Seoul Bike Sharing Demand Prediction

Iqbal Babwane,
Sameer Ansari,
Lukman Haider
Data Science Trainees,
AlmaBetter, Mumbai

Abstract:

This Technical paper presents a rule-based regression predictive model for bike sharing demand prediction. A bike-sharing system provides people with a sustainable mode of transportation and has beneficial effects for both the environment and the user.

In recent days, Pubic rental bike sharing is becoming popular because of is increased comfortableness and environmental sustainability. Data used include Seoul Bike and Capital Bike share program data. Data have weather data associated with it for each hour. For the dataset, we are using linear regression model were train with optimize hyperparameters using a repeated cross validation approach and testing set is used for evaluation. Multiple evaluation indices such as R^2, Adjusted R^2, Root Mean Square Error, Mean Square Error and Mean Absolute Error are used to measure the prediction performance of the regression models. The performance of the model is varied with the time interval used in transforming data.

1. Introduction

The increased usage of private vehicles in metropolitan areas has resulted in significant rise in fuel consumption's that have adverse effect on the climate. It has led people in today's society to accept problems like road traffic as the norm. Therefore, the government and organizations started adopting measures to facilitate sustainable development to address the issue. Many countries have bike sharing system, such as bike sharing system in South Korea, which started to overcome all this issues and to develop a healthy environment for citizen of Seoul to live. In that context, the Bike Share initiative was launched to tackle the public mobility problem. It provided the people with an alternative to using a sustainable mode of transport for a small distance at a minimal cost. And gave people the freedom to utilize the service by themselves. In a bike-share system, a user could lend a bike from any bike stations and return it to a bike station near the destination and since it involves the activity of pedaling the bike it has beneficial health effects. And the city-wide installation of bike stations improved the accessibility of areas by bikes. Docking stations are computerized stands for the purpose of pickup and drop off of the rental bikes. Users of public bikes can rent and return rental bikes at any docking station. Users can verify their trip details (distance, duration) and measure of bodily activities (burnt calories) at My Page > Usage Details. With this kind of smart technology and convenience, the use of Rental bike is increasing every day. So, there is a need to manage the bike rental demand and manage

the continuous and convenient service for the users. This study proposes a data mining-based approach including weather data to predict whole city public bike demand. A rule-based model is used to predict the number of rental bikes needed at each hour.

2. 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.

3. Data Description

Column	**Description**
Date	Date of Rented Bike
Rental Bike count	Number of total rentals
Hour	Hour of the day
Temperature	Temperature in Celsius
Humidity	Humidity of the day %
Windspeed	Wind speed in m/s
Visibility	Atmospherically visibility within10m range
Dew point temperature	Dew point Temperature- T dp in Celsius
Solar radiation	Indicate light and energy that comes from the sun in MJ/m^2
Rainfall	Rain falls in mm
Snowfall	Snow falls in cm
Seasons	Winter, Spring, Summer, Autumn
Holiday	Holiday/No holiday
Functional Day	Neither a weekend nor holiday

4. Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Data cleaning means fixing bad data in your data set.

Bad data could be:

- Empty cells
- Data in wrong format
- Wrong data
- Duplicates

5. Data Visualization

Data visualization is the discipline of trying to understand data by placing it in a visual context so that patterns, trends, and correlations that might not otherwise be detected can be exposed.

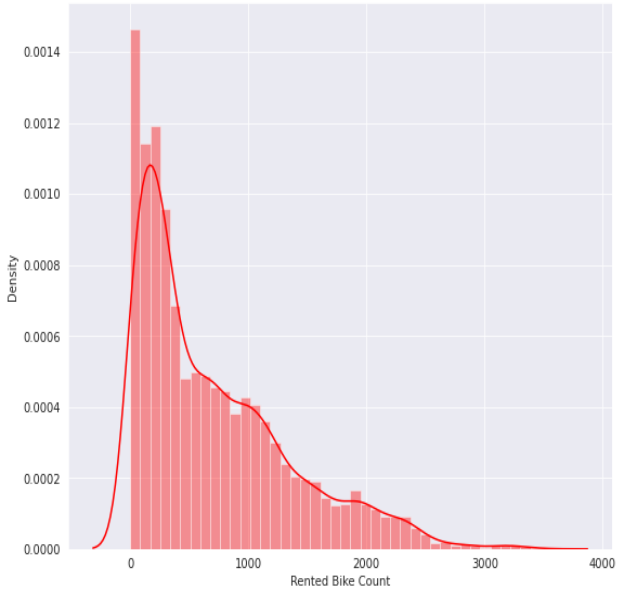
Python offers multiple great graphing libraries packed with lots of different features. Whether you want to create interactive or highly customized plots, Python has an excellent library for you.

