Interpretable AI for Business Applications

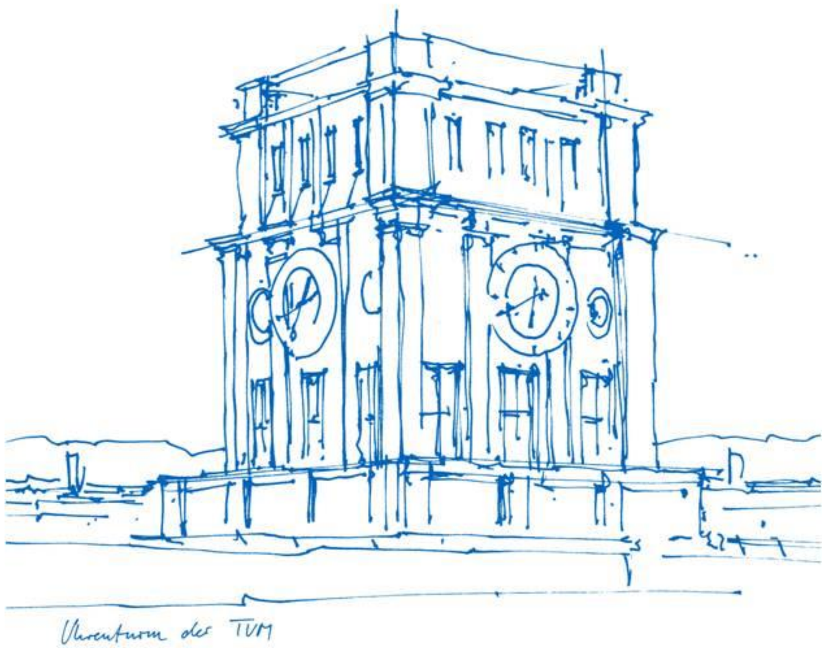
Authors Tuna Acisu,
Soh Yee Lee,
Almut Scheerer

Mentors M.Sc. Olena Schüssler,
Dr. Inna Vasylchuk

Project Lead Dr. Ricardo Acevedo,
M.Sc. Michael Rauchensteiner

Supervisor Prof. Dr. Massimo Fornasier

17th February 2020



Agenda

- Introduction
- Industrial Hydrogen Compressor Dataset
- Explainable AI
- Learning Points

Agenda

- Introduction
- Industrial Hydrogen Compressor Dataset
- Explainable AI
- Learning Points

Goal of the project

HORVÁTH & PARTNERS
MANAGEMENT CONSULTANTS

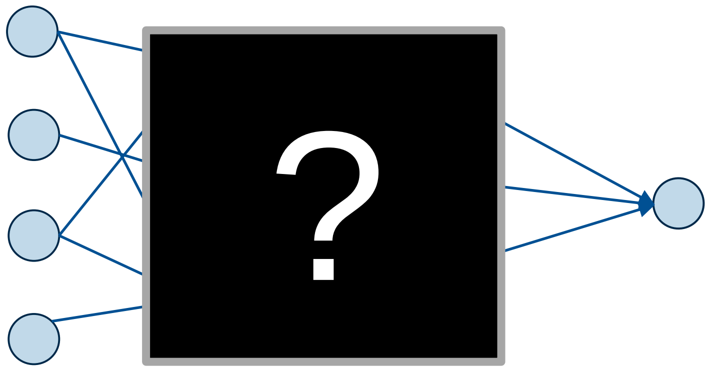
- Deliver a business model for a customer
- Employ and implement eXplainable AI

TUM

- Apply methods and approaches learned in theory into practice
- Work on a real-world data science project from start to finish

- Deliver an accurate model & human interpretable representation
- Build a prototype for a real-world use case

»Interpretability is the degree to which a human can understand the cause of a decision« ***(Miller, 2017)***

