Technical Documentation- Seoul Bike Sharing Demand Prediction

Presented by- Vashu Garg, Palak Bindal, Deepika Gupta, Soumya Jain, Jyoti Singh

ABSTRACT

This research 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 Bikeshare 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 , Root Mean Square error are use to measure the prediction performance of the regression models. The performance of the model is vary with the time interval used in transforming data.

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 organzations 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 pedalling 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.

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.

The main objective is to make predictive model, which could help them in predicting the bike demands proactively. This will help them in stable supply of bike wherever needed.

DATASET DESCRIPTION

Date	Date of Rented Bike
Rental Bike count	Number of tota rentals
Hour	Hour of the day
Temperature	Weather Temperature in Celsius
Humidity	Humidity of the day %
Windspeed	Wind speed in m/s
Visibility	Atmospherical 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/m²
Rainfall	Rain fall in mm
Snowfall	Snow fall in cm
Seasons	Winter, Spring, Summer, Autumn
Holiday	Weather the day is considered a Holiday/No holiday
Functional Day	Whether the day is neither a weekend nor holiday

BREAKDOWN OF DATASETS

Before proceeding to data visualization, we need to perform the following steps:

1. Importing required packages for future analysis.
2. Mounting drive and reading data files from Google drive.
3. Removing future warning seaborn plots.
4. Visualizing all the columns of the respective data frame.
5. Viewing all data information.
6. Checking duplicates if any then drop.
7. Checking unique values, null count, datatypes and null value percentage.
8. Filtering data.
9. Segregation of numerical and categorical data.

EXAMINING NULL / MISSING VALUES

Null values are a big problem in machine learning and deep learning. If you are using sklearn, TensorFlow, or any other machine learning or deep learning packages, it is required to clean up null values before you pass your data to the machine learning or deep learning framework. Otherwise, it will give you a long and ugly error message. So we are checking for null/ missing values. There is no missing value and no null value in provided dataset.

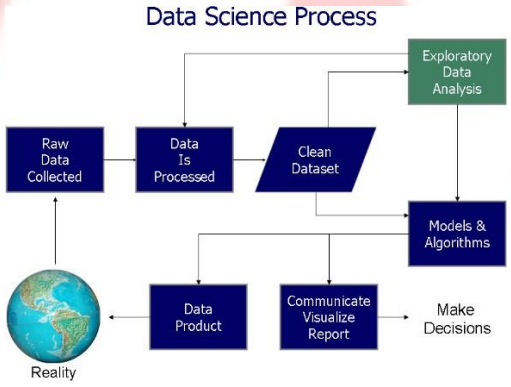
DATA CLEANING

Data cleaning is the foremost step in any data science project. No data is clean, but most is useful. Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing,

modifying, or deleting the dirty or coarse data. To begin with our data cleaning, first we check for duplicate values and there is no duplicate values in given dataset. After doing so we are converting datatypes, and then we have done exploratory data analysis and find best fit model of dataset.

EXPLORATORY DATA ANALYSIS

In statistics, exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing. EDA is helped us figuring out various aspects and relationships among the target and the independent variables.

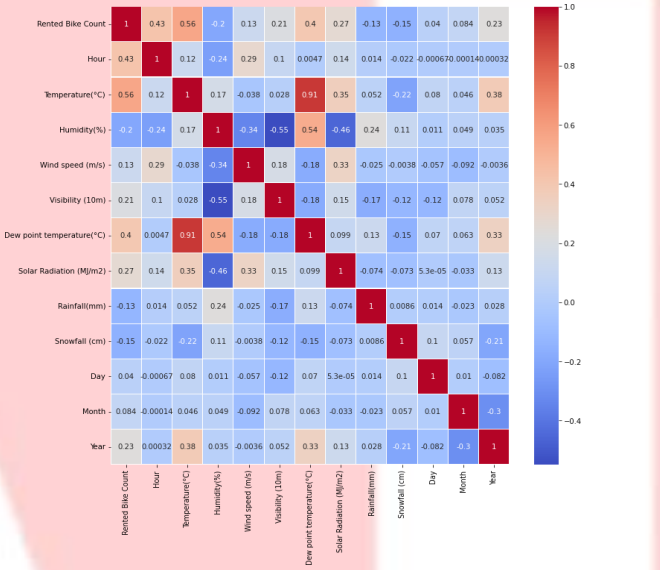


Observation 1:

Correlation is a statistical measure that expresses the strength of the relationship between two variables. Positive correlation occure when two variables move in same direction; as one increases so do the other. Negative correlation occurs when two variables move in opposite directions; as increases, the other decreases.

