Seoul Bike Sharing Demand Prediction(ML Regression)

Self Project by
Onkar A. Pawar

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.

Keywords: Data Cleaning, EDA, Linear Regression, lasso, Ridge , Decision Tree, Random forest Regression, Gradient Boosting Regression.

1.Problem Statement

Data 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.

2. 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.

3. DATASET DESCRIPTION

Date	Date of Rented Bike
Rental Bike count	Number of total 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/m^2
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

4. 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.

5. EXAMINING NULL / MISSING VALUES

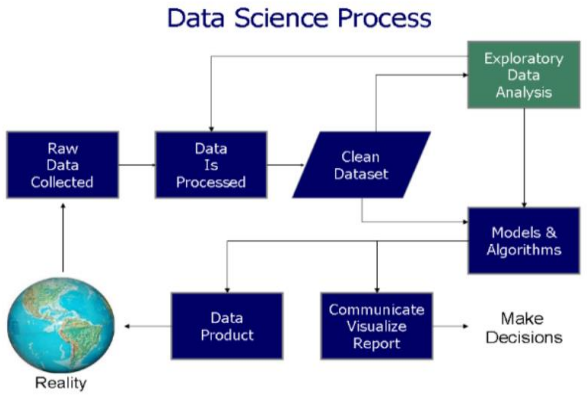
Null values are a big problem in machine learning and deep learning. If you are using sklearn, TensorFlow, or any other machine learning or deep learning packages, it is required to clean up null values before you pass your data to the machine learning or deep learning framework. Otherwise, it will give 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.

6. 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.

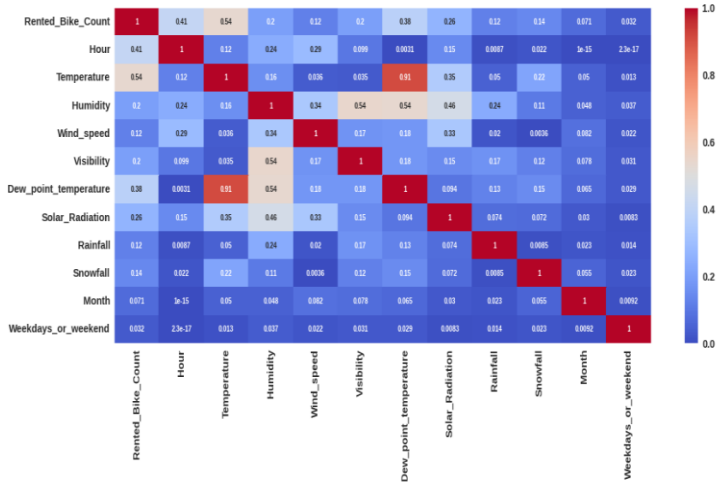
7. EXPLORATORY DATA ANALYSIS

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.

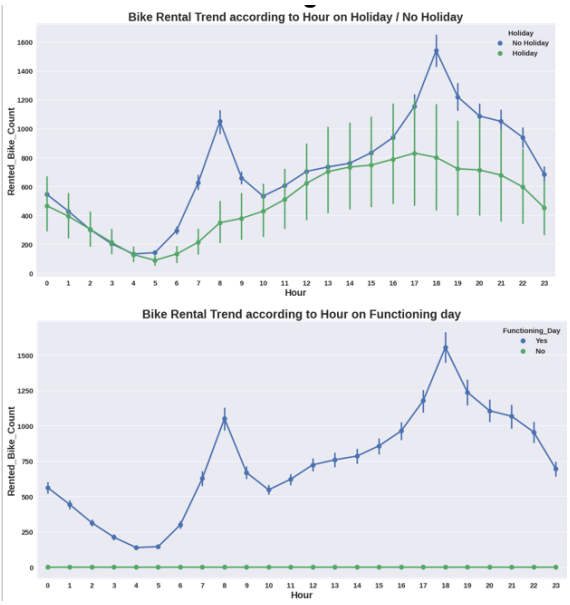


Observation 2:

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.



Observation 3:

Here we observed that, Bike rental trend according to hours is almost similar in all scenarios.

There is sudden peak between 6/7AM to 10 AM. Office /College going time could be the reason for this sudden peak on NO Holiday. But on Holiday the case is different, very less bike rentals happened. Again there is peak between 4PM to 7 PM. may be its office leaving time for the above people.(NO Holiday).

Here the trend for functioning day is same as of No holiday. Only the difference is on No functioning day there were zero bike rentals.

7.1. ADVANTAGE OF VISUALIZATION

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.

8. MODEL SELECTION AND EVALUATION

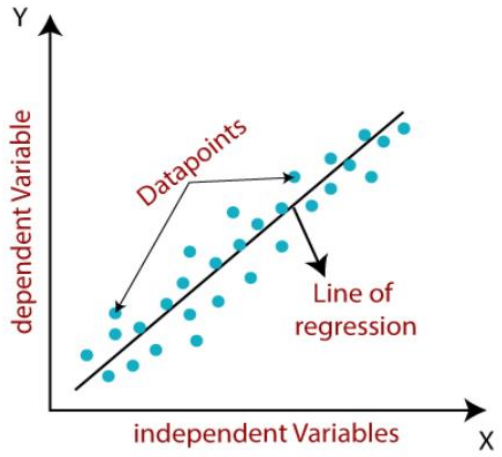
As this is the regression problem we are trying to predict continuous value. For this we used following regression models.

