# Bike Sharing Demand Prediction

**Nikunj Sonule**
**Data Science Trainee**
**AlmaBetter, Bangalore**

## Abstract:

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. We were provided with 'Seoul Bike Demand Prediction' dataset to perform machine learning task, were to get insights from the independent and dependent variable of a dataset.

Our experiment can help understand what could be the reason for the regression of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the correct regression problems.

***Keywords:machine learning, regression, algorithms***

## 1. Problem Statement

Data provided by a Seoul Bike Sharing organizations. The people can have feasible use of bikes and organizations to know how to scale. By analyzing the problems with count of bikes rented at each hour the demand is getting increases with some specific times, seasons and whether.

- Dew point temperature – Celsius
- Solar radiation – MJ/m2
- Rainfall – mm
- Snowfall – cm

The people can find various way to reach the destination but bike sharing is in demand where it's prediction can play a major role to prevent from certain problems and losses. We've a data from 1st December 2017 till 30th November 2018. In this particular time duration we've all record and data with different seasons, whether, temperature, rented bike count, hour, humidity, visibility, solar radiation, rainfall, snowfall, holiday and functioning day.

During this period they captured all the data for data analysis and prediction to done. The main objective is to built a predictive model, which could help them in predicting the 'Rented Bike Count at each hour'. This would help in turn help them in providing bikes at each hour in every possible condition, quickly and efficiently.

- 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

- Seasons – Winter, Spring, Summer, Autumn
- Holiday – Holiday/ No holiday
- Functional Day – NoFunc(Non Functional Hours), Fun(Functional hours)

## 2. Introduction

The Seoul city is been providing with a stable supply of rental bikes. The major concern is about providing rented bike at each hour. Looking towards various factor/feature affecting the rented bike count at each hour (i.e dependent variable). It is important to make rental bike available and accessible to the public at right time without waiting for it.

Generally the algorithms can predict best fit model for the data according to it accuracy score. This rented bike count shows how it varies from every machine learning model. So, generating different accuracy score from various different model to know which is good.

Our goal is to build a predictive model, which could help organizations in predicting the rented bike count at each hour proactively.

## 3. Observations of Rented Bike Count:

- **Reason how it is affecting**

The hour feature is affecting a lot, it varies in an hour for bike. The temperature is too really varying till 40 degree Celsius. There are many factor which gets difficult for people to provide quick and effectively. At particular high wind speed it's harsh to prefer bike at that particular time. So, whether is major interference to both pros and cons. The feature visibility shows great insight for more demand. But certain problem is also concern with dew point temperature and solar radiation.

The increase in rainfall and snowfall has made it impossible. The impact is totally in negative, most of the Seoul folk won't get use for rented bike too. Throughout the month it is necessity for the people. Week_days are more number of chance to get attracted towards the bike sharing.

## 4. Seasons wise rented bike:

- Whenever Seasons gets change with time, it too affect for bike sharing demand. The insights based on the categorical features has revealed.
- The more bike count is for 'No Holiday'.
- There is almost no requirement for no functioning day.

## 5. Steps Involved:

- **Exploratory Data Anlaysis**

  After loading the dataset I performed this method by comparing our target variable (i.e Rented Bike Count) with other independent variables. This process helps us figuring 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 doesn't contains any null values, (if it would be which might tend to disturb accuracy and drop them at the beginning of data preprocessing).

- **Extracting features from column**

  The date column was an important feature to gets smaller insights from its. Extracting month, dayOfweek and weekdays_weekend to create a new feature for dataset.

- **Univariate and Bivariate Analysis**

  The numeric feature has helped with histogram analysis and got insight from heatmap for multi-collinearity. Each feature like hour, temperature, humidity, wind speed, visibility, solar radiation etc. shows count value that how it varies.

  The dependent variable has shown relation with every numeric feature of a dataset. The regression plots has help to understand each variable.

- **Fitting different models**

  For modeling I tried various regression algorithms like

  1. **Linear Regression(Also implemented Lasso, Ridge and ElasticNet but gives nearly similar accuracy)**

2. **Decision Trees**
3. **Random Forest**

- **Feature Importance**

  Looking towards feature importance I noticed ‘temperature’ and ‘Hour’ are the most important variable for Decision tree and Random Forest model. There is slightly competition between ‘Solar Radiation’ and ‘Functioning Day_Yes’ with more necessary feature for model. The ‘temperature’ is most feature importance among all features.

- **Finding best parameters by GridSearchcv**

  While doing regularization it is important to shrink the coefficient values to get better accuracy. So,implementing GridSearchcv in Lasso, Ridge and ElasticNet Regression gives slightly change in adjusted r2 accuracy but their was no good performance. It was quite similar to Linear Regression model accuracy.

