ULB UNIVERSITÉ LIBRE DE BRUXELLES

Anomaly detection SNCB

Data Mining project

Simon Coessens, Md Kamrul Islam, Narmina Mahmudova, José Carlos Lozano

Data mining
École Polytechnique de Bruxelles
Université Libre de Bruxelles

December 17, 2023

Overview

1. Introduction
2. Data Preprocessing
3. Exploratory Data Analysis
4. Data Enrichment
5. Research Questions
6. Anomaly Detection Methods
7. Dashboard Development
8. Streaming Dashboard

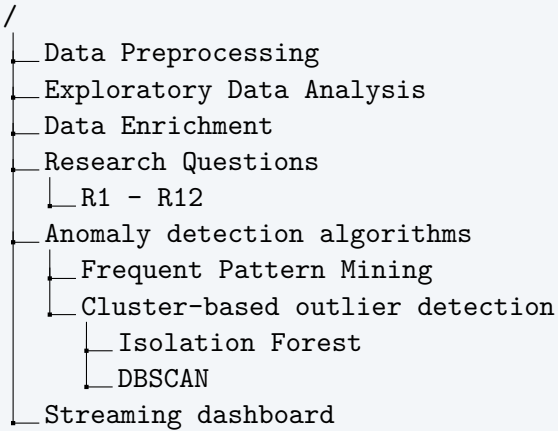
Introduction

Introduction to the project

- Objective: Identifying Anomalies within the Train Data Set
- **Anomalies in our project are found in three ways:**
  - Anomalies found in the Data Preprocessing **A1**
  - Anomalies found through research questions using domain knowledge **A2**
  - Anomalies found using more complex data mining algorithms **A3**

Belgian Railways Class 41

Project structure



Data Preprocessing

Data organization

- Initially, there were difficulties in handling the CSV file.
- We made the choice to have all our data in a Postgres database
- This greatly improved the speed and ease of use of our workflow
- The locations are stored as PostGIS datatypes
- We retrieve the data from the database out of our Python notebook files

Data Cleaning

- Removed NULL values.
- No duplicate rows found.
- Filtered out data based on Time Interval.
- Removed irrelevant Geo-coordinates.
- Converted `timestamps_UTC` to Date time.

```
Unnamed: 0                0
mapped_veh_id             0
timestamps_UTC            0
lat                       0
lon                       0
RS_E_InAirTemp_PC1        0
RS_E_InAirTemp_PC2    12726
RS_E_OilPress_PC1         0
RS_E_OilPress_PC2     12726
RS_E_RPM_PC1              0
RS_E_RPM_PC2          12726
RS_E_WatTemp_PC1          0
RS_E_WatTemp_PC2      12726
RS_T_OilTemp_PC1          0
RS_T_OilTemp_PC2      12726
dtype: int64
```

Exploratory Data Analysis

Statistical Analysis

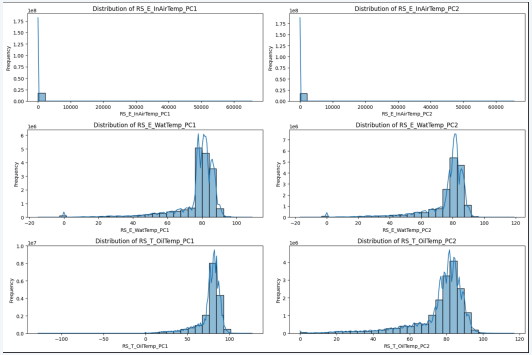
- High standard deviations in `InAirTemp` indicate data dispersion and potential anomalies.
- Extreme max `InAirTemp` values suggest sensor malfunctions or outliers.
- Negative min `WatTemp` indicates erroneous sensor readings.
- Extreme max `RPM_PC2` value indicates data entry error.

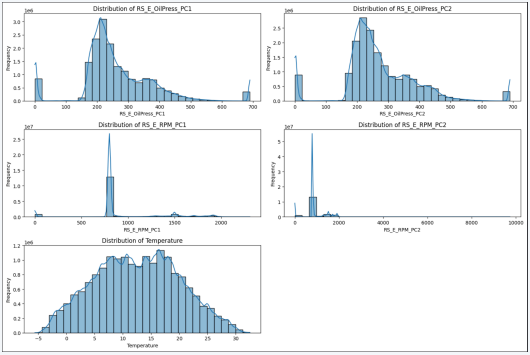
	RS_E_InAirTemp_PC1	RS_E_InAirTemp_PC2	RS_E_OilPress_PC1	\
count	17666508.00000	17666508.00000	17666508.00000	
mean	32.02987	32.33412	263.58789	
std	328.12210	347.99953	115.18088	
min	0.00000	0.00000	0.00000	
25%	22.00000	22.00000	203.00000	
50%	32.00000	33.00000	238.00000	
75%	40.00000	39.00000	320.00000	
max	65535.00000	65535.00000	690.00000	

	RS_E_OilPress_PC2	RS_E_RPM_PC1	RS_E_RPM_PC2	RS_E_WatTemp_PC1	\
count	17666508.00000	17666508.00000	17666508.00000	17666508.00000	
mean	270.68750	912.38797	907.96387	76.94352	
std	116.11669	383.30365	388.47031	13.63081	
min	0.00000	0.00000	0.00000	-15.00000	
25%	210.00000	797.00000	797.00000	77.00000	
50%	248.00000	801.00000	801.00000	81.00000	
75%	331.00000	812.00000	811.00000	84.00000	
max	690.00000	2309.00000	9732.00000	109.00000	

	RS_E_WatTemp_PC2	RS_T_OilTemp_PC1	RS_T_OilTemp_PC2
count	17666508.00000	17666508.00000	17666508.00000
mean	76.14235	76.55832	76.16183
std	14.52867	14.48119	15.35036
min	-17.00000	-128.00000	0.00000
25%	76.00000	74.00000	74.00000
50%	81.00000	81.00000	81.00000
75%	84.00000	85.00000	85.00000
max	119.00000	127.00000	117.00000

Visual Analysis



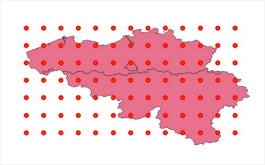


Data Enrichment

Enhancing Data with Weather Information

- We used Open Meteo for weather data
- To reduce the number of requests we pulled data only for certain grid points
-

Temperature	Humidity	Rain
14.4	95	0.9



Research Questions

Research Questions

Clarification

We set out to research some simple research questions first.

