MERCEDES-BENZ GREENER MANUFACTURING

Safety and reliability testing is a crucial step in the automobile manufacturing process. Every new vehicle design must pass a thorough evaluation before it enters the consumer market. Testing can be time-consuming and cost-intensive as a full check of vehicle systems requires subjecting the car to all situations it will encounter in its intended use. Predicting the overall time for a vehicle to pass testing is difficult because each model requires a different test stand configuration. Every possible vehicle combination must undergo the same rigorous testing to ensure the vehicle is robust enough to keep occupants safe and withstand the rigors of daily use. The large array of options offered by Mercedes means a large number of tests for the company's engineers to conduct. More tests result in more time spent on the test stand, increasing costs for Mercedes and generating carbon dioxide, a polluting greenhouse gas. The stated goal of the competition is to reduce the time vehicles spend on the test stand which consequently will decrease carbon dioxide emissions associated with the testing procedure. Although the reduction in carbon dioxide may not be noteworthy on a global scale,4 improvements to Mercedes's process can be passed along to other automakers or even to other industries which could result in a significant decrease in carbon dioxide emissions.

Problem Statement

The objective of the Mercedes-Benz Greener Manufacturing competition is to develop a machine learning model that can accurately predict the time a car will spend on the test bench based on the vehicle configuration. The vehicle configuration is defined as the set of customization options and features selected for the particular vehicle. The motivation behind the problem is that an accurate model will be able to reduce the total time spent testing vehicles by allowing cars with similar testing configurations to be run successively. This problem is an example of a machine learning regression task because it requires predicting a continuous target variable based on one or more explanatory variables.

Analysis

Two data files are provided by Mercedes-Benz for use in the competition: a training dataset, and a testing dataset. Both files are provided in the comma separated value (CSV) format. The training and testing data both contain 4209 vehicle tests obtained by Mercedes for a range of vehicle configurations. The training data also contains the target variable, or testing duration in seconds, for each vehicle test. No target is provided for the testing data as the testing durations are known only by Mercedes and are used to determine the winner of the competition. Each vehicle test is defined by the vehicle configuration, which is encoded in a set of features. Both the training and the testing dataset contain 376 different vehicle features with names such as 'X0', 'X1', and 'X2' and so on. All of the features have been anonymized meaning that they do not come with a physical representation. The description of the data does indicate that the vehicle features are configuration options such as suspension setting, adaptive cruise control, all-wheel drive, and a number of different options that together define a car model. There are 8 categorical features, with values encoded as strings such as 'a', 'b', 'c', etc. The other 368 features are binary, meaning they either have a value of 1 or 0. Each vehicle test has also been assigned an ID which was not treated as a feature for this analysis.

In order for the machine learning model to process the categorical variables, they must be one-hot encoded. This means that the unique categories contained in the categorical variable are transformed into a set of new variables. Each instance is assigned a one for the new variable corresponding to its original categorical variable and a zero in all other new variables. After one-hot encoding and alignment of the training and testing data there were a total of 553 binary features in both datasets.