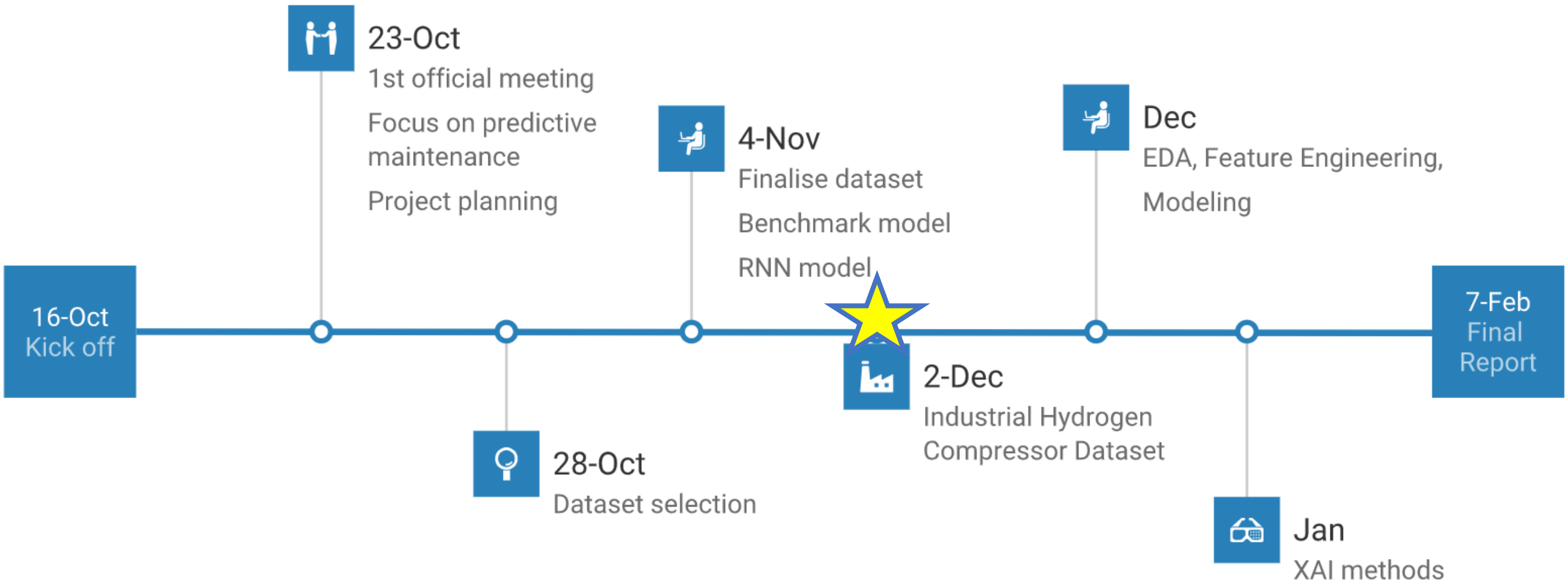


- AI methods are used for business applications
- Blackbox model → highly accurate, but not very interpretable
- Lack of transparency and trust

Goal: explain black box model output

- Gain insight on how the model works
- Detect biases
- In this project: LIME and SHAP, methods for feature influence

Project Plan



Agenda

- Introduction
- Industrial Hydrogen Compressor Dataset
  - Introduction to the dataset
  - EDA
  - Temperature Removal
  - Dataset Preprocessing
  - Modeling
- Explainable AI
- Learning Points

Industrial Hydrogen Compressor Dataset

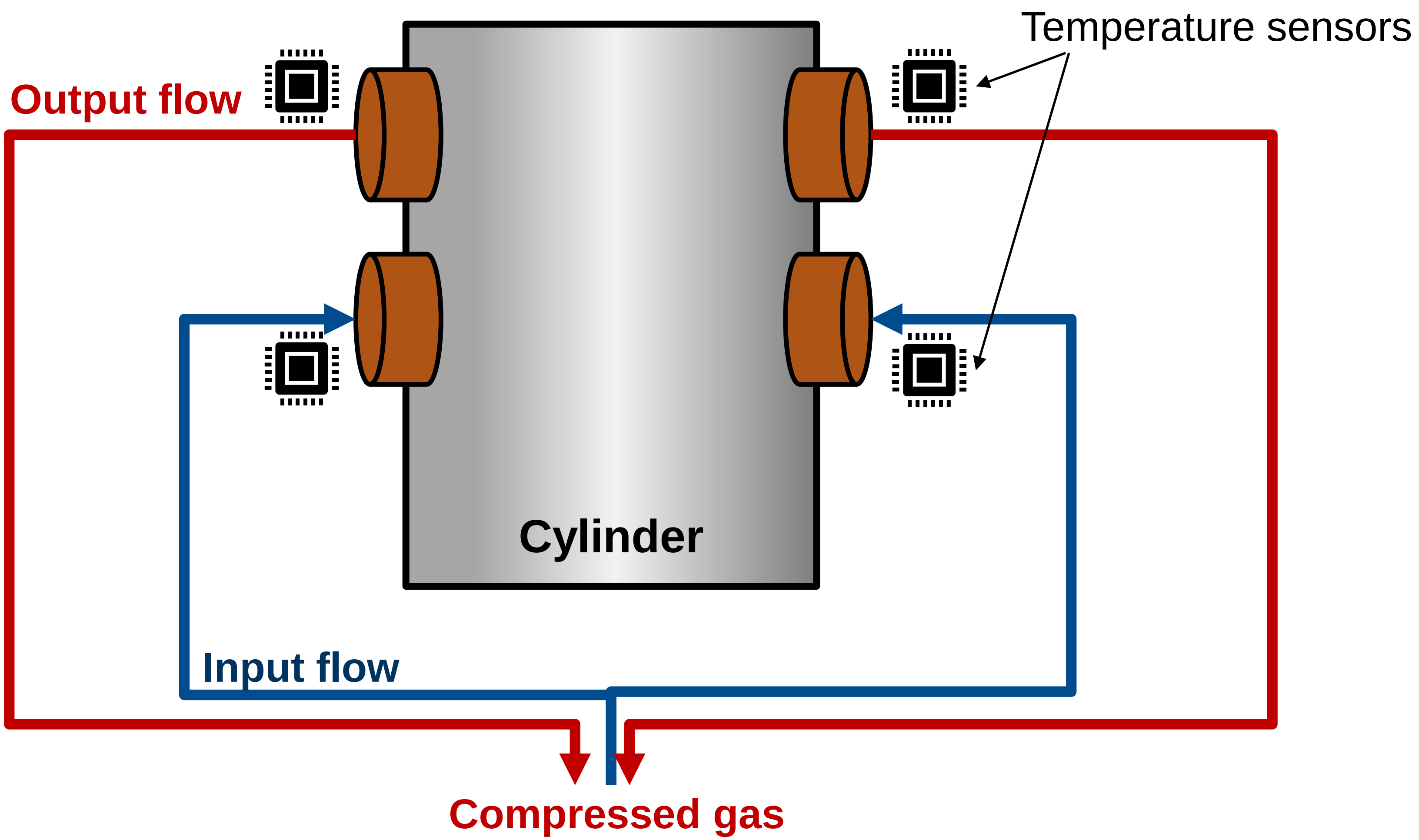
- **Real world dataset**: Sensor data from turbine and compressor
- More than 80 different sensors measuring e.g. pressure, temperature
- Measurement only taken at certain deviation from baseline
- Data collected over 19 years, between 0.13 million and 5.1 million measurements *per sensor*

- **Event:** Valve breakage, leads to machine downtime

- Very imbalanced dataset, very few events (<1%)