To get a little overview, here are a few popular plotting libraries:

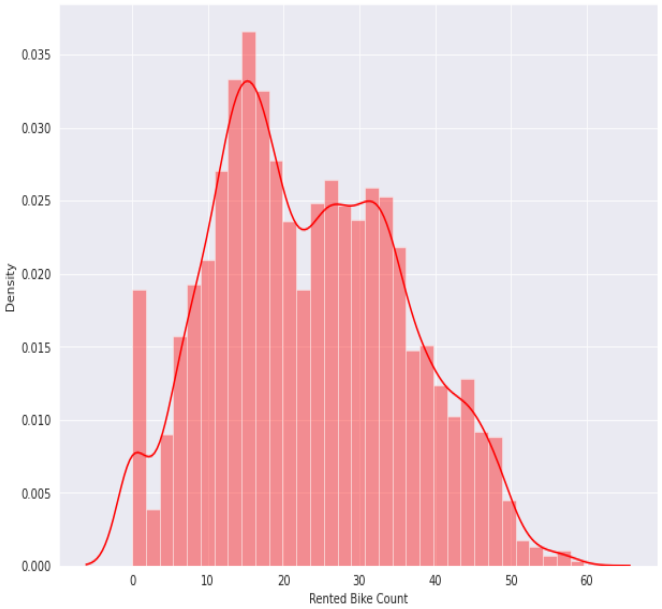
- Matplotlib: low level, provides lots of freedom
- Pandas Visualization: easy to use interface, built on Matplotlib
- Seaborn: high-level interface, great default styles

i. Observation - 1

In this Observation, we plot distplot for rented bike count and we observe that the plot performs right skew distribution.

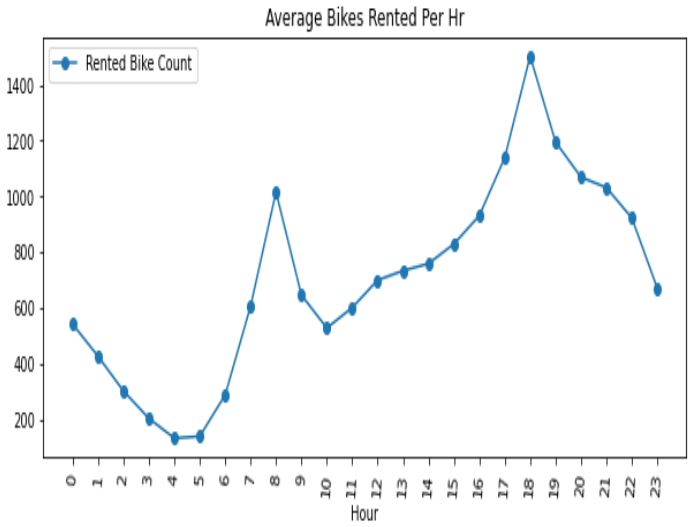


Now we use square root transformation and we see some improvement in plot. Now plot is behaving like normal distribution.



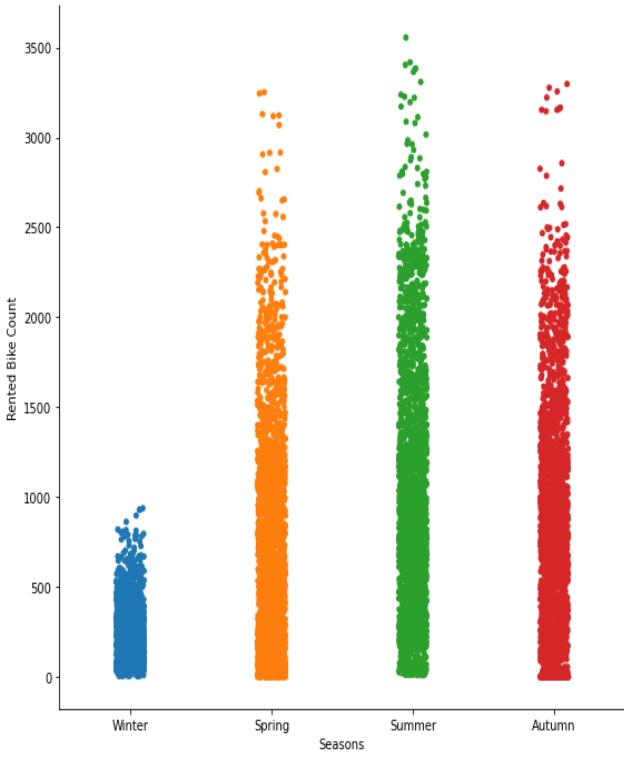
ii. Observation – 3

In this observation, we see that maximum rented bike count is at 6 pm and the minimum count is 4 am and 5 am.