Correlation can be used to test hypotheses about cause effect relationships between variables. Correlation is often used in the real world to predict trends.

Temperature and Dew point temperature are almost 0.91 correlated, so it's generate multicollinearity issue. So we drop Dew point temperature feature.



Observation 2:

Data types are an important aspect of statistical analysis, which needs to be understood to correctly apply statistical methods to your data.

During the data collection phase, the researcher may collect both numerical and categorical data when investigating to explore different perspectives. However, one needs to understand the differences between these two data types to properly use it in research.

We treat numeric and categorical variables differently in Data Wrangling. So, we should always make at least two sets of data: one contains numeric variables and other contains categorical variables.

We plot numerical data to analysis data distribution:

Here we can see some attributes are normally distributed but some are positively or negatively skewed.

Right/Positive Skewed Distribution: Mode < Median < Mean: Rented Bike Count, Wind Speed(m/s), Solar Radiation(MJ/m2)

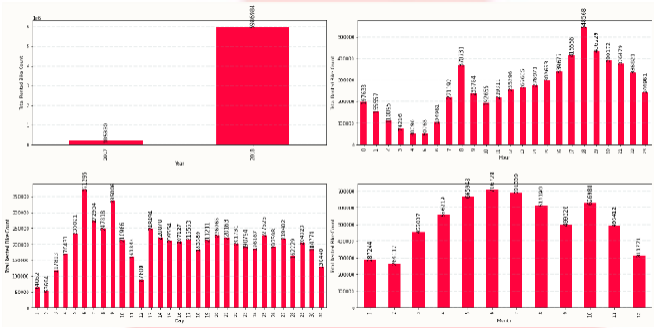
No Skew: Mean = Median = Mode : Hour, Temperature, Humidity(%),Rainfall(mm),Snowfall(cm)

Left/Negative Skewed Distribution: Mean < Median < Mode: visibility(10m)

Observation 3:

Discrete data is the type of numerical data with countable(finite/infinite) elements. i.e. they have one-to-one mapping with natural numbers.

Lets analyze the discrete values by creating histogram to understand the distribution. Discrete variable count is 4.

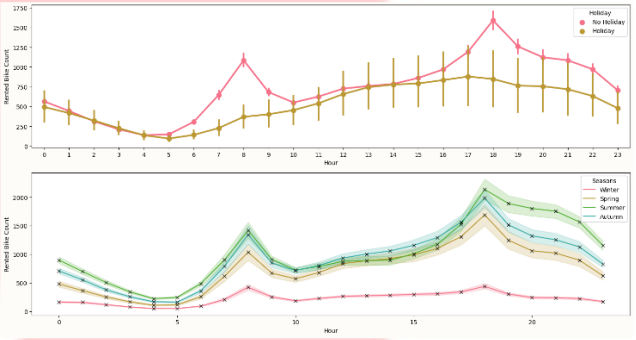


Observation 4:

Categorical data is a type of data that can be stored into groups or categories with the aid of names or labels. This grouping usually made according to the data characteristics and similarities of these characteristics through a method known as matching.

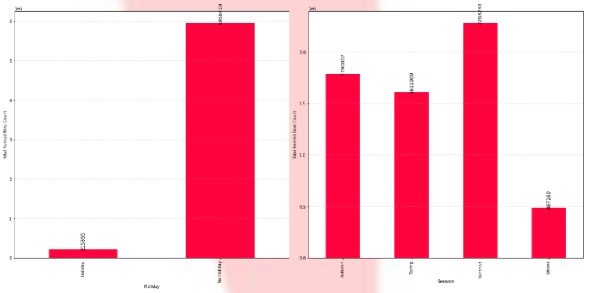
In our data only two categorical features are there:

Seasons: Summer, Autumn, Spring, Winter are 4 categories.
Holiday: Holiday/ No Holiday are two categories.



