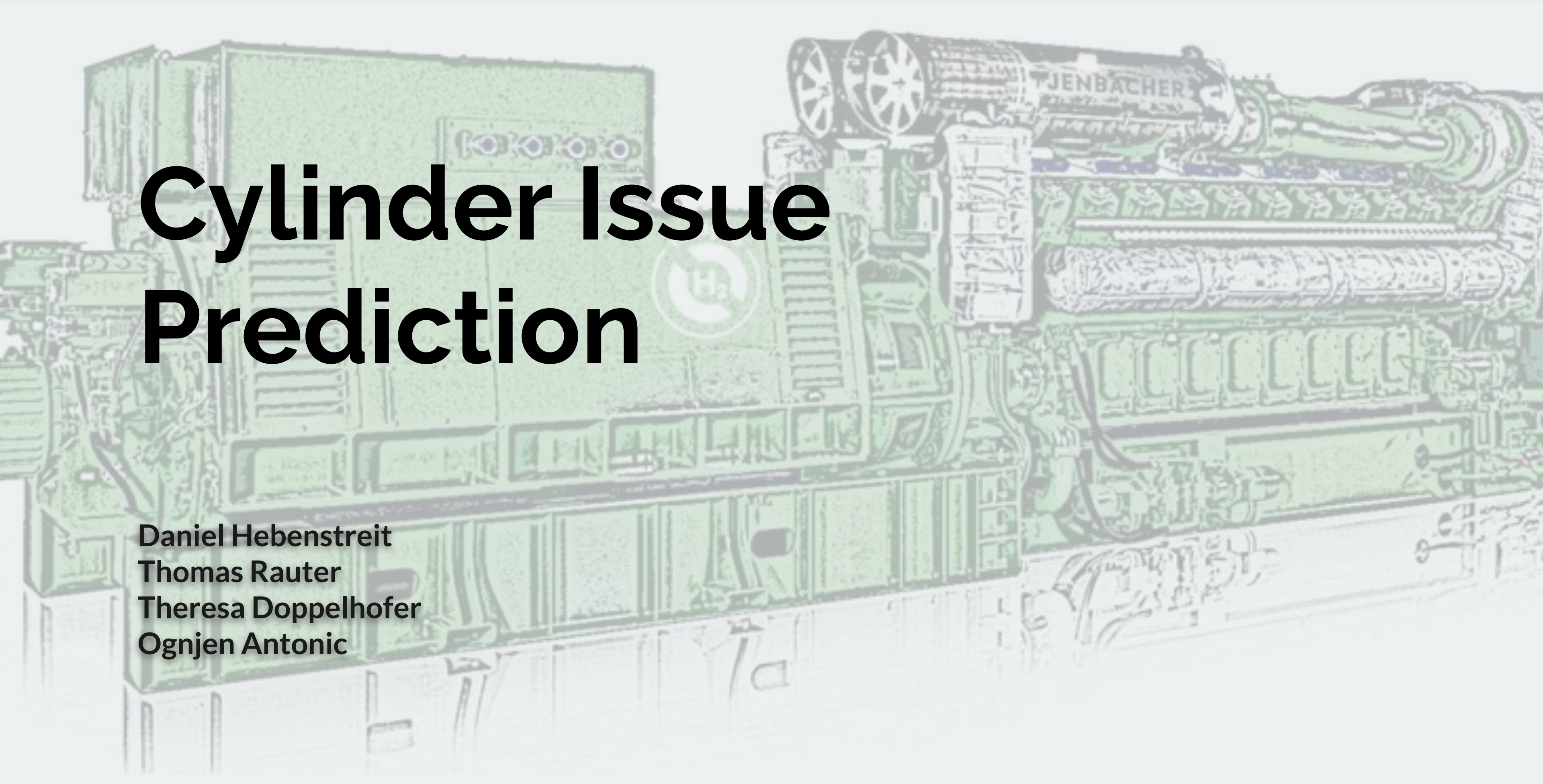


Cylinder Issue Prediction

Daniel Hebenstreit
Thomas Rauter
Theresa Doppelhofer
Ognjen Antonic

Daniel Hebenstreit
daniel.hebenstreit@student.tugraz.at
Data Science Master's Student

Theresa Doppelhofer
theresa.doppelhofer@student.tugraz.at
Data Science Master's Student