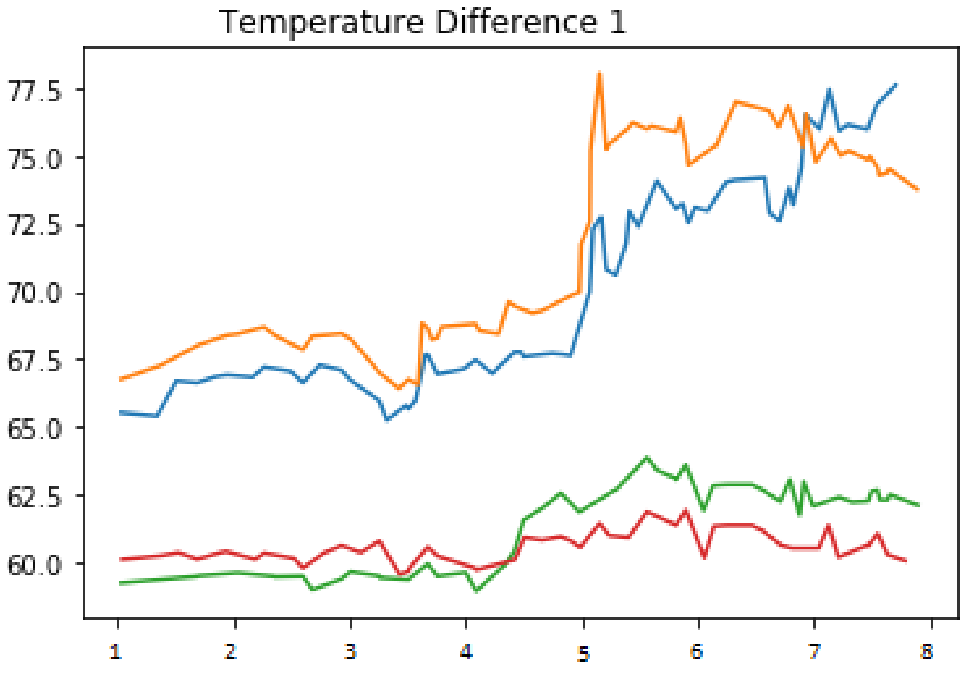
Understanding the data

Inside the compressor: Cylinders and Valves



Industrial Hydrogen compressor dataset: Events

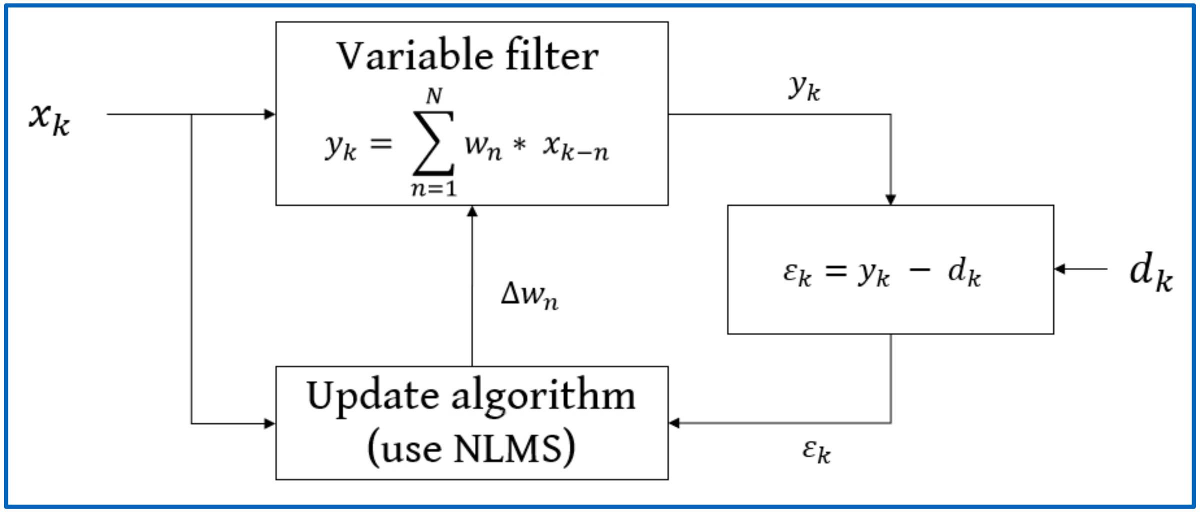
Can we see patterns in the data?



- Differences in temperature go up before valve breaks

Removal of outside influence - adaptive filtering

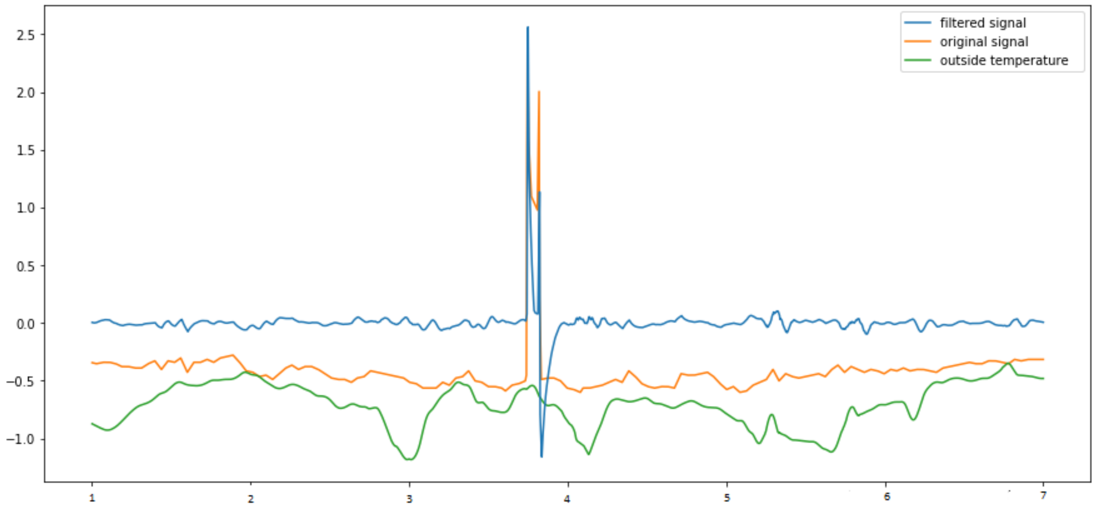
- Temperature sensors are crucial features
- But: fluctuation from outside temperature affect measurements