Here we can see at 18th hour of the day traffic for rental bike count is more. And in every season traffic for rental bike count is more in 17-19th hour of the day.

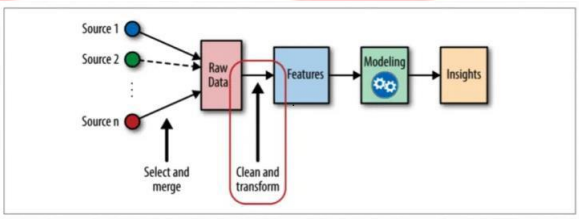
LATION



1. Visualized data is processed faster and easier.
2. Better insights of the data are drawn which may be missed in traditional reports
3. Helps us visualize trends which improve performance.
4. Data visualization increase productivity and sale.

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.



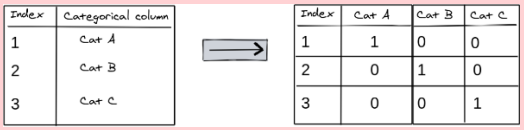
- **Feature coding**

We are encoding categorical data in both encoder and check accuracy of encoders:

1. One Hot Encoder Data
2. Label Encoder Data

- One Hot Encoder:

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 LabelEncoding

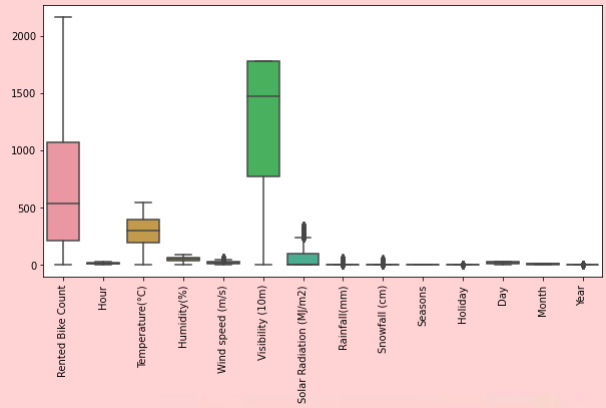
- Label Encoder

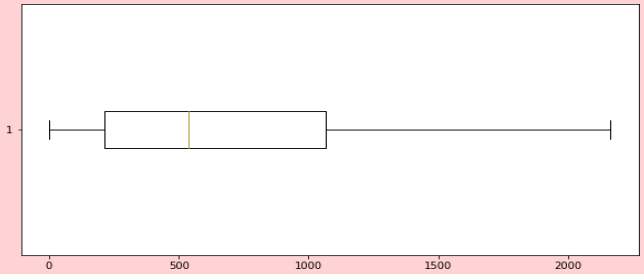
Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

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

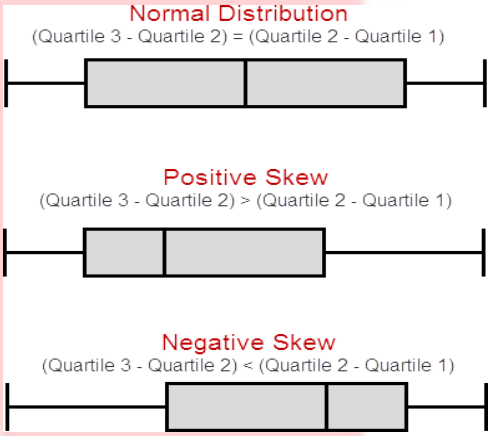
outliers.



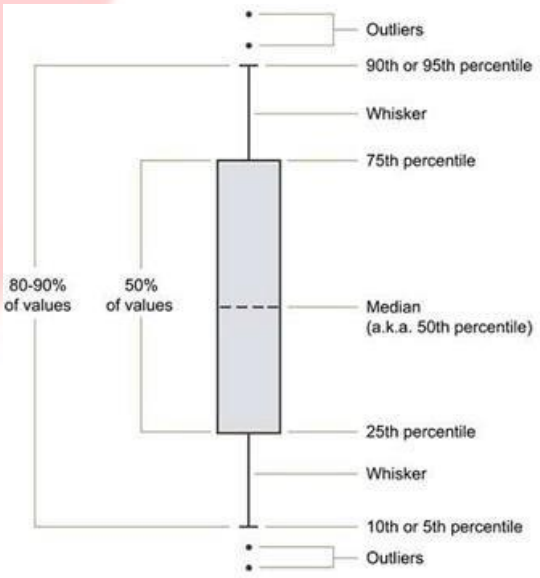


So the data from all the other columns (except first and last column) can be used as inputs to the model.