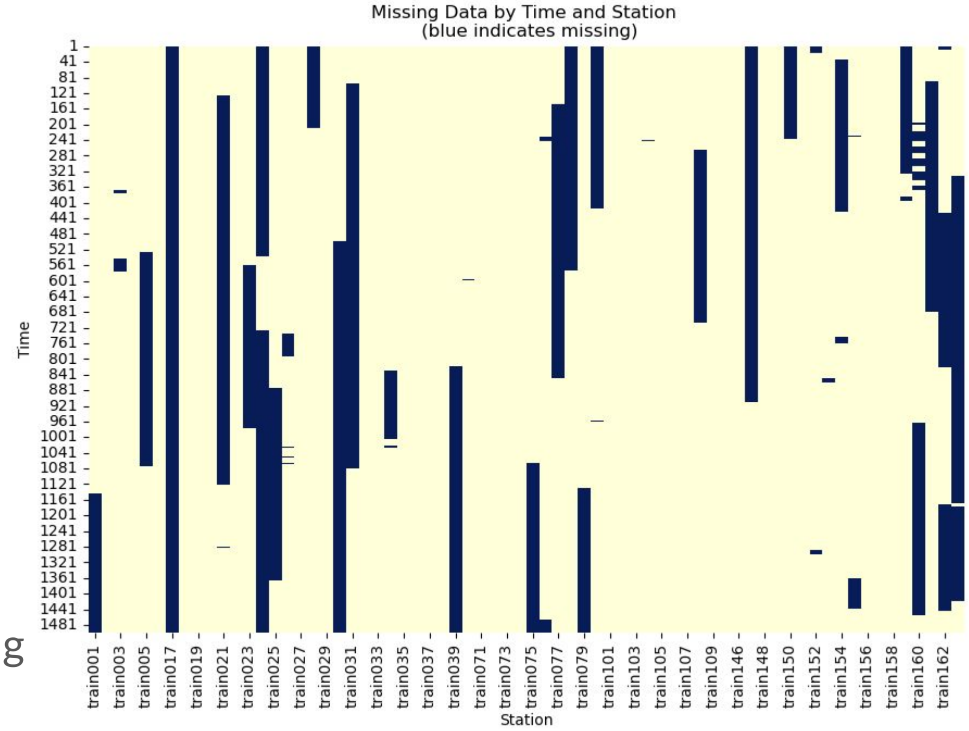
Thomas Rauter
t.rauter@student.tugraz.at
Biotechnology Master's Student

Ognjen Antonic
ognjen.antonic@student.tugraz.at
Computer Science Bachelor's Student

Table of Contents

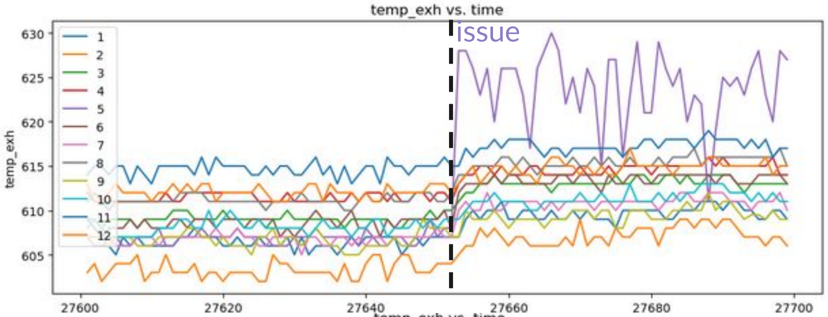
- Exploratory Data Analysis
- Feature Engineering & Pre-processing
- Model
- Experiments
- Conclusions

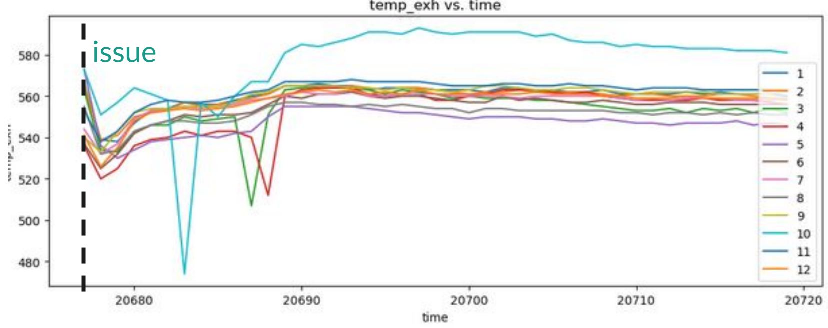
Exploratory Data Analysis


Missing Data by Time and Station (blue indicates missing)

- Missing data
  - Multiple engines affected
  - Multiple intervals of data missing
- Solution: Ignore missing data
  - Due to the disadvantages of imputation schemes (e.g. increased bias)
  - Too large and frequent gaps

Exploratory Data Analysis (cont.)