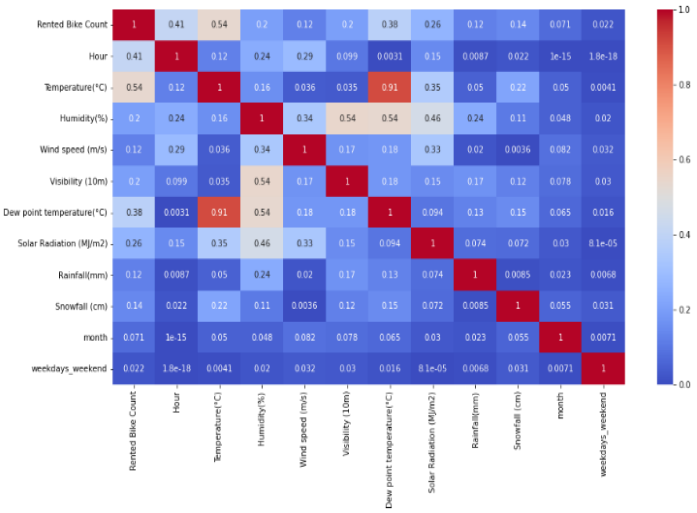
iii. Observation – 3

In this observation, we see bar plot for season wise count of rented bike.
Spring, Summer and Autumn are high count and winter is low rented bike count.



iv. Observation – 4

In this plot, we see the correlation between for all the features.
Temperature and dew point temperature are very strong positive correlated with each other.

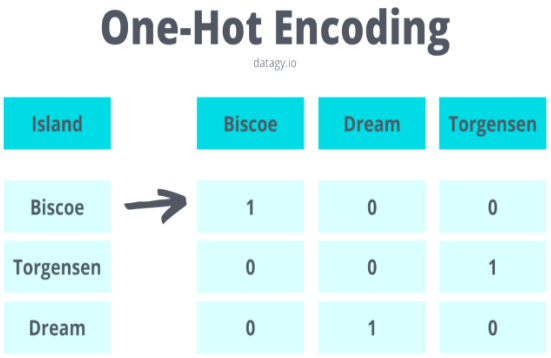


5. Feature Engineering

Feature engineering is the act of converting raw observation into desired features using statistical or machine learning approaches. Feature engineering refers to manipulation- addition, deletion, combination, mutation of our dataset to improve machine learning model training, leading to better performance and greater accuracy. Effective feature engineering is based on sound knowledge of business problem and the available data sources.

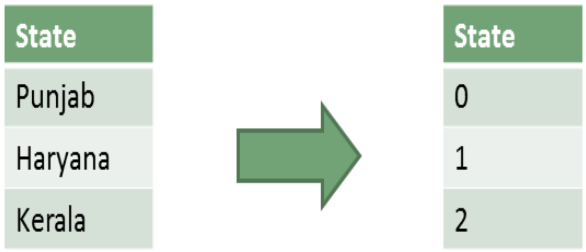
i. One hot encoder data

One-Hot encoding is used in machine learning as a method to quantify categorical data.
One-hot encoding approach eliminates the order but it causes the number of columns to expand vastly. So, for columns with more unique values try using other techniques like Label Encoding



