Abstract

Team members:
Kai Mo (kam455) Tianqi Xie (tix19) Li Shi (lis112)

1. Data source

The data we use is “APS Failure at Scania Trucks Data Set”, which is sourced from UCI Machine Learning Repository.
Link to data:
https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks

2. Details

The dataset's positive class consists of component failures for a specific component of the APS system. The negative class consists of trucks with failures for components not related to the APS. The problem we are trying to solve is to classify whether APS give false or true alarm. The dataset contains 171 attributes and all of them are anonymous, which significantly increases the difficulty of interpreting the dataset.

3. Our methods

As it is a binary classification problem, we chose to apply classification models to it, such as logistic regression model, tree-based models, and support vector machine model.

4. Hypothesis

Unbalanced dataset. If we don’t oversample or downsample the dataset, we might overestimate/underestimate the accuracy. For instance, we have 59,000 positive samples and only 1,000 negative samples, a dummy classifier could achieve over 98% accuracy, so it is hard to justify whether the trained model performs well.

5. Why interesting

With technology developing, more and more cars are used on the road. So driving safety is essential to us. Using data mining methods, we can know more about if the APS alarm is reporting correctly in a certain condition. Furthermore, the result can be applied to the real industry to improve the driving service experience as well as automatic driving.