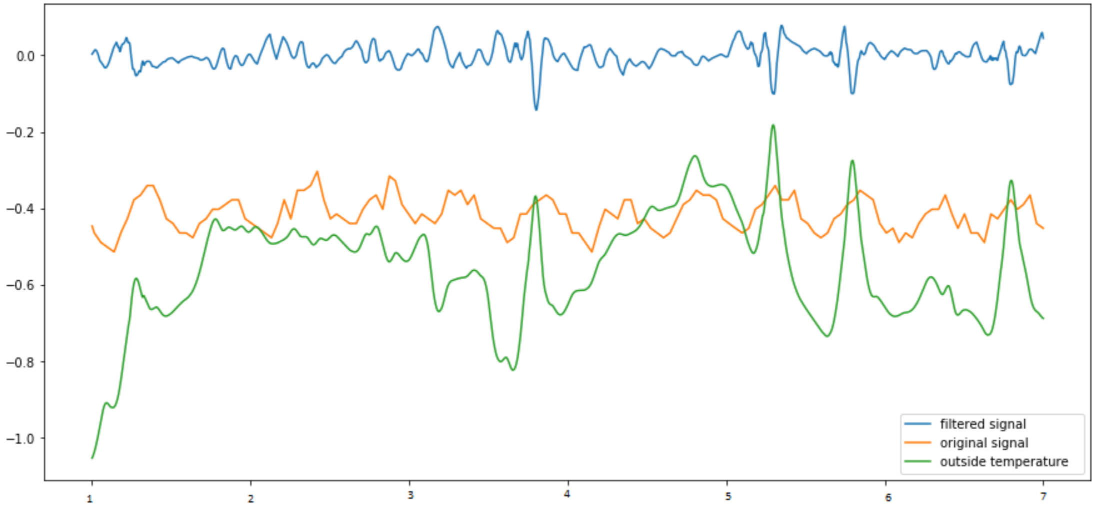


- Two inputs to filter: x_k $(observation)$ and $d_k(desired\ signal)$
- Filter tries to find d_k in $x_k \rightarrow y_k$
- Residual signal: $r_k = x_k - y_k$
- x_k raw valve measurements, d_k outside temperature, r_k filtered valve measurements

  → use adaptive filtering to remove influence from outside temperature!

Removal of influence - adaptive filtering





Correlation before filtering

	outside	valve
outside	1.0	0.47
valve	0.47	1.0

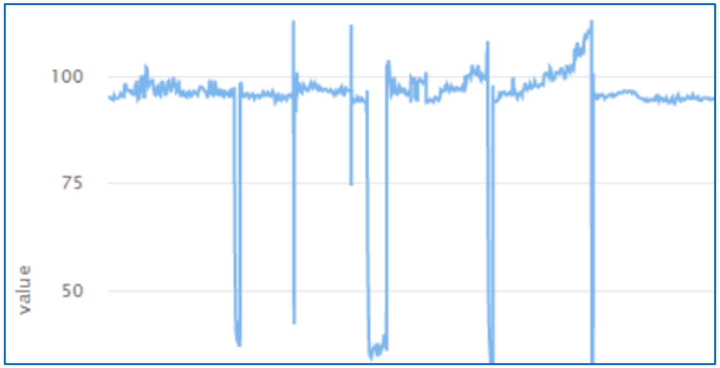
Correlation after filtering

	outside	valve	filtered signal
outside	1.0	0.47	0.01
valve	0.47	1.0	0.09
filtered signal	0.01	0.09	1.0

Important Data Preprocessing Summary

Sync time interval → Feature engineering → **Remove shutdown period** → Aggregate features per day → **Append events & "Manual" oversample** → Train model

Machine downtime



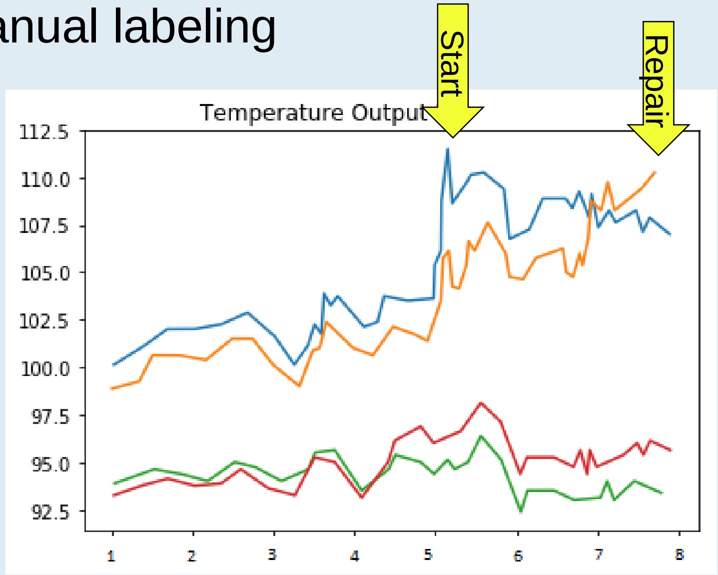
Temperature sensor reading

- Derive interval, mathematically and domain knowledge from client

Sensor reading ≥ t	Sensor reading < t	Downtime (Sensor reading = room temperature)	The turbine is powered up again Sensor reading ≥ t

