Mercedes-Benz Greener Manufacturing.

Project 1

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.

**Following actions are actually performed:**

1. **Data Exploration**

Learned that both training and test dataset mainly consist of 368 binary features and 8 categorical features, out of which there are 12 binary features that only contain a single value.

More importantly, for there 8 categorical features, there are 24 test records that contain values that are not present in training dataset as pointed in **section 1.7**. This also lays the foundation to explore alternative model to handle these records.

1. **Data Wrangling**

Removed features that have zero variance as well as the ‘ID’ and check that the datasets contain no null value, followed by label encoding of the categorical features.

Dimension reduction is performed by using PCA.

1. **Model Training and Prediction**

A model is trained by utilizing XGBoost with reduced-dimensions data and used to predict the test dataset in **section 3.1**.

As PCA limits the ability to understand which are the key features to the trained model, an alternative model is also trained by using the original features in **section 3.2**.

Bearing in mind the issue pointed in **section 1.7**, one more model is trained by removed features that the test data set having values not present in train dataset. This will be jointly used to predict the test dataset in full.