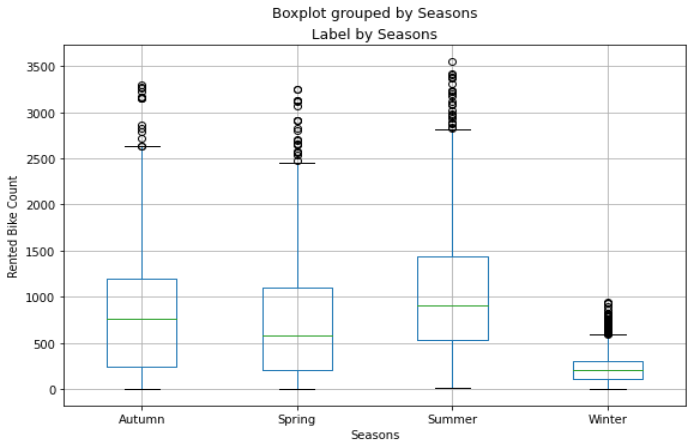
ii. Label Encoder

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

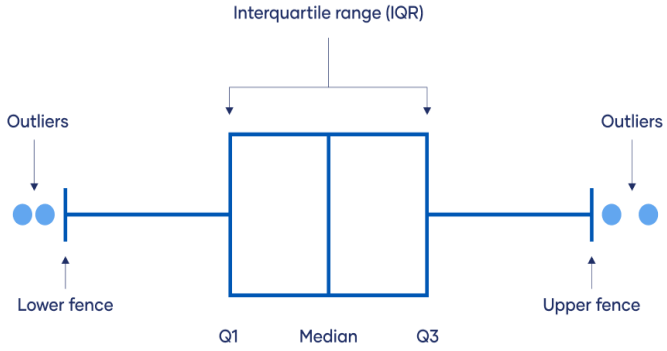


6. Outlier

Outliers is a data point in the dataset that differs significantly from the other data or observation. The thing to remember that, not all outliers are the same. Some have a strong influence, some not at all. Some are valid and important data values. Some are simply errors or noise. Many parametric statistics like mean, correlations, and every statistic based on these is sensitive to



Analysis of outlier



- **Outlier detection**

  We use following methods to detect Outlier using Interquartile Range.

i. Square root

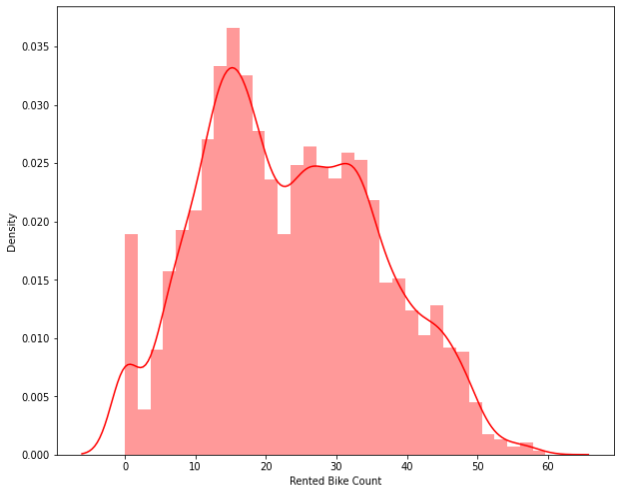
The square root method is typically used when your data is moderately skewed. Now using the square root (e.g., sqrt(x)) is a transformation that has a moderate effect on distribution shape. It is generally used to reduce right skewed data. Finally, the square root can be applied on zero values and is most commonly used on counted data. Square Root Transformation: Transform the values from y to √y.

ii. **Log Transformation**

The logarithmic is a strong transformation that has a major effect on distribution shape. This technique is, as the square root method, octenyl used for reducing right skewness. Worth noting, however, is that it cannot be applied to zero or negative values. Log Transformation: Transform the values from y to log(y).

iii. **Cube root transformation**

Cube root transformation involves converting x to $x13$. This is a fairly strong transformation with a substantial effect on distribution shape: but is weaker than the logarithm. It can be applied to negative and zero values too. Negatively skewed data Cube Root Transformation: Transform the values from y to $y^{1/3}$.



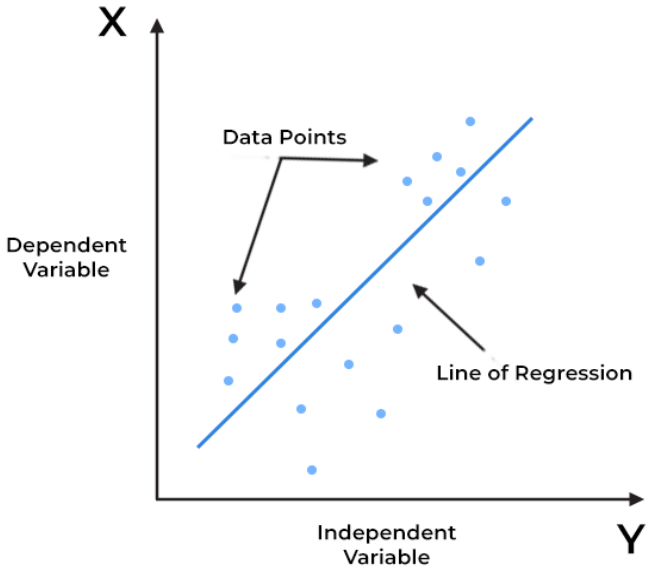
After applying square root transformation our 'Rented bike counted' plot looks like normal distribution