"Healthy" | Exclude 2 hours | "Downtime" | Exclude 2 days | "Healthy"

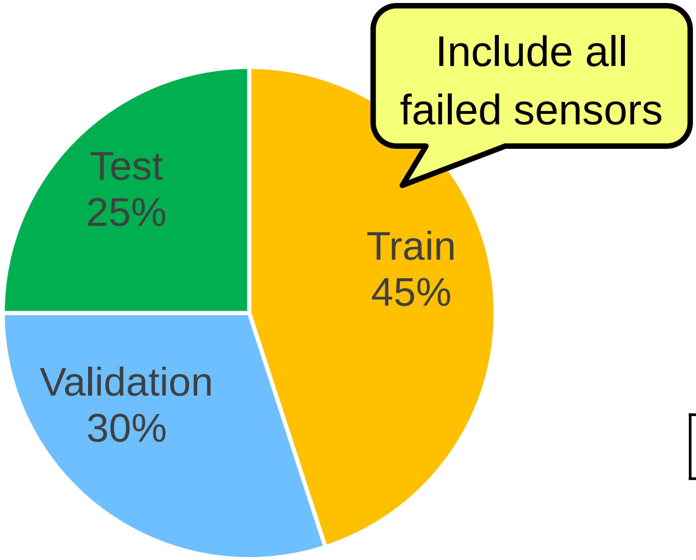
Important Data Preprocessing Summary (1)

Challenges	**Solutions**
Messy Repair Data - Unknown broken valves - Unknown start time of abnormality - Maintenance or real breakage?	Manual labeling 
Extremely imbalanced dataset	• Cost-sensitive learning • Manually "oversample"

Problem Definition

- Predict **daily** valve breakage abnormally as a whole ⇒ 1
- Apply XAI methods to find abnormal sensors

Train / Validation / Test Split



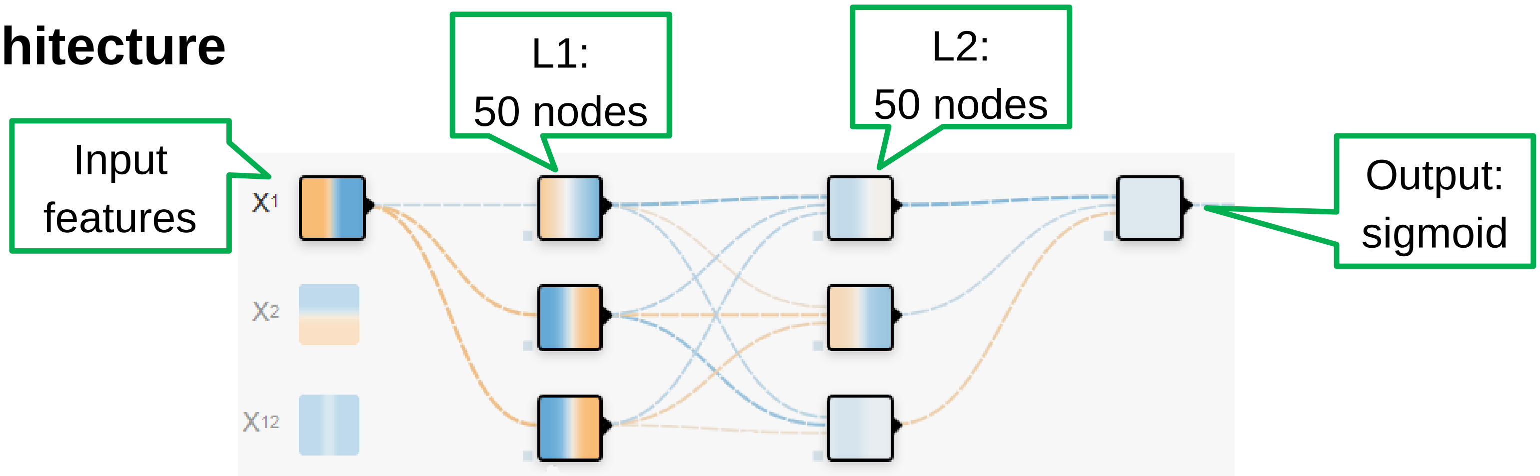
Holdout

Cross Validation ⇐ Overfitting

Breakage 1	Day 1	Day 2	Day 3

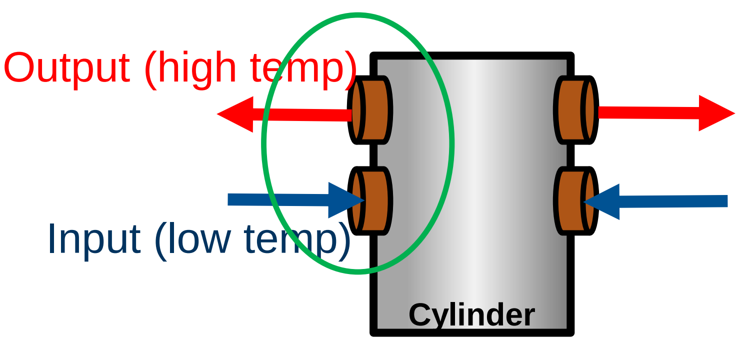
Modeling

NN Architecture

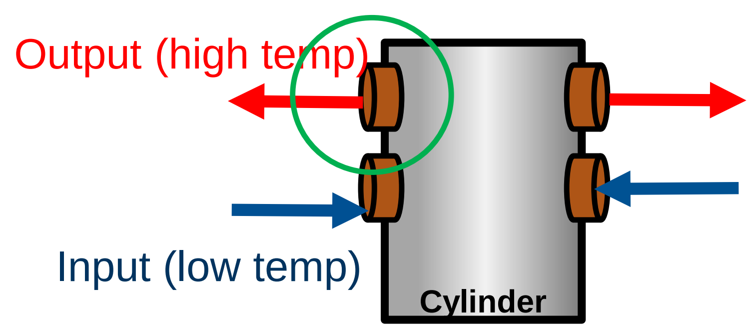


Feature Engineering

(1) "Paired"

