**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.

The above given is the description of the data that was provided to complete the project. The dataset had 2 zip files of csv format. The first one had the train data in it and was also named the same. Similarly, the second file had the test data.

The train data consists of almost 378 inputs and one output. The inputs are the names of the models of the cars that are produced by Mercedes-Benz and the output is the time calculated by the testing systems for optimization. For the given task we were required to find the variance of all the inputs in the data and eliminate the columns whose variance is almost zero. On calculation it was found that almost 12 columns of the data set had variance equal to zero so that were immediately removed and then the total inputs of the dataframe were 365 as the ID column was also dropped due to its insignificant nature. Further normal values were found in the data set but almost 8 columns in the set were having categorical values so label encoding was done on those columns as mentioned in the task above. To make the data set easier and more convenient to understand MinMaxScaler was also applied on the datasets which had more varied values. Further the data set was split into a train test and evaluation sets.

The test data consists of almost 377 inputs that are used to predict the future values. The test data had five columns that had variance near 0 so that were also removed in order to refactor the data. Three towers also check for null and unique values and no no values are found in their data set bird on observation heat columns like that of the train data set were also categorical you do this level and coding was also applied on the test set. Similarly, MinMaxScaler has also applied to make the data easily readable.

The train datasets 5 column that is the output column was also plotted in a graph to understand its nature all the outliers for the same column were also found weatherproof box plots embedded with the seaborn library. Only one value of the total values in the column was extremely out of the normal range of data set hence that was removed and as it was considered as an extreme outlier. A number of other outliers were also observed with the help of the boxplot after the removal of the extremist outlier but they were not removed from the data set as the number of outliers was large and their removal could result in overfitting of for model.

Further a principal component analysis was used to reduce the dimensions of the data and make it more approachable depending on the variance coefficients found with the help of PCA. After this the train set which is denoted by X in the notebook was fitted and transformed. This in result give two new sets of X\_train and X\_valuation with reduced dimensions.

The newly formed reduced dimensioned sets are used in the XG boost model for predictions. The root mean square value found for the test and the predicted datasets is in the range of 8 – 9. And this is how the project was completed and all the tasks listed above for the project were executed