7. Fitting different models

i. Linear Regression
ii. Lasso Regression
iii. Ridge Regression
iv. Elastic net Regression
v. Decision trees
vi. Bagging Regressor
vii. Random Forest
viii. Gradient Boosting
ix. Extreme Gradient Boosting
x. Light Gradient Boosting Machine

Linear Regression:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. LR makes prediction for continuous as well as numeric variables.



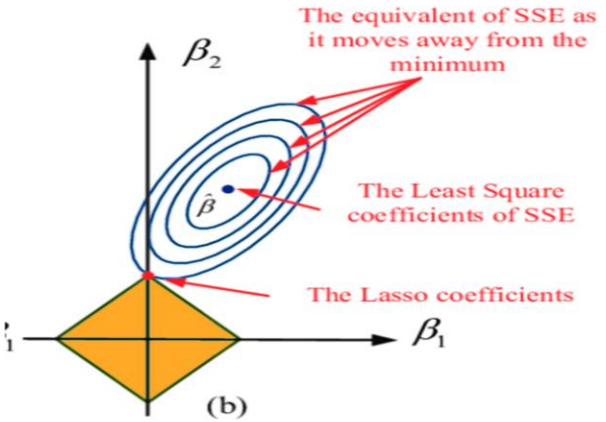
Linear Regression: Single Variable

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

Predicted output — Coefficients — Input — Error

Linear Regression: Multiple Variables

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$



Linear regression shows the relationship between a dependent and one or more independent variables. The following equation defines an LR line: $Y=a+bX$

It is done by fitting a linear equation of line to the observed data. For fitting the model, it is more important to check, whether there is a connection between the variables or features of interest, which is supposed to use the numerical variables, i.e., the correlation coefficient.

Lasso Regression:

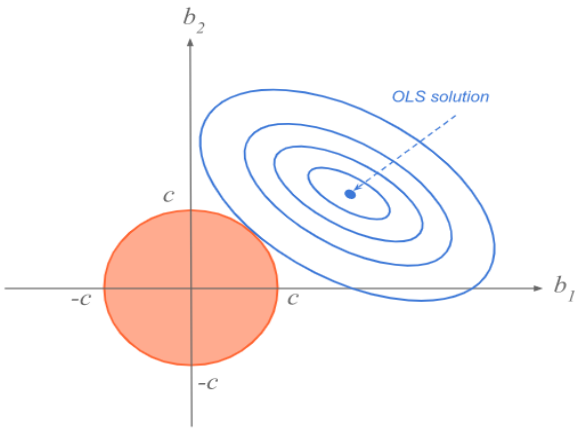
LASSO stands for Least Absolute Shrinkage and Selection Operator.

The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero. Is also used as L1 regularization. The equation for the cost function of Lasso regression will be:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Ridge Regression:

Ridge regression is a model method that is used to analyses any data that suffers from multicollinearity and it reduce the complexity of the model. 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. It is also used as L2 Regularization. The equation for the cost function in ridge regression will be:



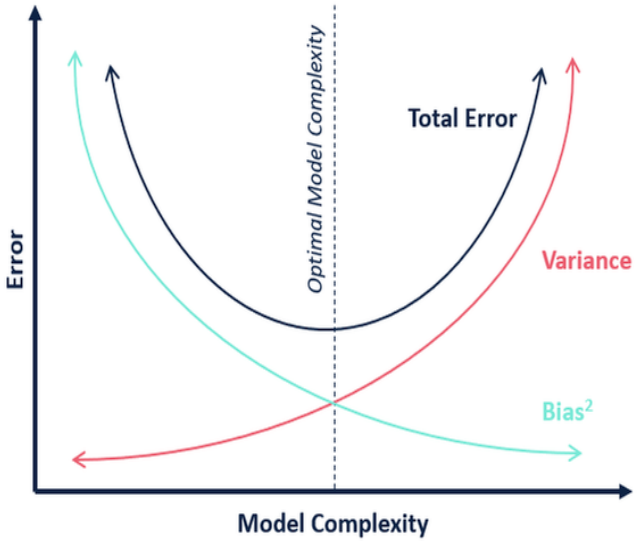
$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} {\beta_j}^2$$