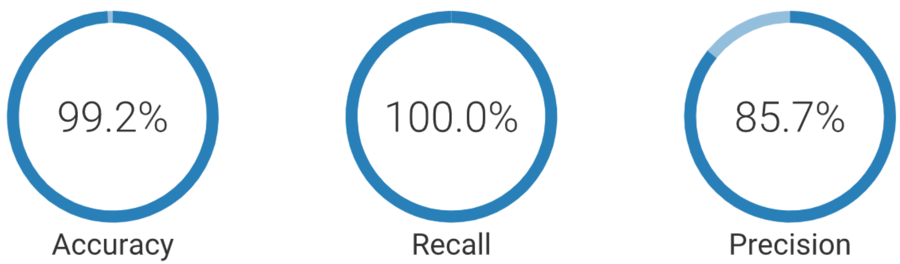


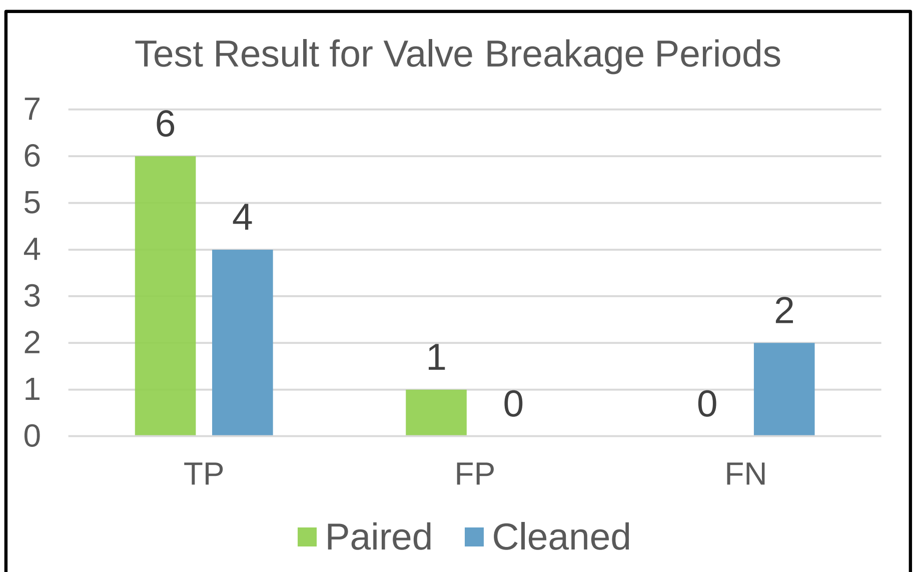
(2) "Cleaned"

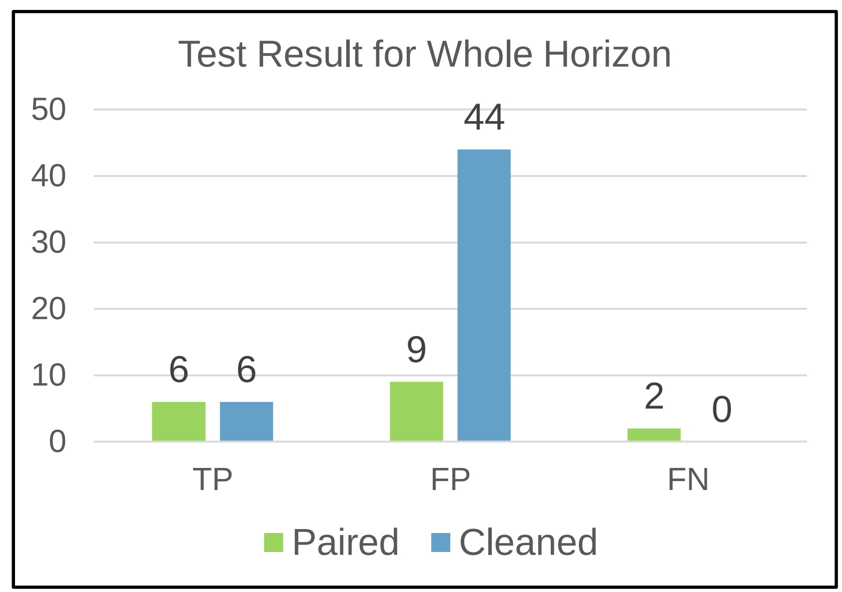


Result

- Experimented with more than 1000 models.
- Best 100 epochs with early stopping for valve breakage periods.