Examples Of Box Plot



Distribution analysis of outlier



Target Variable

The target variable of the dataset is the feature of a dataset about which you want to gain a deeper understanding.

In our data, the column ‘Rental Bike Count’ contains the value we need to predict i.e. the target variable is ‘Rental Bike Count’. Here is the target parameter Rental Bike Count distribution analysis plot and plot is positively skewed.

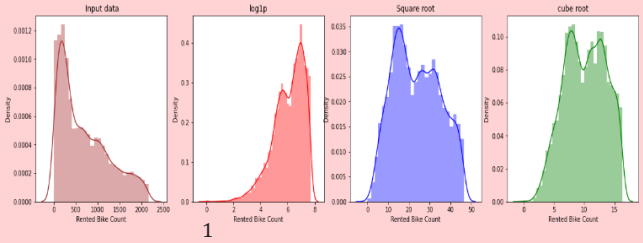
OUTLIER DETECTION

We use following methods to detect Outlier using Interquartile Range

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.

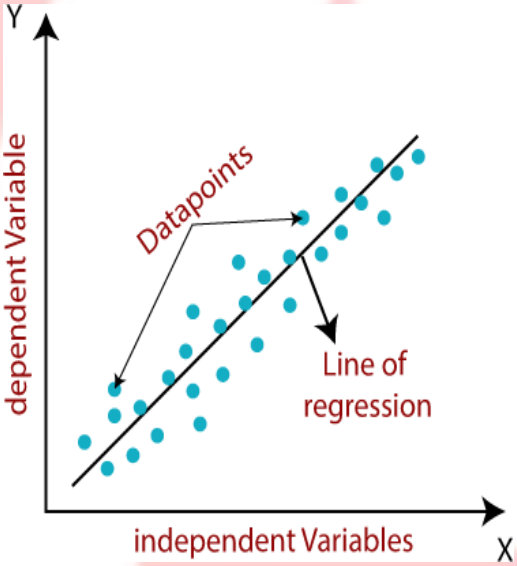


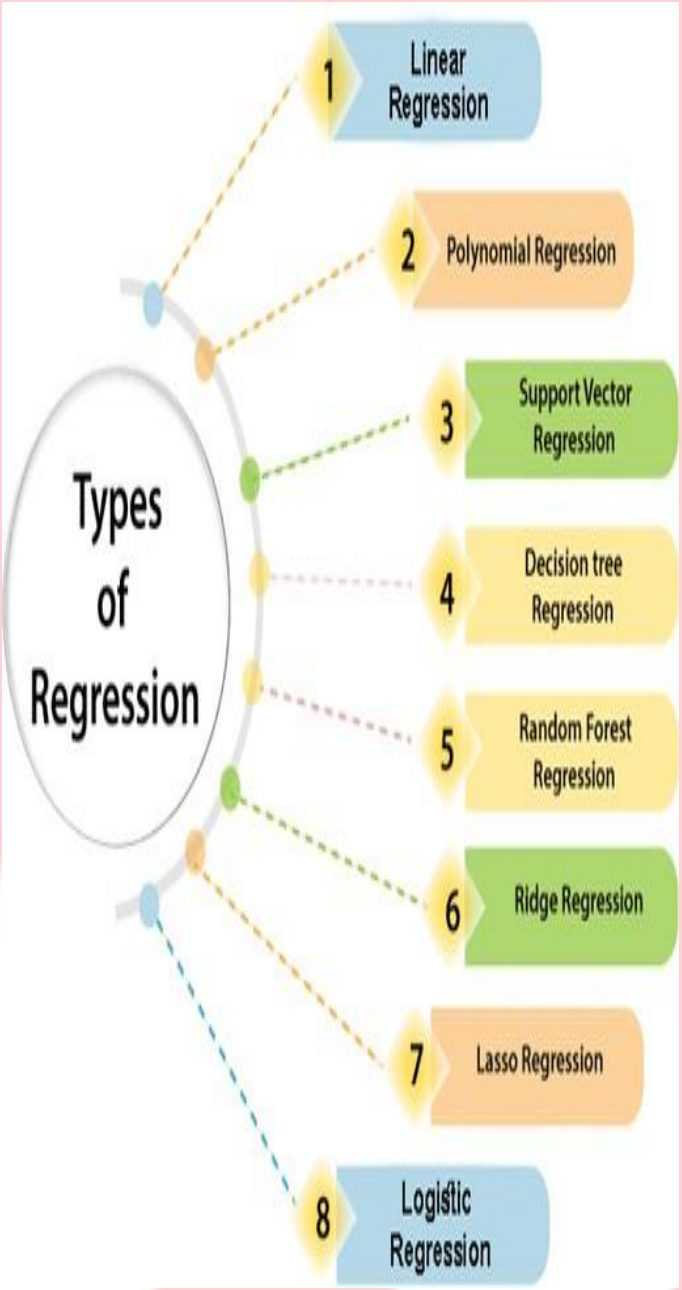
1

FITTING DIFFERENT MODELS ALGORITHMS

I. 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 shows the relationship between a dependent and one or more independent variables. The following equation defines an LR line:

$$Y = a + bX$$

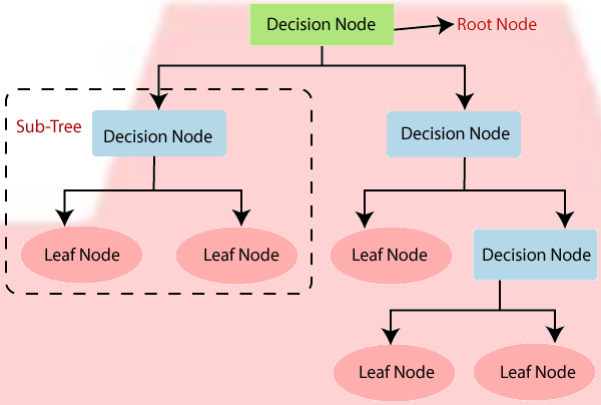
