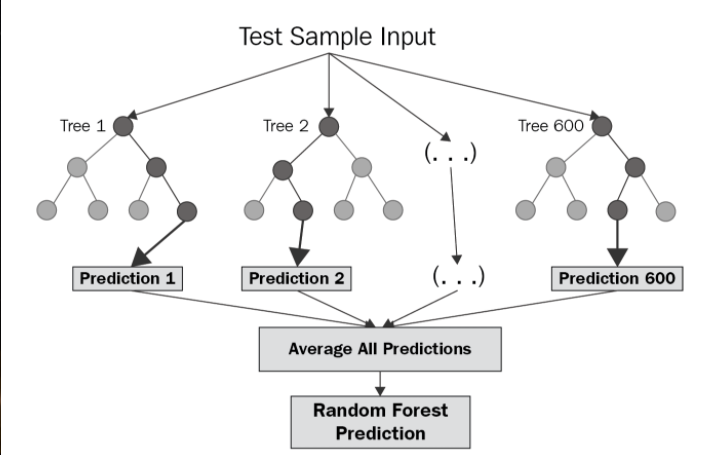
1. **Decision Tree**

A decision tree is a graphical representation of all the possible solutions to a decision based on certain conditions. Tree models where the target variable can take a finite set of values are called classification trees and target variable can take continuous values (numbers) are called regression trees.



## Random Forest Regression:

Random Forest is a bagging type of Decision Tree Algorithm that creates several decision trees from a randomly selected subset of the training set and n features, collects the values from these subsets, and then averages the final prediction out of all n number of decision trees



# Model Performance:

The model can be evaluated by various metrics such as:

## Mean square error

The MSE of an [estimator](https://en.wikipedia.org/wiki/Estimator) measures the [average](https://en.wikipedia.org/wiki/Expected_value) of the squares of the [errors](https://en.wikipedia.org/wiki/Error_(statistics))—that is, the average squared difference between the estimated values and the actual value.

## 

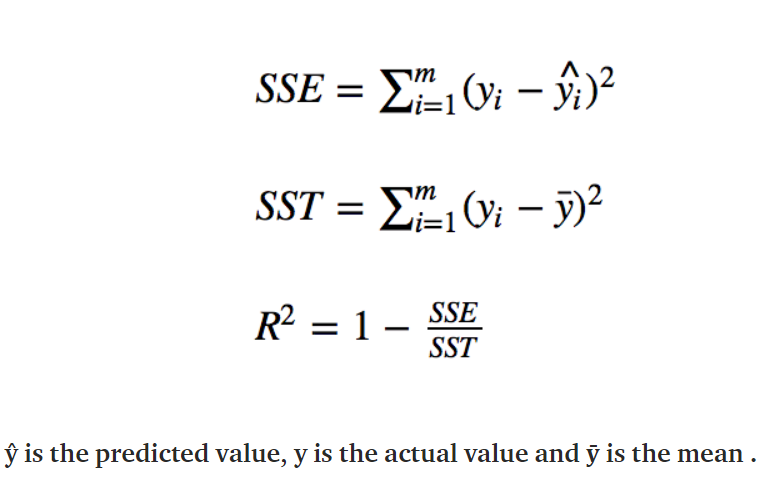
## Root mean square error

RMSE is just the root of MSE. It is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient. One can compare the RMSE to observed variation in measurements of a typical point

## 

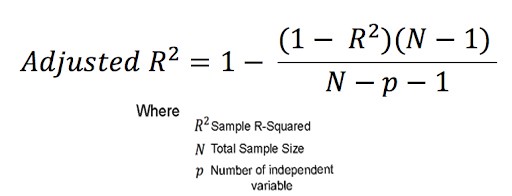
## R square:

R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model. It has one limitation that its value increases as the number of Parameters increase even if that parameter does not improve model



## Adjusted R Square:

Adjusted R-squared is a modified version of R-squared that overcomes the problem of r2 and has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

****

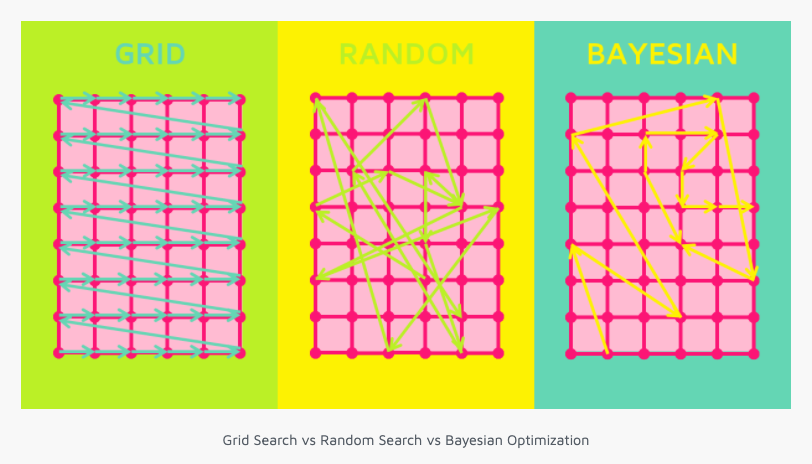
**Hyperparameter tuning**

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects the performance, stability, and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV for hyperparameter tuning. This also results in cross-validation and in our case, we divided the dataset into different folds.

**Grid Search CV-Grid:**

Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.



**Conclusion:**

* The temperature, hours, and solar radiation features were found to be more relevant for the bike count required at each hour for the stable supply of rental bikes.
* As we have analyzed the various features, we have seen that people prefer to take bikes on rent when temperature is near about 25 degrees Celsius.
* Other factors such as rainfall and snowfall also have an impact on the requirement of bikes for rent. Because in heavy rainfall and snowfall bike riding sometime becomes dangerous.
* As we have analyzed that the rental bike demands are high in the evening and in the morning. So, bikes should be available at that time to fulfill the bike demands.
* The Bike demand increases with an increase in visibility and decreases with an increase with humidity.
* We tried adding possible columns to make the model a bit more complex but for Linear Regression model it is still too general.
* We have to make our model more complex for better discretion or move to tree based and ensembling algorithm for better results.
* Random forest gives predictions better than a decision tree model. Predictions made by Random Forest is better than all the models used. The value of the Adjusted R-squared for the Random Forest is 0.796, which is good.

**References**

1. Stack Overflow
2. Medium
3. GeeksforGeeks