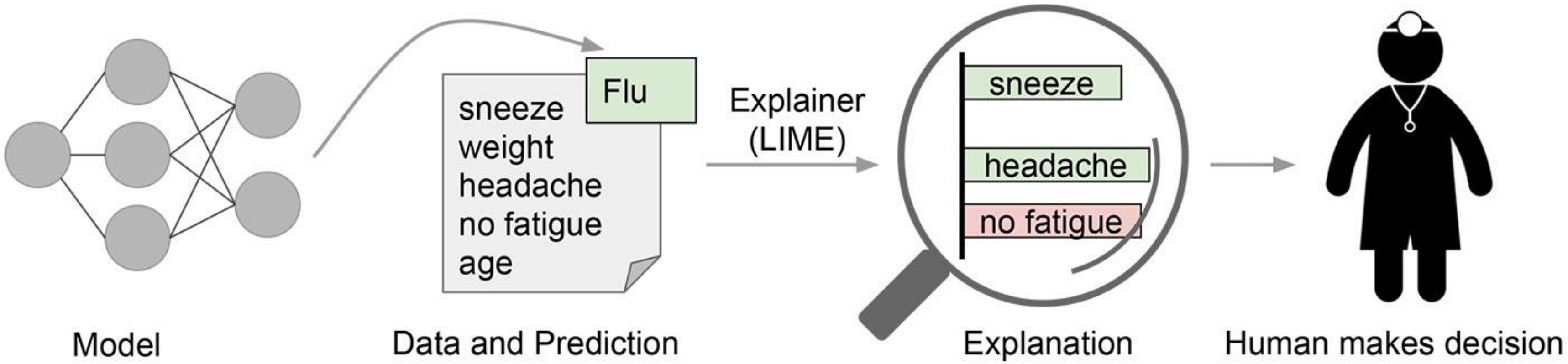


Agenda

- Introduction
- Industrial Hydrogen Compressor Dataset
- Explainable AI
  - LIME
  - SHAP
- Learning Points

Local Interpretable Model-Agnostic Explanations (LIME)

Goal: Explain a prediction by learning a linear model locally around it



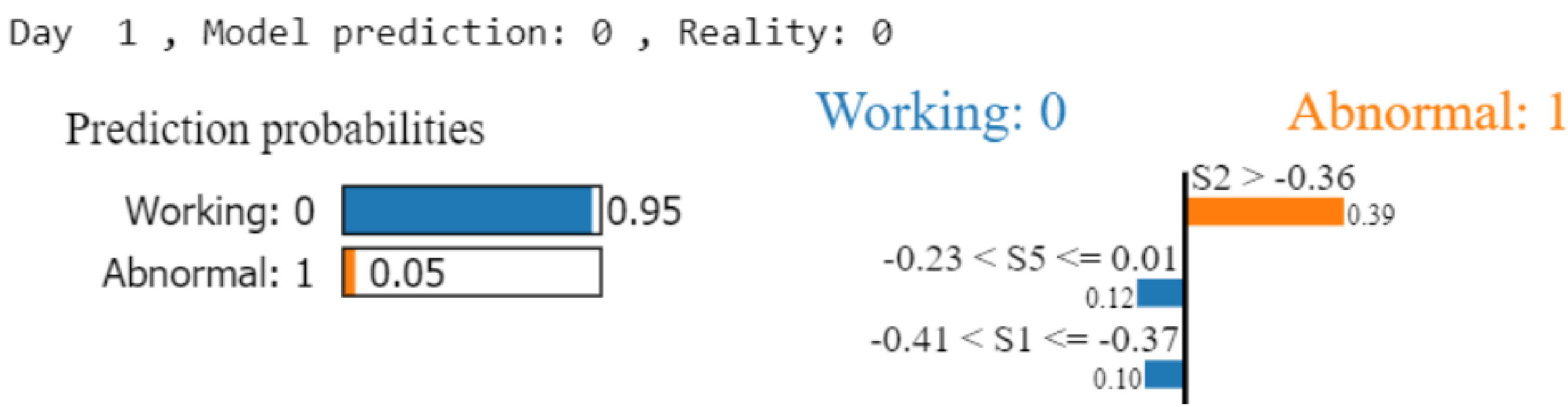
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). “Why should i trust you” Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144)

Model uses the paired temperature features

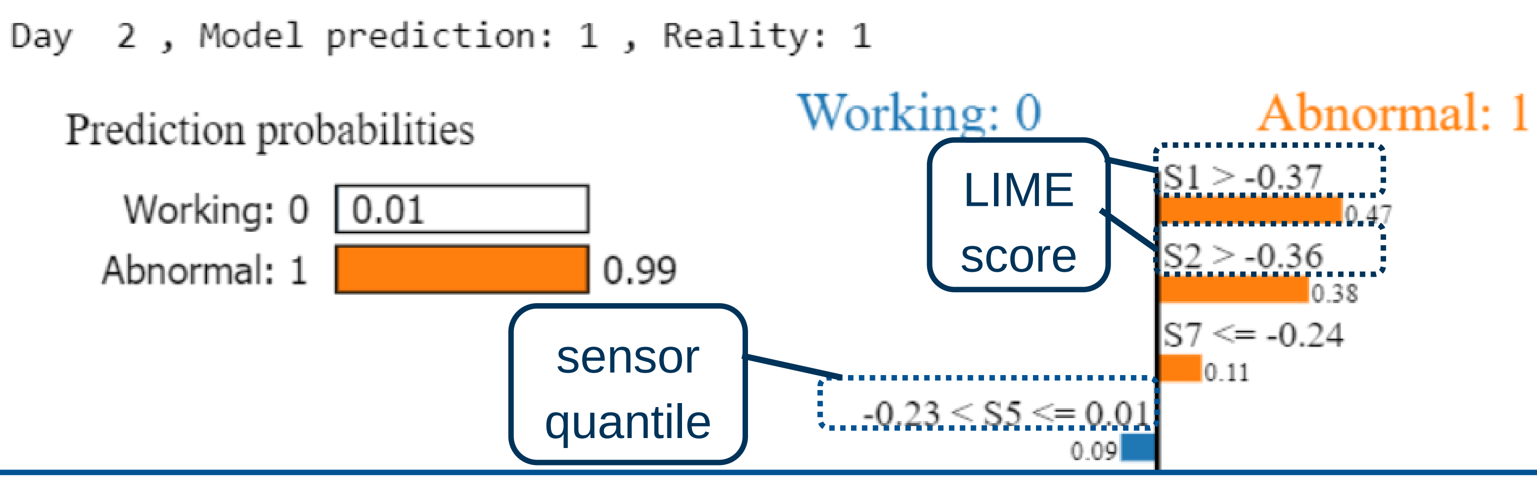
Explanations are created using discretized features (in quantiles)