It is done by fitting a linear equation of line to the observed data. For fitting the model, it is more important to check, whether there is a connection between the variables or features of interest, which is supposed to use the numerical variables, i.e. the correlation coefficient.

gistic regression or linear regression, SVMs can handle highly non-linear data using an amazing technique called kernel trick.

Kernel transforms linearly inseparable data to separable data by adding more dimensions to it.

II. Decision Tree Regression

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.

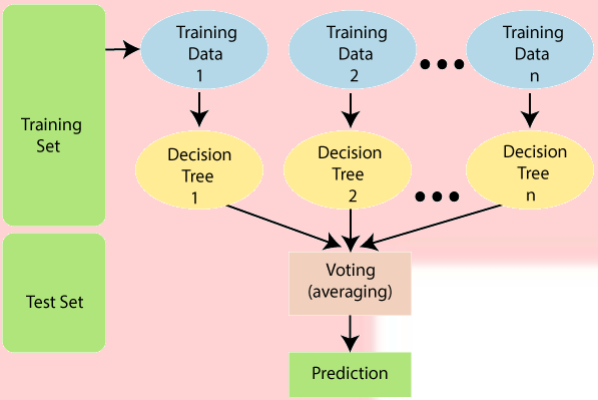


III. Random Forest Regression

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."



The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

IV. 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 fuction in ridge regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}{\beta_j}^2$$

V. Lasso Regression

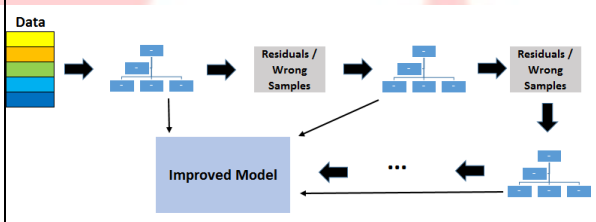
LASSO stands for Least Absolute Shrinkage and Selection Operator.