- **Model Evaluation Metrics**

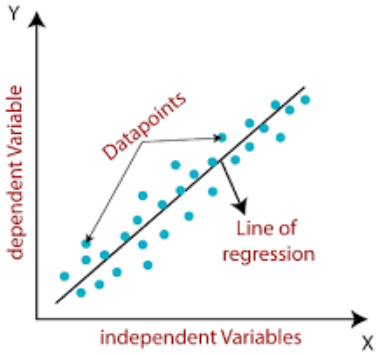
  By model evaluation metrics, the calculation value of Mean Square Error(MSE), Root Mean Square Error(RMSE), R2 and Adjusted R2 has shown clear vision of performance of each model. Among all others Random Forest has shown a better result in accuracy in Adjusted R2 with 84%.

# 6. Algorithms:

## 1. Linear Regression:

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line. Linear regression is used to estimate the relationship between two quantitative variables. Simple linear regression when :

- How strong the relationship is between two variables.
- The value of the dependent variable at a certain value of the independent variable.




**Straight Line Formula**: $y = mx + c$
***- Where {m} is the slope and {c} is the intercept***

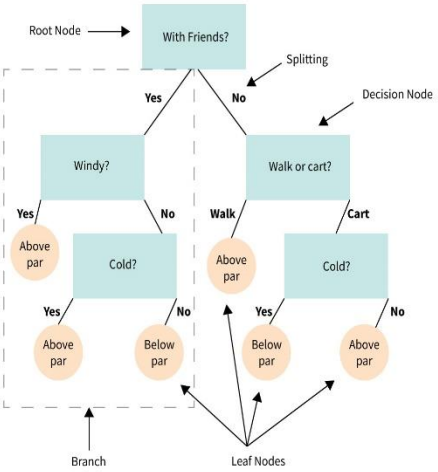
- $\hat{y} = \hat{m}x + \hat{c}$

  **or**

- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

## 2. Decision Tree

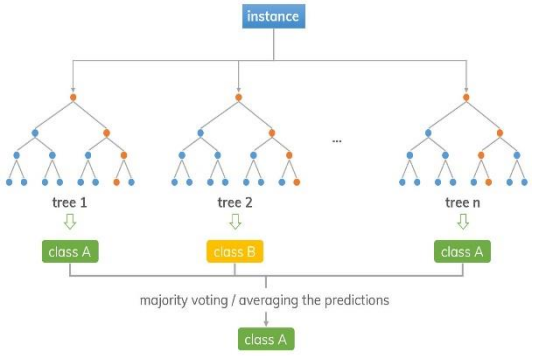
A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered.




- **Root node:** The base of the decision tree.
- **Splitting:** The process of dividing a node into multiple sub-nodes.
- **Decision node:** When a sub-node is further split into additional sub-nodes.
- **Leaf node:** When a sub-node does not further split into additional sub-nodes; represents possible outcomes.
- **Pruning:** The process of removing sub-nodes of a decision tree.
- **Branch:** A subsection of the decision tree consisting of multiple nodes.

## 3. Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It




builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Important Hyperparameter:

- n_estimators– number of trees the algorithm builds before averaging the predictions.
- max_features– maximum number of features random forest considers splitting a node.
- mini_sample_leaf– determines the minimum number of leaves required to split an internal node.

## 4. Model Performance:

Model can be evaluated by various metrics such as :

- **Mean square error**

The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points. For every data point, you take the distance vertically from the point to the corresponding y value on the curve fit (the error), and square the value.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$\text{MSE}$ = mean squared error
$n$ = number of data points
$Y_i$ = observed values
$\hat{Y}_i$ = predicted values

- **Root Mean Square Error**

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. Formally it is define as:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

- **R2**

R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

$$\text{R2 Squared} = 1 - \frac{SSr}{SSm}$$

SSr = Squared sum error of regression line

SSm = Squared sum error of mean line

- **Adjusted R2**

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where

$R^2$ Sample R-Squared

$N$ Total Sample Size

$p$ Number of independent variable

# 5. Conclusion:

That's it! I reached the end of exercise.

Starting with loading the data so far I have done EDA, null values treatment, extracting features, univariate and bivariate analysis, model building, feature importance and model evaluation metrices.

Even after Gridsearchcv in linear regression we don't have good accuracy score in linear regression.

So accuracy of the best model is 84% for this dataset. This performance could be due to various reasons : no proper pattern of data, different model give different accuracy score.

**References-**


1. MastersIndatascience
2. Towardsdatascience
3. Analytics Vidhya