Mercedes-Benz Greener Manufacturing Project

1. Data exploration

On this section I imported the ‘train.zip’ and ‘test.zip’ data sets and explored several features of the data. For instance, the shape of the data sets, the data types of the columns, the basic statistical information of the data with the ‘describe’ method. I also eliminated the ID columns here, as well as created the ‘target’ data frame from the train data, which is what we want to predict on the test data.

2. Data wrangling

On this section I started by verifying if there were any null data entries, and there weren’t. After I checked which were the unique values in the data frame. These were mostly categorical and from the first 8 columns, with the exception of only 0’s and 1’s in the remaining columns. Using LabelEncoder I transformed all categorical data into numerical data. Then I removed all columns that had zero variance in both the training and test data. I proceed by verifying the correlation matrix of the training data and removing highly correlated columns from both the training and test data. Afterwards I removed the outlier value rows for each of the original categorical data columns, the first eight, and the corresponding rows on the target data frame.

3. Validation of XGBoost model

To determine which parameters to use in dimensionality reduction and XGBoost I created a function allowing for the variance of several parameters: number of components for the PCA with respect to the percentage of variance captured, the tree depth and the number of estimators for the XGBoost. Then I used a script to run along several values of these parameters to determine which would give the lowest root mean squared error and simultaneously the biggest R2 score.

4. Application of the model on 'test.zip' data

Using the parameters determined on the previous step I used dimensionality reduction with PCA on the test data and the XGBoost model on the whole training data to determine the prediction for the test data.