Intelligent Data Analysis Industrial Challenge 2024 - Developing an Effective Predictive Model for Imminent Component X Failures in Heavy-Duty Scania Trucks

The goal is to develop a predictive model (or models) that can accurately predict whether a specific engine component (hereinafter referred to as 'Component X') in a vehicle is at risk of imminent failure. The dataset includes information gathered from Scania trucks used in heavy-duty applications.

The contest:

Welcome as a participant in the IDA 2024 Industrial Challenge. The task is to come up with a good prediction model (or models) for judging whether or not a vehicle faces imminent failure of the specific engine Component X or not. The contestants will need to submit an csv file named 'IDA_Industrial_challenge_2024_predictions.csv', with 5045 number of predictions (with values in {0,1,2,3,4}), i.e., one prediction (at the last readout) for each vehicle in the 'test_operational_readouts.csv' file. This file should be sent as an attachment to tony@dsv.su.se, with the e-mail subject being 'IDA 2024 Industrial Challenge Submission' using a CSV file in the following format:

vehicle_id, predicted_class

All competitors must also write a paper describing the methods they used when creating their predictive model(s). It is encouraged that the paper is written in a clear and inspirational style in which the authors motivate their decision(s) when developing their model(s). The papers should use the same template as the other papers in the conference, see (link) for more details. This paper should also be attached to the contest submission.

The timeline to participate in the contest is available on the IDA 2024 Industrial Challenge website (link).The top three contestants will be included in the IDA conference proceedings and they will also present their methodology and results at the conference in the industrial challenge session. Scania has generously provided prize money to the top three contenders, where the first-placed contender/team will receive an amount of 5000 SEK, the second-place contender/team will receive an amount of 3000 SEK, and the third-place contender/team will receive an amount of 2000 SEK.

Training Data:

Disclaimer - repair frequencies and readout frequencies may have been modified and are not necessarily representative for actual truck usage.

The data for the challenge comprises of multiple files, each containing different types of information. Where the operational training-, validation- and test-data have identical information content layout. Hence, the content explanation for the train readout files is also applicable to validation readout files and test readout files, but the observation sampling is different and explained in section *Validation and Test data*.

The file 'train_operational_readouts.csv' includes operational data, which consists of sporadic and unevenly sampled time series of readouts from each vehicle (identified by the **vehicle_id** column). Each row corresponds to a unique operational data readout, and ***time_step*** indicates the time when the readout happened in relation to the specific vehicle's start of operation. In other words, ***time_step*** represents the number of time units the vehicle and Component X have been in use. The data consists of a subset of all available operational data, selected by experts. There are 14 different variables, and each variable can have one or multiple bins described

by an index. The naming convention for the variables is ***variableid_binindex***, e.g., a 1-dimensional histogram would have only one index “1234_0”, while a multi-dimensional histogram (with 3 bins) would have three indices, “2345_0”, “2345_1”, and “2345_2”.

The file with the name ‘train_tte.csv’ (time to event (tte)) includes repair records for each vehicle (identified by the **vehicle_id** column), providing details on when repairs were conducted on Component X (if any) during the study period. If ***in_study_repair*** is set to 1, it means that Component X was repaired (for the first time) ***length_of_study_time_step*** time units after the start of operation of the vehicle with Component X (factory mounted). On the other hand, if ***in_study_repair*** is set to 0, no repair event occurred during the first ***length_of_study_time_step*** time units of operation.

The file ‘*_specifications.csv’ contains information about the specifications of the vehicles, such as their engine type and wheel configuration. The specification data also represent a subset of all available data, selected by experts.

Validation and Test data:

‘train_operational_readouts.csv’ data contain all information within the study time, while the readout data for validation and test have been selected such that, at random, a readout for a vehicle has been selected as the *last* readout. Hence, they do *not* contain all information about a vehicle’s life. This is done to simulate the usage of a prediction model in a real-world setting, when the model only has information about a vehicle up until the present time.

Below is an illustration of the difference between train and validation/test sets. For the former cases, all information about the vehicles is available, but in the latter cases, the third readout has been selected (at random), illustrated by the green arrow. Hence the information available in this case is only the three leftmost readouts (excluding the fourth and final readout):

Figure 1: In train data the complete sequence of readouts up until the end of the study is included, while in the validation and test data the last readout has been selected at random, indicated by the green arrow.

Finally, the file ‘validation_labels.csv’ have two columns where the first contain information about **vehicle_id** while the second named ***class_label*** which corresponds to the class for the last readout, how these classes should be interpreted is described in the section *Evaluation* below.

For the test set on the other hand, there are no labels available since this is the set for which you are expected to provide the labels. The test and validation sets include the same operational data and specification details as the training set for the vehicles.

Attributes:

The attribute names of the operational data have been anonymized for proprietary reasons. It consists of both single numerical counters and histograms consisting of bins with different conditions. The histograms have open-ended conditions at each end. For example, if we are measuring the ambient temperature "T," then the histogram could be defined with 4 bins where bin 1 collects values for temperature T < -20, bin 2 collects values for temperature T >= -20 and T < 0, bin 3 collects values for temperature T >= 0 and T < 20, and bin 4 collects values for temperature T > 20. The attributes describing the specifications is all categorical.

Evaluation:

The classes in the class column in the training and validation data have been created by a time window before a failure event of Component X. Readouts within a time window of 48 to 24 time units before the failure of Component X belongs to Class = 1. Readouts within a time window of 24 to 12 time units before failure are labeled as Class = 2. Class = 3 and Class = 4, have time windows of 12 to 6 time units and 6 to 0 time units, respectively. The performance will be evaluated using a cost metric of miss-classification, shown in the table below.

	Predicted: 0	Predicted: 1	Predicted: 2	Predicted: 3	Predicted: 4
Actual: 0		Cost_0_1=7	Cost_0_2=8	Cost_0_3=9	Cost_0_4=10
Actual: 1	Cost_1_0=200		Cost_1_2=7	Cost_1_3=8	Cost_1_4=9
Actual: 2	Cost_2_0=300	Cost_2_1=200		Cost_2_3=7	Cost_2_4=8
Actual: 3	Cost_3_0=400	Cost_3_1=300	Cost_3_2=200		Cost_3_4=7
Actual: 4	Cost_4_0=500	Cost_4_1=400	Cost_4_2=300	Cost_4_3=200	

The total cost of a prediction model is the sum of the different “Cost_n_m” multiplied by the number of instances, resulting in a summarized cost (Total_cost). Cost_0_{1,2,3,4} refers to the cost of an unnecessary check being done by a mechanic at a workshop, while Cost_{1,2,3,4}_m refers to the cost of potentially missing a faulty truck or signaling an alarm too late, which may cause a breakdown or expensive replanning of transport mission for the customer. The total cost is calculated as Total_cost = Cost_n_m * No_Instances, where n, m $\in$ {0,1,2,3,4}.

The goal is to minimize the ‘Total_cost’, and the winner is the submission with the lowest ‘Total_cost’. Note, that one person is only allowed to participate in one submission. Hence, only one submission is allowed per team/constellation of members. Also, note that the submission should contain exactly one prediction for each vehicle, available in the test_operational_readouts.csv file, i.e., information up until the last (randomly selected) readout for each vehicle.

Good luck!

Tony Lindgren

2024 IDA Industrial Challenge Chair