Overfitting: So, what is overfitting? Well, to put it in more simple terms it's when we built a model that is too complex that it matches the training data "too closely" or we can say that the model has started to learn not only the signal, but also the noise in the data. The result of this is that our model will do well on the training data, but won't generalize to out-of-sample data, data that we have not seen before.

Bias-Variance tradeoff: When we discuss prediction models, prediction errors can be decomposed into two main subcomponents we care about: error due to "bias" and error due to "variance". Understanding these two types of error can help us diagnose model results and avoid the mistake of over/under fitting. A typical graph of discussing this is shown below:

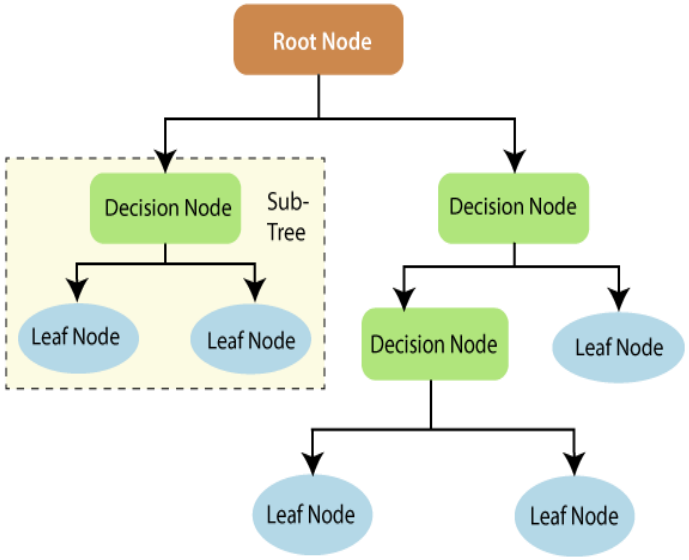
Bias: The red line, measures how far off in general our models' predictions are from the correct value. Thus, as our model gets more and more complex, we will become more and more accurate about our predictions (Error steadily decreases).

Variance: The cyan line, measures how different can our model be from one to another, as we're looking at different possible data sets. If the estimated model will vary dramatically from one data set to the other, then we will have very erratic predictions, because our prediction will be extremely sensitive to what data set, we obtain. As the complexity of our model rises, variance becomes our primary concern.



Decision trees:

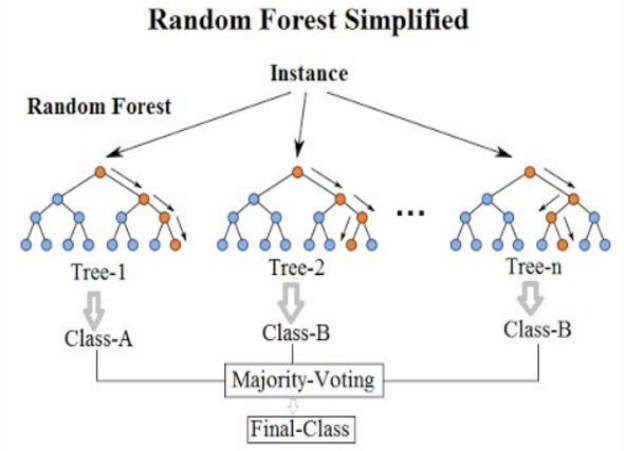
Decision Tree is a supervised learning method used in data mining for classification and regression methods. It is a tree that helps us in decision-making purposes. It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed. The final tree is a tree with the decision nodes and leaf nodes.



Random Forest:

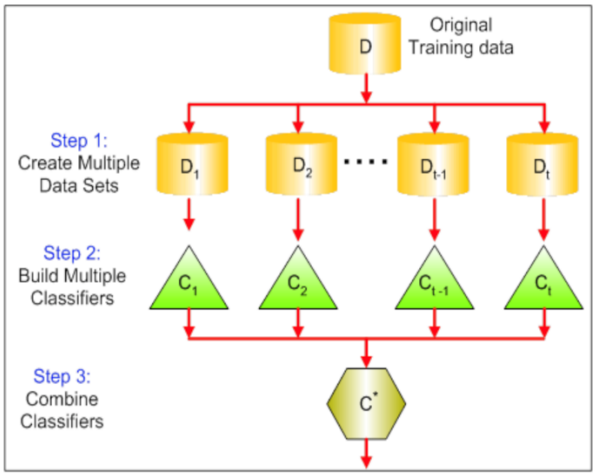
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique.

