Cool Train
INFOH423 Data Mining Project 2023/24

SNCB: Georges Tod, Pieter Moelans, Christophe Vanderelst
ULB: Mahmoud SAKR, Raphaël GYORI
georges.tod@belgiantrain.be, pieter.moelans@belgiantrain.be, christophe.vanderelst@belgiantrain.be, mahmoud.sakr@ulb.be, raphael.gyory@ulb.be

The National Railway Company of Belgium SNCB is responsible for organizing and operating the rail service in Belgium. Within SNCB, the Technics directorate is tasked with acquiring, updating, and maintaining the rolling stock. Their objective is to deliver a superior service by ensuring the availability of an ample supply of safe, dependable, and comfortable rolling stock tailored to meet both operational and commercial requirements of customers[1]. This project is sponsored by the rolling stock team, who provide the data and the technical challenge.

The context of this challenge is the cooling of diesel trains[2]. SNCB's diesel trains are AR41 (Belgian Railways Class 41 - Wikipedia) and consist of 2 vehicles which are always coupled. Each of these vehicles is referenced in the data: some_variable_name_PC1 and some_variable_name_PC2. Each PC is equipped with 2 cooling systems: one for the engine air intake and one for the oil of the transmission. In the latter, the water is the coolant (water-oil intercooler) while in the former air is used as coolant (air-air intercooler).

Maintaining these systems working properly is important for the health of the diesel engines and transmission and to avoid incidents. In operations, if one of the engines fails, then the other can serve as a backup. However, if the backup fails then there will be a train delay. It is of everybody's interest to keep both engines working properly and making sure the temperatures and pressures stay within acceptable ranges.

[1] https://www.belgiantrain.be/en/about-sncb/enterprise/management-structure/directions/technics
[2] 5% of SNCB trains are diesel-powered, 95% electric.

The maximum acceptable temperatures are for the: air: 65°C, water: 100°C, oil: 115°C. Above those thresholds the engines are stopped automatically to avoid damage.

Just like in life, problems arise. Here the anomalies could come from: a sensor, a cooling system, an engine, both engines, or external sources: weather is too hot, pollen gets stuck into the radiators, etc. Being able to capture and to distinguish some of these potential failures and to predict them in advance will help preserve vehicles' health and prevent train incidents.

For this challenge, you will receive a 2GB csv file containing real life timeseries from the diesel trains. The data contains the timestamp of each sample for both PC1 and PC2. Beware the sampling time is not uniform, and there might be some duplicates in the data. The data contains real life errors: GPS positions reporting zeros or temperatures at exactly zero, etc. The temperatures are in °C, the pressures in kPa. Every sample reports:

- the temperatures from the two redundant engine cooling systems
- the RPMs of the engines. We are interested in issues when the train is in operation: if both engines are turned off this is less interesting. If a single engine is off, then there is probably an issue with the one that is off - as we typically try to have both of them operating at the same time
- gps locations of the vehicles

The temporal extent of the data covers from January 2023 till September 2023, so the effect of weather conditions can be explored. The data is raw so some processing will be necessary before any analysis. An excerpt of the file is as follows (zoom-in to see):

	mapped_veh_id	timestamps_UTC	lat	lon	RS_E_InAirTemp_PC1	RS_E_InAirTemp_PC2	RS_E_OilPress_PC1	RS_E_OilPress_PC2	RS_E_RPM_PC1	RS_E_RPM_PC2	RS_E_WatTemp_PC1	RS_E_WatTemp_PC2	RS_T_OilTemp_PC1	RS_T_OilTemp_PC2
0	181	2023-08-01 3:44:12	50.7698183	3.8721144	27	23	255	238	794	801	83	81	76	77
1	143	2023-08-01 6:36:29	51.0399934	3.6934285	33	32	272	324	802	804	78	78	73	74
2	183	2023-08-24 6:53:54	50.7422026	3.6020347	31	33	234	182	799	802	82	82	85	87
3	177	2023-08-01 13:53:38	50.9309143	5.3271318	35	38	220	244	794	801	77	81	78	82
4	143	2023-08-24 7:02:30	51.1807725	3.5752586	41	34	227	282	806	800	85	78	82	79

Your team is challenged to analyze the given data, and develop a robust method to report anomalies. Your method needs to detect anomalies and to distinguish them, e.g., (1) noise, (2) deviations in only one of the two cooling systems, which may indicate a problem in one set of sensors, (3) anomalies in the two cooling systems which may indicate a problem in the train engine, etc. Obviously, not everything can be predicted and/or explained.

This overall challenge is broken in the following tasks:

1. Data loading, exploration and preprocessing
2. Enrichment with weather data. One possible source is https://openweathermap.org/, but you may as well use other available sources
3. Developing and comparing multiple anomaly detection methods
4. Now as you have anomalies detected, it doesn't help much to list them in a table. Instead, you should develop a dashboard for providing a comprehensive view of anomalies in relation to dimensions like weather, time, location, etc. The dashboard should help the SNCB rolling stock team to make sense of the presented anomalies. The role of the dashboard is to help explain the found anomalies by spotting visual patterns. For instance, the dashboard may help visualize that many anomalies occur in certain regions, weather conditions, etc.
5. Bonus Task: Deployment in streaming mode. It should be possible to deploy your method to run in a streaming mode. That is, while here you are given a historical data file, your method should accept a continuous stream of new observations, and report anomalies on the fly.

Deliverables

You should deliver a presentation of your work and results (as a .pdf) containing the following elements:

1. A cover page with the list of group members, including student ID
2. A description of dataset loading and preprocessing
3. A description of your data exploration activity; better accompanied with statistics, figures, screenshots, etc

4. A clear presentation of your anomaly detection methods, and how they compare. Special attention should be given to the correctness of the results
5. A live demo of the dashboard, optimally accompanied with a scenario that shows some anomalies and the steps to use the dashboard to understand them.

Evaluation

The evaluation jury consists of the two course instructors. SNCB representatives may also attend, depending on their availability. You will be asked to present your solution and defend it. The grading will consider the following factors:

- Your data management: loading, integration, exploration (10 points)
- The choice and parameterization of the relevant data mining methods (10 points)
- The evaluation and comparison of the proposed analyses (10 points)
- Your interpretation of the found anomalies. Think of answers to questions like: how far are your results accurate? have you cross checked your results with some ground truth? What would limit the use of your solution in SNCB (10 points)
- The bonus task (5 points)

Good luck and have fun !