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 is also used as L1 regularization. The equation for the cost fuction of Lasso regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}|\beta_j|$$

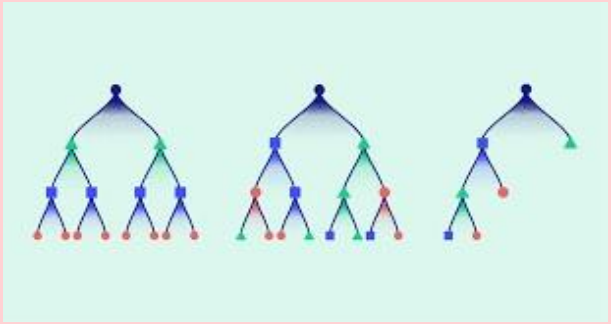
VI. Gradient Boosting Regressor:

Gradient boosting is one of the most popular machine learning algorithms for tabular datasets. It is powerful enough to find any nonlinear relationship between your model target and features and has great usability that can deal with missing values, outliers, and high cardinality categorical values on your features without any special treatment. While you can build barebone gradient boosting trees using some popular libraries such as XGBoost or LightGBM without knowing any details of the algorithm, you still want to know how it works when you start tuning hyper-parameters, customizing the loss functions, etc., to get better quality on your model.



VII. XGBoost Regressor

XGBoost stans for Extreme Gradient Boosting is an open source library that provides an efficient and effective implementation of the gradient boosting algorithm. . Shortly after its development and initial release, XGBoost became the go-to method and often the key component in winning solutions for a range of problems in machine learning competitions. Regression predictive modeling problems involve predicting a numerical value such as a dollar amount or a height. XGBoost can be used directly for regression predictive modeling



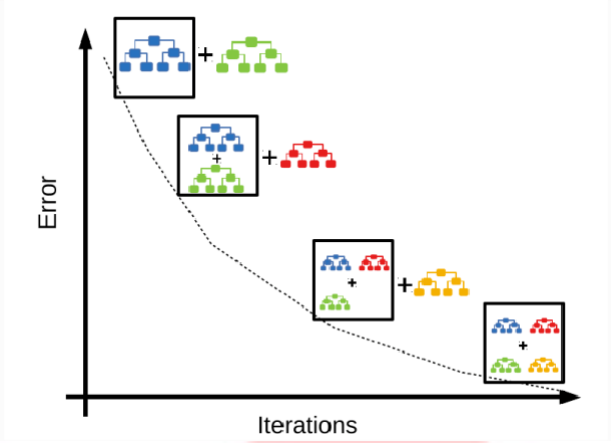
XGBoost is one of the fastest implementations of gradient boosting. trees.

It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits).
XGBoost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits.

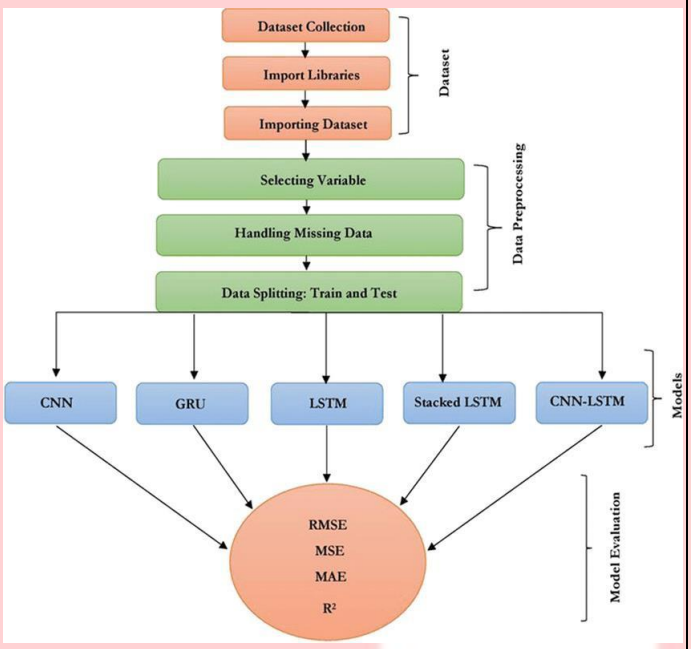
VIII. LightGBM

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

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



MODEL EVALUATION



I. Root Mean Square Error (RMSE)

RSME (Root mean square error) calculates the transformation between values predicted by a model and actual values. In other words, it is one such error in the technique of measuring the precision and error rate of any machine learning algorithm of a regression problem.

