Nadiem Ahmed October 1, 2019

Project Deliverable 1

1. Dataset

The dataset I chose is from a Kaggle competition called "Mercedes-Benz Greener Manufacturing". As it is part of a competition, the objective is already set and I will aim to improve upon others' scores. The goal of the competition is to accurately predict the time it takes for Mercedes to test a car given a set of custom features that the car will have.

2. Methodology

i. Data preprocessing

The dataset consists of anonymized variables that represent custom features Mercedes would put in a car (4WD, added air suspension,etc.), and the target variable is the time it takes to test the car. For the anonymized variables, a few are alphabetical (categorical) while the majority are binary. The data has already been split into train and test sets.

ii. Machine learning model

Since the target variable is continuous, I think it is appropriate to use a regression model. I am not sure whether a linear or polynomial regression will work better, but I have a feeling polynomial regression will be better.

iii. Final conceptualization

I plan on presenting a poster. In the competition leaderboard, the highest R^2 value is .5555 and the average is approximately .55. I will try to get an R^2 value higher than .55.