**Mercedes-Benz Greener Manufacturing**

DESCRIPTION

Reduce the time a Mercedes-Benz spends on the test bench.

Problem Statement Scenario:  
Since the first automobile, the Benz Patent Motor Car in 1886, Mercedes-Benz has stood for important automotive innovations. These include the passenger safety cell with a crumple zone, the airbag, and intelligent assistance systems. Mercedes-Benz applies for nearly 2000 patents per year, making the brand the European leader among premium carmakers. Mercedes-Benz is the leader in the premium car industry. With a huge selection of features and options, customers can choose the customized Mercedes-Benz of their dreams.

To ensure the safety and reliability of every unique car configuration before they hit the road, the company’s engineers have developed a robust testing system. As one of the world’s biggest manufacturers of premium cars, safety and efficiency are paramount on Mercedes-Benz’s production lines. However, optimizing the speed of their testing system for many possible feature combinations is complex and time-consuming without a powerful algorithmic approach.

You are required to reduce the time that cars spend on the test bench. Others will work with a dataset representing different permutations of features in a Mercedes-Benz car to predict the time it takes to pass testing. Optimal algorithms will contribute to faster testing, resulting in lower carbon dioxide emissions without reducing Mercedes-Benz’s standards.

Following actions should be performed:

* If for any column(s), the variance is equal to zero, then you need to remove those variable(s).
* Check for null and unique values for test and train sets.
* Apply label encoder.
* Perform dimensionality reduction.
* Predict your test\_df values using XGBoost.

Find the datasets [here.](https://github.com/Simplilearn-Edu/Machine-Learning--Projects)

<https://medium.com/@williamkoehrsen/capstone-project-mercedes-benz-greener-manufacturing-competition-4798153e2476>

**Data Exploration**

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 and are available for download on the [Kaggle competition data page](https://www.kaggle.com/c/mercedes-benz-greener-manufacturing/data). [11] 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’, ‘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.

Dear Kuntal  
  
PCA part is working appropirately  
  
the columns can be sorted to be passed to the regression analysis.  
  
I would recommend to see the multicollinearity present in the database using technique like heat map  
  
and then you may apply Ridge/Lass with XGBoost