These questions are posed based on the domain knowledge that we have about the project.

Examples

- Count of temperature boundary exceedances per vehicle ID
- Mapping temperature exceedings by location for each vehicle ID
- Analyzing RPM readings with relation to temperature deviations.
- Detecting anomalies in speed readings.

Research Questions

Question 1:
Where are the locations with temperatures exceeding set boundaries?

Analysis:

- Air temperature anomalies show varying occurrence counts.
- Oil pressure issues exhibit an uneven distribution.
- Water temperature anomalies are infrequent and have sparse peaks.
- Vehicle 128 experiences frequent occurrences in all three cases.

Research Questions

Question 2:
What are the absolute occurrences of temperatures exceeding boundaries for each vehicle_id?

Analysis:

Research Questions

Question 3:
Is the engine shut down when air, water, and oil temperatures exceed the threshold?

Analysis:

- Recorded 3212 RPM values for PC1 seem inaccurate.
- RPM anomalies are prominent during summer, especially in June and May.
- Peaks observed on Mondays and Sundays.

Research Questions

Question 4:
Are there specific times of day, days of the week, or months where temperature anomalies are more frequent?

Analysis:

- Peaks occur around noon and late evening.
- Weekly distribution shows evenness, with a slight increase on Thursdays.
- Highest numbers recorded in summer; fewest observed in winter and early spring.

Research Questions

Question 5:

What is the correlation between various engine parameters, external weather conditions and anomalies in train cooling systems?

Analysis:

- Temperature anomalies show a very low correlation with external weather and engine parameter
- Indicates temperature anomalies occur independently

Research Questions

Question 6:
How do temperatures differ between internal and outside temperature sensors?

Analysis:

- Analyzing buckets, its common to have a 5 to 20 degrees difference with outside temperature
- Values with absolute difference from outside temperature PC1: 13541
- Values with absolute difference from outside temperature PC2: 17568

Research Questions

Question 7:
Are there any variations in internal air temperature readings between pairs of sensors?

”InAir” Sensors

Analysis:

- Sensors have differences that eventually balance out -> Consider persisting differences
- Around 15,000 instances with persisting unusual variances in internal air temperature sensor readings

Research Questions

Question 8:
Are there any variations in sensor readings of oil and water temperatures PC1?

Analysis:

- These values have correlation, hence the analysis.
- Again we look for persisting differences across time for every train.
- Approximately 140,000 instances characterized by an unusual consistent variance among their sensor readings for oil and water temperatures.

Anomaly Detection Methods

Frequent Pattern Mining

- Apriori and FP GROWTH
- To have a better understanding of the Data: what is occuring frequently together?
- Not an ideal method for anomaly detection

Support	Itemsets
0.6262	'80-95°C water', '25-45°C air'
0.6003	'80-95°C water', '80-95°C oil'
0.5631	'80-95°C oil', '25-45°C air'
0.5438	'80-95°C water', '80-95°C oil', '25-45°C air'
0.5300	'Medium RPM', '25-45°C air'
0.4828	'Medium RPM', '80-95°C water'
0.4642	'Medium RPM', '<80°C oil'

Clustering methods

Clarification

- Construct clusters using a certain clustering algorithm
- Find data points outside the clusters that are flagged as anomalies

Algorithms used

- DBSCAN
- Proximity-based clustering
- Kmeans

Isolation Forest for Anomaly Detection

Overview

- An unsupervised learning algorithm for anomaly detection.
- Isolates anomalies instead of profiling normal data points.
- Not memory intensive, so use full for our large data set

Observations

- We filtered out the data points that we already discussed

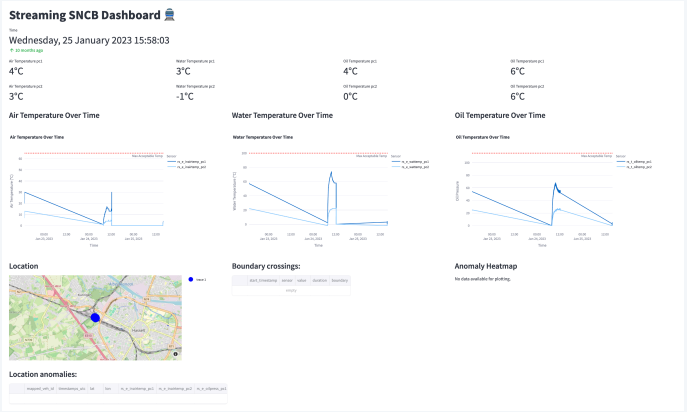
Dashboard Development

Streaming Dashboard

ULB

Dashboard

- Dashboard built by streaming the historical data
- Built using `Python` and the `Streamlit` library.



ULB UNIVERSITÉ LIBRE DE BRUXELLES

Thank you for your attention

Simon Coessens, Md Kamrul Islam, Narmina Mahmudova, José Carlos Lozano

Data mining
École Polytechnique de Bruxelles
Université Libre de Bruxelles

December 17, 2023