1. Linear Regression
2. Lasso regression (regularized regression)

3. Ridge Regression(regularized regression)
4. Decision Tree regression.
5. Random forest regression
6. Gradient Boosting regression.

1.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.



2.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:

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

3.Ridge Regression:

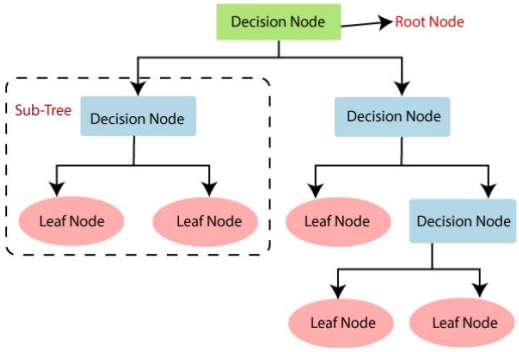
Ridge regression is a model method that is used to analyses any data that suffers from multicollinearity and it reduce the complexity of the model. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values. It is also used as L2 Regularization.

The equation for the cost function in ridge regression will be:

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

4.Decision Tree Regression

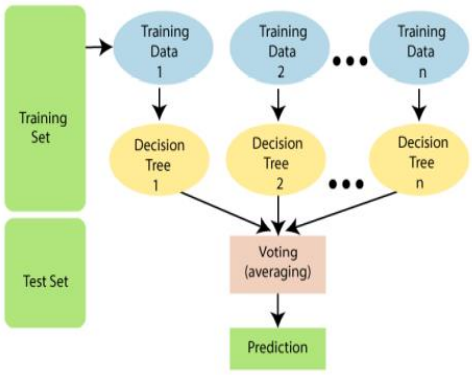
Decision Tree is a supervised learning method used in data mining for classification and regression methods. It is a tree that helps us in decision-making purposes. It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed. The final tree is a tree with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



5.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 the dataset".

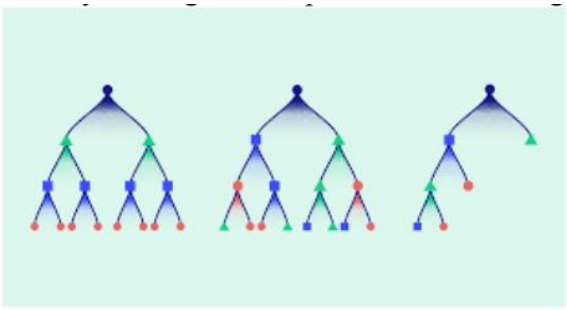


6. Gradient Boosting Regressor

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.

7.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.

8.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.

8.2 Mean Square Error (MSE):

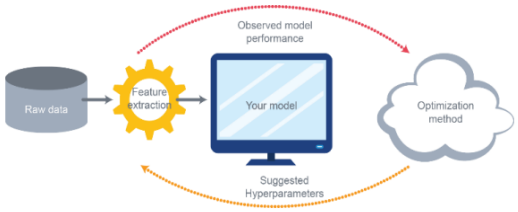
MSE is a risk method that facilitates us to signify the average squared difference between the predicted and the actual value of a feature or variable.
Mean Squared Error is calculated in much the same way as the general loss equation from earlier. We will consider the bias value as well since that is also a parameter that needs to be updated during the training process.
The mean squared error is best explained with an illustration.

8.3 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.

8.4 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:

1.These are usually defined manually by the machine learning engineer.

2.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.

3.Some examples of Hyperparameters are the learning rate for training a neural network, K in the KNN algorithm.

8.5 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.

8. Conclusion:

As we have calculated MAE,MSE,RMSE and R2 score for each model. Based on r2 score will decide our model performance.
Our assumption: if the difference of R2 score between Train data and Test is more than 5 % we will consider it as over fitting.
Linear, Lasso, Ridge and Elastic Net:
Linear, Lasso, Ridge and Elastic regression models have almost similar R2 scores(61%) on both training and test data.(Even after using Grid serach CV we have got similar results as of base models).
Decision Tree Regression:
On Decision tree regressor model, without hyper-parameter tuning, we got r2 score as 100% on training data and on test data it was very less. Thus our model memorized the data. So it was a over fitted model.
After hyper-parameter tuning we got r2 score as 88% on training data and 83% on test data which is quite good for us.
Random Forest:
On Random Forest regressor model, without hyper-parameter tuning we got r2 score as 98% on training data and 90% on test data. Thus our model memorized the data. So it was a over fitted model, as per our assumption
After hyper-parameter tuning we got r2 score as 90% on training data and 87% on test data which is very good for us.
Gradient Boosting Regression(Gradient Boosting Machine): On Random Forest regressor model, without hyper-parameter tuning we got r2 score as 86% on training data and 85% on test data. Our model performed well without hyper-parameter tuning. After hyper-parameter tuning we got r2 score as 96% on training data and 91% on test data, thus we improved the model performance by hyper-parameter tuning.