SUPERVISED MACHINE LEARNING (Regression)

SEOUL BIKE SHARING DEMAND PREDICTION

AKASH GAWANDE

DATA SCIENCE TRAINEES

(ALMABETTER, BANGALORE)

Introduction

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.

Problem

Focusing on the ultimate goal, which comprehends to combine the actual and past bike usage pattern with season in order to forecast Bike Sharing Demand.

DATA DESCRIPTION

Numbers of Columns given in our data are as follows:

- 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 – NoFunc(Non Functional Hours), Fun(Functional hours)

STEPS INVOLVED:

1. **Exploratory Data Analysis**
    A. Checking the shape of the data.
    B. Examine the head and tail of data to learn about the data served.
    C. Showing data-types and features of columns.
    D. Examining the statistics of feature data set.
    E. Using the missingno (msno) to identify the null value via matrix visualization.
    F. Checking missing values
    G. Changing the "Date" Column into three "year", "month", "day" column.
    H. Extracting correlation heatmap and calculate Variance Inflation Factor (VIF) to check correlation and multicollinear variables.

2. **Extracting Conclusions from the data:**
    A. Plotting graph to visualize the distribution of Rented Bike Count.
    B. Plotting graph for square root transformation of Rented Bike Count.
    C. Analysis of the numeric features.
    D. Plotting distplot of Count Vs. Rented Bike Count/Hour/Temperature/Humidity(%)/Windspeed(m/s)/Visibilty(10m)/Dew Point Temperature/Solar Radiation (MJ/m2)/Rainfall(mm)/Snowfall(cm)/month/weekdays_weekend
    E. Plotting the Regression plot of each column dataset v/s Rented Bike Count Columns.
    F. Extricating the categorical features.
    G. Running a test to check the average bike rented per hours.
    H. Analysis of Monthly Distribution of Rented Bike Count.
    I. Plotting Regression plots with respect to Temperature, Humidity and Windspeed.

3. **Creating a Function to Train the Model and Calculating the Score:**
    A. Getting MSE, RMSE, R2-SCORE, ADJUSTED-R2 SCORE for different models used.
    B. Assigning the dependent and independent variables.
    C. Splitting the model into train and test sets.
    D. Fitting linear regression on train set.

HYPERPARAMETER TUNING:

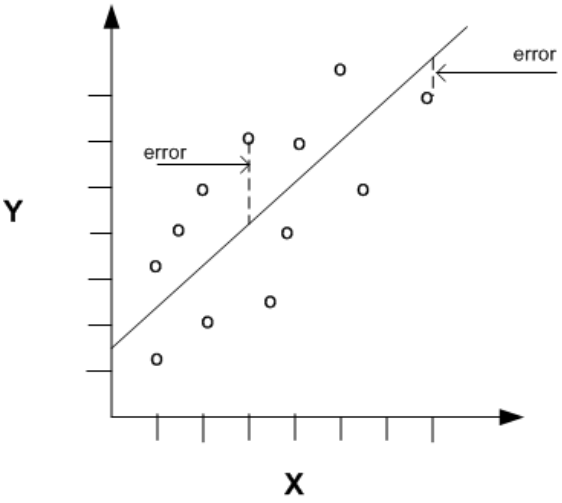
In machine learning, hyperparameter optimization or tuning is the problem of

choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

MODELS USED IN OUR PROJECT

LINEAR REGRESSION:

Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line.

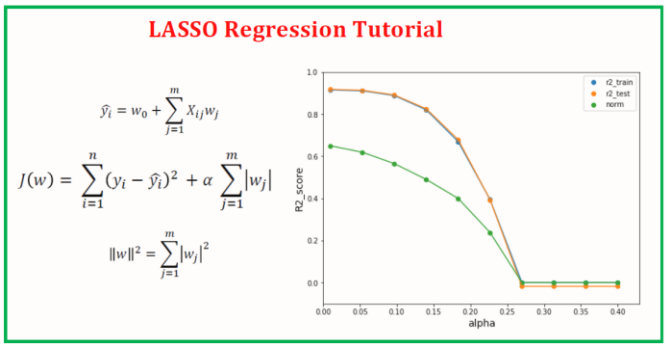


LASSO REGRESSION:

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction.

This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso Regression uses L1 regularization technique. It is used when we have more features because it automatically performs feature selection.



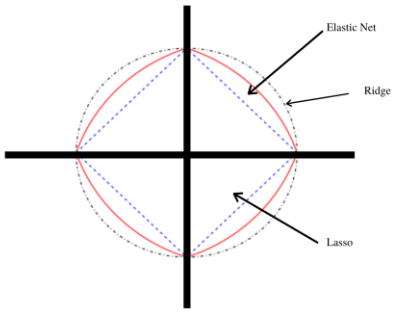
RIDGE REGRESSION:

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable.

