**Seoul Bike Sharing Demand Prediction**

**Abhay Deep Singh**

Data Science Trainees

**AlmaBetter, Bangalore**

**Abstract:**

Currently, the bike-sharing scheme is well-received throughout the world. It is a shared bike service to individuals, which is free of charge and for a short term basis at a minimal rate. Most bike-sharing systems permit people to borrow and return a bike from a bike station to another station that belongs to the same network.

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

**1. 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.

## 2. Introduction

Currently, the bike-sharing scheme is well-received throughout the world. It is a shared bike service to individuals, which is free of charge and for a short term basis at a minimal rate. Most bike-sharing systems permit people to borrow and return a bike from a bike station to another station that belongs to the same network.

Bike sharing involves the provision of a pool of bike across a network of strategically positioned ‘bike sharing stations’, typically distributed in an urban area, which can be accessed by different types of users (i.e., registered members or occasional/casual users) for short-term rentals allowing point-to-point journeys.

Bike-sharing gains a vast range of attention in recent years as part of initiatives to boost the use of cycle, improve the first mile/last mile link to other modes of transportation, and to minimize the negative effect of transport activities on the environment. Bike-sharing has significant impacts on establishing a larger cycling community, increasing the use of transportation, minimizing greenhouse gas emissions, enhancing public health and also traffic troubles.

So the constant raise of users necessitates the prediction of the number of rental bikes that were needed to make the bike sharing system to consistently work. Therefore, this notebook aims to use machine learning to predict the required number of rental bikes required at each hour.

## 3. Data Description

This work comprehends the relation between rental bike used in each hour and the different predictors such as weather information and time information. The dataset contains count of public bikes rented at each hour in Seoul Bike System with the corresponding weather data 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 -**** mm

### ****Snowfall -**** cm

### ****Seasons -**** Winter, Spring, Summer, Autumn

### ****Holiday -**** Holiday/No holiday

### ****Functional Day -**** (Non Functional Hours), Fun(Functional hours)

**4. Step involved:**

* Exploratory Data Analysis
* Top of Form
* Top of Form

Exploratory data analysis (EDA) is a term for certain kinds of initial analysis and findings done with data sets, usually early on in an analytical process. Some experts describe it as “taking a peek” at the data to understand more about what it represents and how to apply it. Exploratory data analysis is often a precursor to other kinds of work with statistics and data.

* Encoding of Categorical columns

We used One Hot Encoding 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 numerical format.

* Standardization of features

Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning.

* Fitting different models

For modeling we tried various classification algorithms like.

* Linear Regression
* Random Forest Classifier
* Lasso Regression
* Ridge Regression
* Elastic Regression
* Decision Tree
* Random Forest
* Hyper parameter Tunning

**Algorithms:**

1. **Linear Regression:**

The linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models. We will use this model to predict the required number of rental bikes required at each hour.

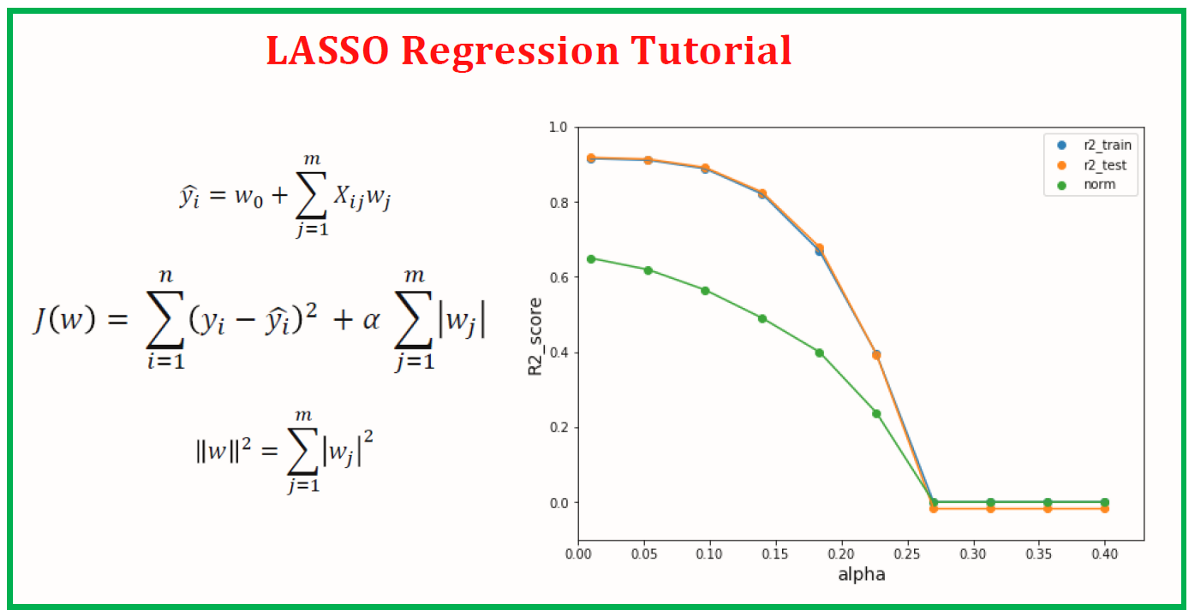
1. **Random forest classifier:**

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicated out of all.

****

**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)

****

**Hyper parameter Tunning:**

Hyper parameters 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 hyper parameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyper parameters grid that can be adjusted according to the business problem. Hyper parameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyper parameter tuning. This also results in cross validation and in our case we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

**Grid Search CV**

Grid Search combines a selection of hyper parameters 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.

It helps to loop through predefined hyper parameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyper parameters.

**Randomized Search CV**

Random search is a technique where random combinations of the hyper parameters are used to find the best solution for the built model. It is similar to grid search, and yet it has proven to yield better results comparatively.

**Bayesian Optimization**

Bayesian optimization is a global optimization method for noisy black-box functions. Applied to hyper parameter optimization, Bayesian optimization builds a probabilistic model of the function mapping from hyper parameter values to the objective evaluated on a validation set.

**Conclusion:**

Starting with loading the data so far we have done EDA null values treatment, encoding of categorical columns, feature selection and then model building.

After comparing the root mean squared error and mean absolute error of train and test result of all the applied model on the dataset, I found that Random forest gave the highest R2 score of 98% and 91% on train and test data respectively.

Hence can be concluded that random forest is best model for predicting bike rental for each hour and lessens the waiting time and enhancing the. mobility comfort.