Mercedes-BenZ Greener Manufacturing

Submitted By:

Punit Bhalla (pxb173830)

Rajat Koti (rxk174330)

Spandan Dey (sxd174230)

Final Project – Machine Learning (CS 6375.003)

The University of Texas at Dallas

Spring 2018

Introduction

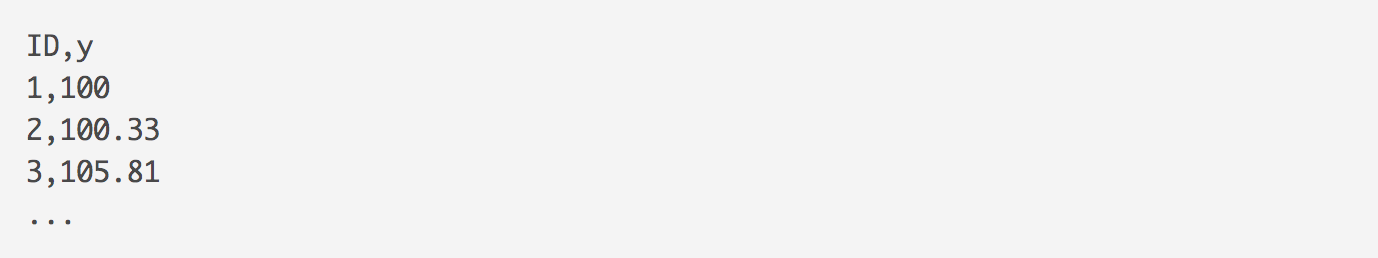
Mercedes-Benz Greener Manufacturing is basically a challenge taken up from kaggle.

Daimler’s Mercedes-Benz cars are leaders in the premium car industry. To ensure the safety and reliability of each and every unique car configuration before they hit the road, Daimler’s engineers have developed a robust testing system. The primary concern of this challenge is to reduce the time that car spends on the test bench. As one of the world’s biggest manufacturers of premium cars, safety and efficiency are paramount on Daimler’s production lines.

Dataset DESCRIPTION

|  |  |
| --- | --- |
| **Data Set** | Multivariate |
| **Default Task** | Regression |
| **Attributes Task** | Categorical, Real, Integer |
| **# Instances** | 4209 |
| **# Attributes** | 377 |
| **Missing Values** | No |

This data set consists of 377 anonymous attributes, consisting of categorical, real and integral values. For each 'ID' in the test set, task is to predict the label (‘y’) i.e. time a car spends on the test bench, given car’s features. The output should contain a header and have the following format:



PRE-PROCESSING TECHNIQUES

Following steps have been taken in order to pre-process the dataset:

1. We checked if there are any missing values (i.e. ‘?’ or ‘NaN’) in both train and test datasets. Fortunately, both the datasets do not comprise of any missing values.
2. We looked around the datatypes for each of the features (attributes) provided in the datasets. We found there are 8 columns ('X0', 'X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X8') that comprise of categorical data.
3. Out of those 8 columns, we found that three columns ('X0', 'X2', 'X5') in test dataset hold in total 16 different categorical values when compared with train dataset.
4. In order to convert the categorical data into numerical values, we first appended the test dataset into train dataset and then encoded those categorical values.
5. After, converting the categorical dataset to numerical values, we split the test data (that was earlier appended to the train dataset) from the train dataset.
6. Now, we checked if any feature(s) in the train dataset has constant values for every instance. We found there were in total 13 features for which values are constant across all the instances. Therefore, we removed those features from both the datasets.
7. Finally, we standardized both the datasets using Scikit’s standard scaler method.

ALGORITHMS USED

We are going to use following models in order to predict the label (‘y’) for the test dataset:

1. SVR (Support Vector Regression)
2. MLP (Multi-layer Perceptron Regression)
3. AdaBoost Regressor
4. Gradient Boosting Regressor