Cost Function (J) = $\frac{1}{n}\sum_{i=0}^{n}(h_\theta(x^i) - y^i)^2$

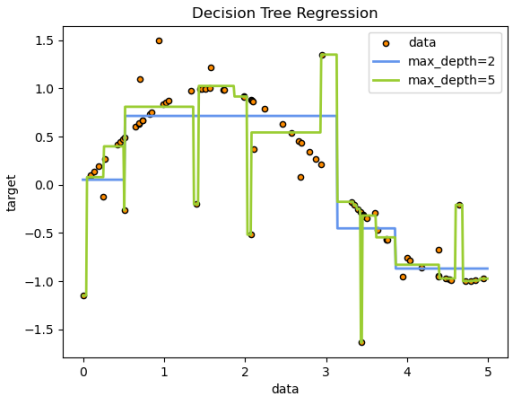
ELASTIC NET REGRESSION:

Elastic net is a penalized linear regression model that includes both the L1 and L2 penalties during training. Using the terminology from "The Elements of Statistical Learning," a hyper parameter "alpha" is provided to assign how much weight is given to each of the L1 and L2 penalties.



DECISION TREE REGRESSION:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.



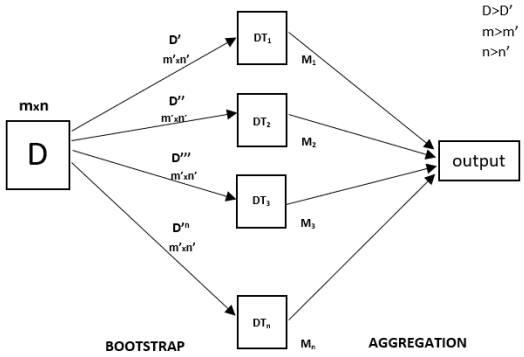
RANDOM FOREST:

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

We need to approach the Random Forest regression technique like any other machine learning technique

- Design a specific question or data and get the source to determine the required data.
- Make sure the data is in an accessible format else convert it to the required format.

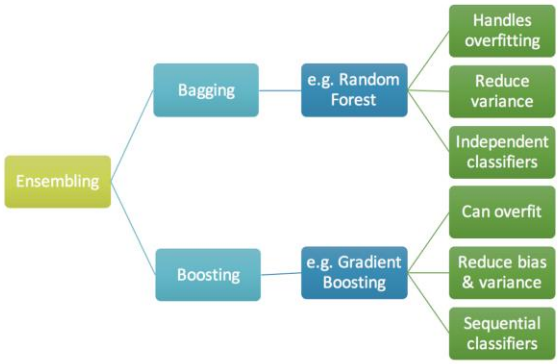
- Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
- Create a machine learning model
- Set the baseline model that you want to achieve
- Train the data machine learning model.
- Provide an insight into the model with test data
- Now compare the performance metrics of both the test data and the predicted data from the model.
- If it doesn't satisfy your expectations, you can try improving your model accordingly or dating your data, or using another data modeling technique.
- At this stage, you interpret the data you have gained and report accordingly.



GRADIENT BOOSTING:

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e., Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.

Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier). When it is used as a regressor, the cost function is Mean Square Error (MSE) and when it is used as a classifier then the cost function is Log loss.



eXtreme Gradient Boosting:

Extreme Gradient Boosting, or XGBoost for short is an efficient open-source implementation of the gradient boosting algorithm. As such, XGBoost is an algorithm, an open-source project, and a Python library.

It was initially developed by Tiangi Chen and was described by Chen and Carlos Guestrin in their 2016 paper titled "XGBoost: A Scalable Tree Boosting System."

It is designed to be both computationally efficient (e.g., fast to execute) and highly effective, perhaps more effective than other open-source implementations.

CONCLUSION:

In the comparison between "hour vs rented count of bikes" we can clearly notice a high demand in the rush hour of 8:00 am to 9:00 pm.

In the comparison between "holiday-non holiday vs rented count of bikes" we get the notion of high demand of bikes during non-holiday i.e., working days compared to holidays i.e., non-working days

Demand of Rented Bike gradually decreases with increase in rainfall.

Same pattern of decrease in demand is observed with the increase in snowfall.

In the Winter Months the Demand decreases i.e, December, January, February, however demand spikes up in summer months i.e May, June, July.

Modeling Conclusions We used 8 Regression Models to predict the bike rental count at any hour of the day - 'Linear','Lasso','Ridge','Elasticnet','Decision_Tree','Random_Forest','Gradient_Boosting','Xtreme_GB'. Using the predictions made by these level 1 individual models as features, we trained 4 level 2 stacking algorithms (Linear Regression, Random Forest Gradient Boost and Xtreme Gradient Boosting) to make more refined predictions. Below is a summary of the model performances

Of all the models, we found a simple XGBoost Model providing the best/lowest RMSE score and the adjusted_r2 of 99% which made the model deployable.