RMSE is a square root of value gathered from the mean square error function. It helps us plot a difference between the estimate and actual value of a parameter of the model.

Using RSME, we can easily measure the efficiency of the model.

II. Mean Square Error (MSE)

MSE is a risk method that facilitates us to signify the average squared difference between the predicted and the actual value of a feature or variable.

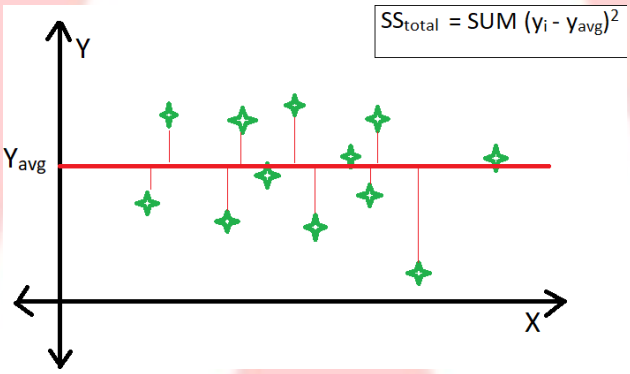
Mean Squared Error is calculated in much the same way as the general loss equation from earlier. We will consider the bias value as well since that is also a parameter that needs to be updated during the training process.

The mean squared error is best explained with an illustration.

III. R-squared (R^2)

R-squared is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.

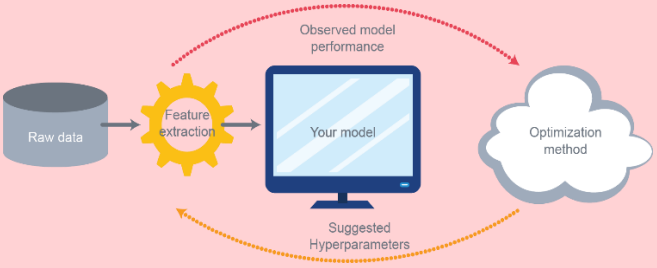
R-square is a comparison of the residual sum of squares with the total sum of squares. The total sum of squares is calculated by summation of squares of perpendicular distance between data points and the average line.



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:

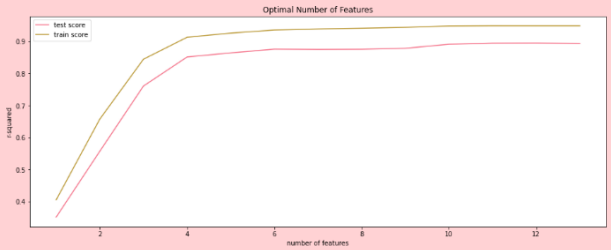
- These are usually defined manually by the machine learning engineer.
- One cannot know the exact best value for hyperparameters for the given problem. The best value can be determined either by the rule of thumb or by trial and error.
- Some examples of Hyperparameters are the learning rate for training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

I. 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.



II. 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.

CONCLUSION

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

The experimental result shows that:

- Heat map shows Temperature and Dew point temperature is highly correlated.
- Bike is rented when functioning day is there otherwise not.
- More number of bike are rented in **year 2018**
- Most number of bike are rented 17 to **19th hour of the day** and in morning at 8 pm.
- Most number of bikes are rented on **6th and 9th day of month.**
- Most numbers of Bikes were rented in **summer**, followed by **autumn**, **spring**, and **winter**. **May-July** is the peak Bike renting Season, and **Dec-Feb** is the least preferred month for bike renting.
- 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.

Feature and Labels had a weak linear relationship; hence the prediction from the linear model was very low. Best predictions are obtained with a **LightGBM** model with an R^2 score of **0.907** and RMSE score of **3.354**