- Hard Problem:
  - Often no indicative pattern visible
- Assume Cylinder difference as key factor
- Prominent changes seem to happen in short time intervals

Singular Model Hypothesis

We can predict for each timestep well enough to get an accurate overall event classification

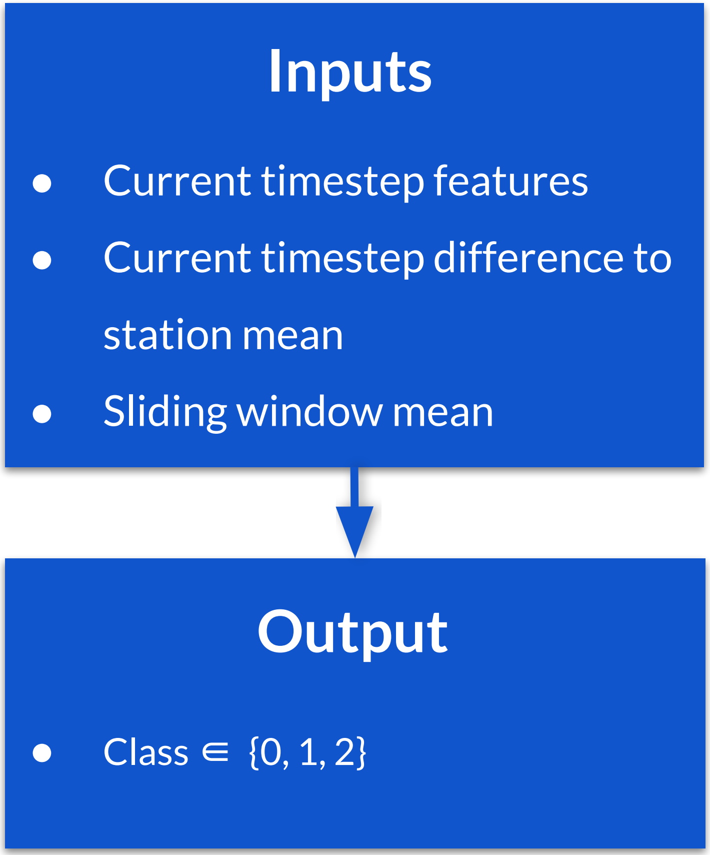
- Focus purely on predicting individual timesteps
- Post-processing for the event classification
    - Reduces number of False Positives

Feature Engineering & Pre-processing

- Information of current timestep (load, knock_control, …)
- Information w.r.t. other cylinders of the same engine
  - Difference to station mean for all features of that timestep
- Compute sliding window mean (`window_size = 30 timesteps`)
  - Contains short term history

XGBoost Classification Model

- Empirically outperformed other models
  - Naïve bayes, linear classifier, random forest
- Hyperparameters: 100 estimators
- Postprocessed outputs
  - < 5 non-consecutive $\in \{1, 2\}$ set to class 0



Inputs

- Current timestep features
- Current timestep difference to station mean
- Sliding window mean

Output

- Class $\in \{0, 1, 2\}$

Experiments

- Model that takes more than just station mean into account
  - features of current + 11 random cylinders of the same station as input
  - overcomes non-homogeneous cylinder amount
  - Inconclusive results, large variance based on seed (0.1-0.5 bmcc)
- Windowed Approaches for Event Classification
  - use mixtures of window sizes (long and short range)
  - did not perform better than single timestep prediction with postprocessing

Experiments (cont.)

- Indicators for nan values as additional feature
  - Bool if a previous value was nan
  - Number of previous values that were nan
- GMM Anomaly Detection
  - Label 0 overrepresented, A/B very sparse

Conclusions and Future Possibilities

- Our XGBoost model with window and engine information achieved **0.31** total score
  - Binary MCC: 0.15
  - Multiclass MCC: 0.06
  - Normalized Hamming Distance: 0.004
- Explore DL models (LSTM, transformer)
  - Use padding to overcome non-homogeneous number of cylinders