"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."



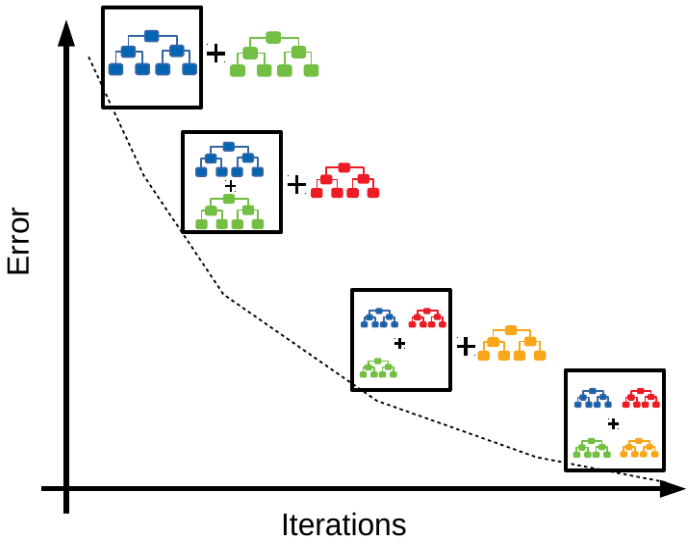
Bagging Regressor:

Bagging is an ensemble technique used to reduce the variance of our predictions by combining the result of multiple classifiers modelled on different sub-samples of the same data set. The following figure will make it clearer:



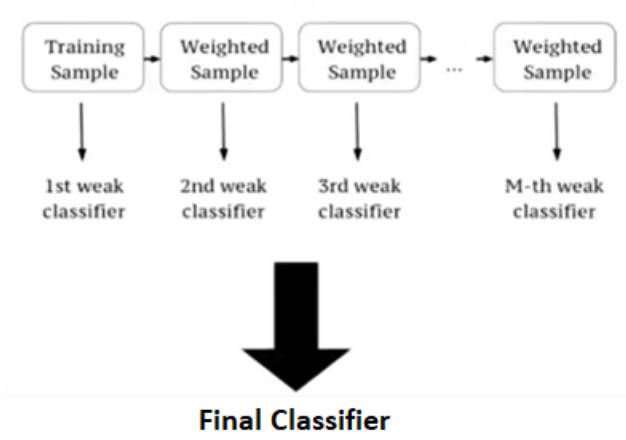
Gradient Boosting:

Gradient Boosting algorithm is used to generate an ensemble model by combining the weak learners or weak predictive models. Gradient boosting algorithm can be used to train models for both regression and classification problem. Gradient Boosting Regression algorithm is used to fit the model which predicts the continuous value.



Extreme Gradient Boosting

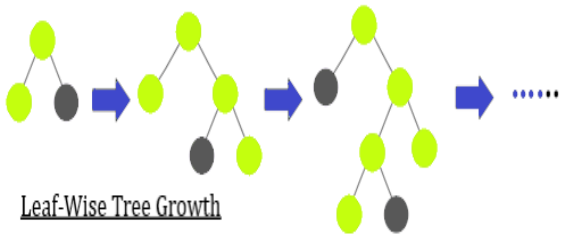
XGBoost (eXtreme Gradient Boosting) is one of the most loved machine learning algorithms. Teams with this algorithm keep winning the competitions. It can be used for supervised learning tasks such as Regression, Classification, and Ranking. It is built on the principles of gradient boosting framework and designed to "push the extreme of the computation limits of machines to provide a scalable, portable and accurate library."



