Machine Learning Engineer Nanodegree

Capstone Proposal

Hajime Kawata
February 26st, 2019

Proposal

Predict failure of heavy duty trucks out of sensors data

Domain Background

Reducing the failure rate of trucks in logictics is an important issue for improving profitability and customer satisfaction in logistics business. Until now, logistics companies rely on regular maintenance to prevent the occurrence of failures. However, since the utilization rate of trucks exceeds the days what has been assumed to be covered by periodic maintenance, the percentage of post-maintenance caused by failures increasing, which has been a problem for some time.

For this reason, it is strongly required that inspection and maintenance work be performed before regular failure, by appropriately detecting signs of failure during regular maintenance In order to properly capture signs of failure, it has been considered to use data collected from sensors installed in trucks, bu sensor data is gathered in milliseconds at longest, and collection targets are also diverse. As a result, several hundred megabytes of data may be generated from one track on a day. A large data cost has been anticipated.

In this project in order to quickly and flexibly arrange trucks based on failure prediction,

- Reduce the amount of features derived from the target sensor to be collected and suppress the arithmetic cost to be generated
- Perform a failure based on the logic derived from the data pattern leading to the failure on the truck side.

This would make it possible to appropriately send alerts before the actual failure occurrence to the monitoring center and to adjust the arrangement of the trucks.

Problem Statement

Here, using the training data, we derive the prediction model of the occurrence of the failure (label) from the sensor data (features). This prediction logic is applied to the test data and the ability of the prediction logic is evaluated. The sensor data is anonymized for confidentiality reasons, and so we have to make the prediction model purely on mathematical and statistical approach.

Datasets and Inputs

Data is provided in : *Kaggle* : https://www.kaggle.com/uciml/aps-failure-at-scania-trucks-data-set/home

The dataset consists of data collected from heavy Scania trucks in everyday usage. The system in focus is the Air Pressure system (APS) which generates pressurized air that is utilized in various functions in a truck, such as braking and gear changes. (*Kaggle*)

Category	Description
positive class	*consists of component failures for a specific component of the APS system.*
negative class	*consists of trucks with failures for components not related to the APS.*

Following is the data volume information.

_	Total	Positive	Negative	features
Train	60000	1000	59000	171
Test	16000			171

Solution Statement

For the criteria listed in **Domain Background***, take the following solution approach in this project

- Reduce sensor data features from 171 by PCA (Primary Components Analysis)
- Construct the model from reduced features with XGBoost classifier, by adjusting hyperparameters with Grid Search CV

Benchmark Model

As benchmark, adopts 60% correct answer rate. This is based on a hearing from an interview myself conducted that the failure rate prevented by regular maintenance is about 60% by experience. It is worth considering the devided model in this project, if failure prediction is more than 60%.

Evaluation Metrics

In evaluating the prediction logic, it is a challenge to extract failure events as much as possible while maintaining the operation rate. For that reason, it is necessary to accurately derive the judgment of correctness. Here, to make the balance of Confusion Matrix, adopt a model with a high result of scoring by **ROC-AUC**.

Project Design

1. Split the training data set to establish model and evaluate
2. Apply PCA to reduce the dimension of the model.
3. Device predictor with XGBoost classifier, while seeking the best hyperparameter values with Grid Search by ROC-AUC.
4. Evaluate with the model with given test set, with confusion matrix, and accuracy, precision, and f1 scores.
5. Using the given test set to evaluate the model
6. Evaluate the contribution of feature, to discuss the possibility of reducing the sensors to predict failures.