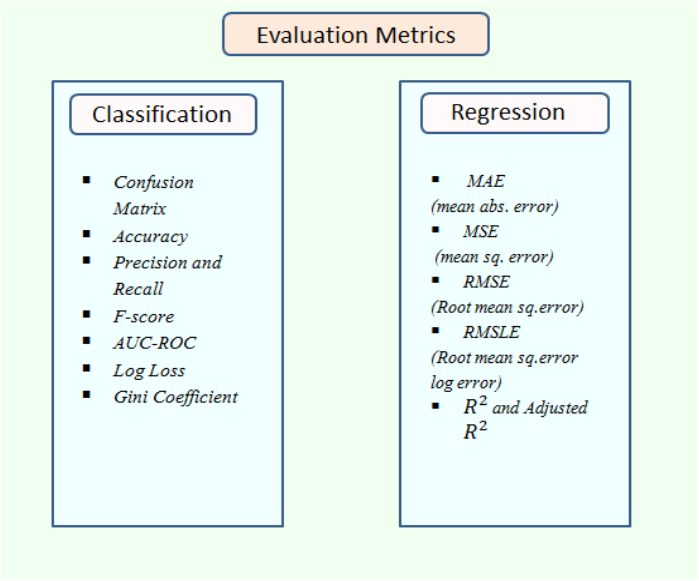
Light Gradient Boosting Machine:

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages:

i. Faster training speed and higher efficiency.
ii. Lower memory usage.
iii. Better accuracy.
iv. Support of parallel, distributed, and GPU learning.
v. Capable of handling large-scale data. MODEL

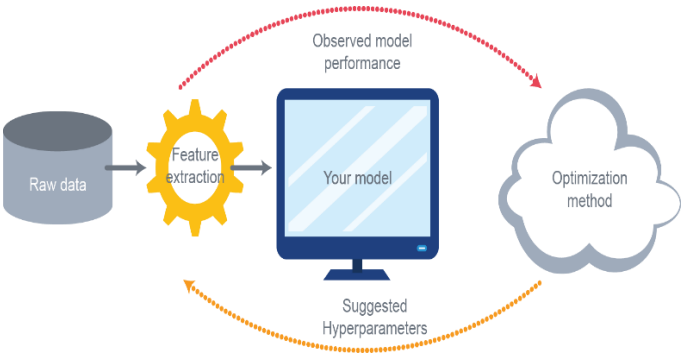


8. Model Evaluation:



Hyper parameter tuning:

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fit the model parameters. However, there is another kind of parameter, known as Hyperparameters, that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.



Hyperparameters are those parameters that are explicitly defined by the user to control the learning process. Some key points for model parameters are as follows:

- These are usually defined manually by the machine learning engineer.
- One cannot know the exact best value for hyperparameters for the given problem. The best value can be determined either by the rule of thumb or by trial and error.
- Some examples of Hyperparameters are the learning rate for training a neural network, K in the KNN algorithm.

Grid Search CV:

The Grid Search Method considers some hyperparameter combinations and selects the one returning a lower error score. This method is specifically useful when there are only some hyperparameters in order to optimize. However, it is outperformed by other weighted-random search methods when the Machine Learning model grows in complexity.
Grid Search is an optimization algorithm that allows us to select the best parameters to optimize the issue from a list of parameter choices we are providing, thus automating the 'trial-and-error' method. Although we can apply it to multiple optimization issues; however, it is most commonly known for its utilization in machine learning in order to obtain the parameters at which the model provides the best accuracy.

Randomized Search CV:

In Random Search, the hyperparameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions of the cost minimization space with more suitable hyperparameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyperparameters is beyond the scientist's control.

9. Conclusion:

This study focused on predicting the bike sharing demand using given dataset. Regression techniques Linear Regression, Lasso Regression Ridge Regression, Elastic Net, Decision Tree, Bagging Regression, Random Forest, Gradient Boosting Regressor, XGB Regressor, Light-GBM, MLP Regressor are used to predict. This statistical data analysis shows interesting outcomes in prediction method and also in an exploratory analysis.

- Heat map shows Temperature and Dew point temperature is highly correlated.
- Bike is rented when functioning day is there otherwise not.
- Most number of bikes are rented 17 to **19th hour of the day** and in morning at 8 pm.
- Most numbers of Bikes were rented in **summer**, followed by **autumn**, **spring**, and **winter**.
- Most number of bikes are rented on **Working day** instead of holiday.

This is evident from EDA analysis where bike demand is more on weekdays, working days in Seoul.

hence the prediction from the linear model was very low. Best predictions are obtained with a **LightGBM** model with an R^2 score of **0.919** and RMSE score of **183.21**

10. References:

1. Stackoverflow
2. Almabetter
3. GeeksforGeeks
4. w3school