LIME – Results

Day 1



Day 2



LIME – Results

- checked all events:
  ⇒ not all top explanations match the affected valve pair(s)
- this linear approximation is not good

⇒ tried another method, SHAP

SHapley Additive exPlanations (SHAP)

Goal: Explain a prediction by the Shapley Value

Shapley Value

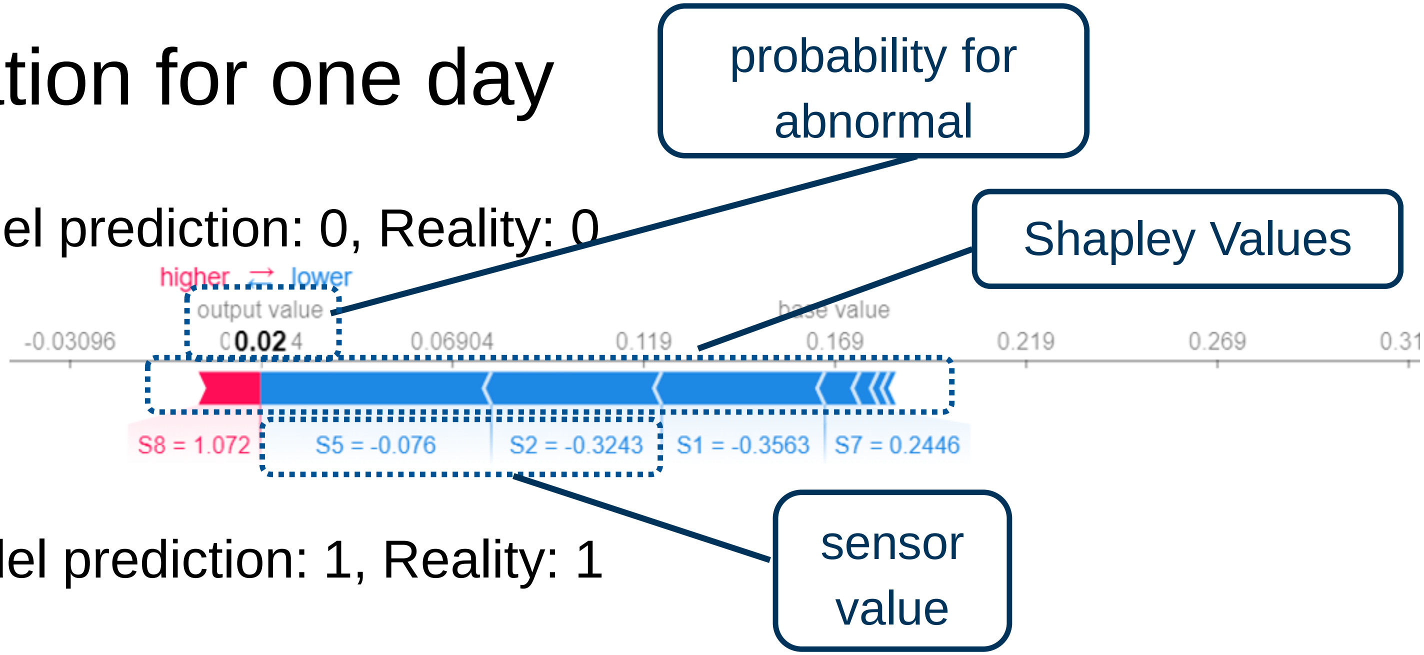
- In Game Theory: Method for fair allocation of output among the members of a coalition

$$\Phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left(v(S \cup \{i\}) - v(S)\right)$$

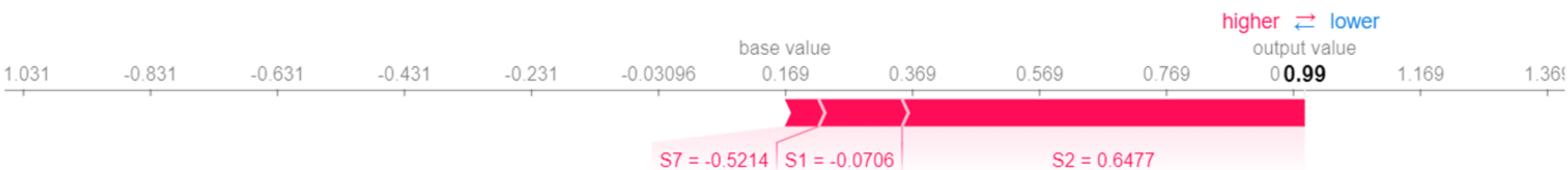
- Here: each feature value = player in a game; prediction = payout
- Use approximation method: KernelSHAP

Explanation for one day

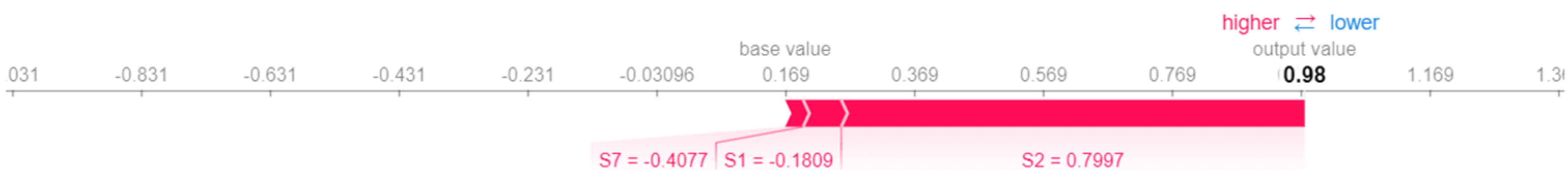
Day: 1, Model prediction: 0, Reality: 0



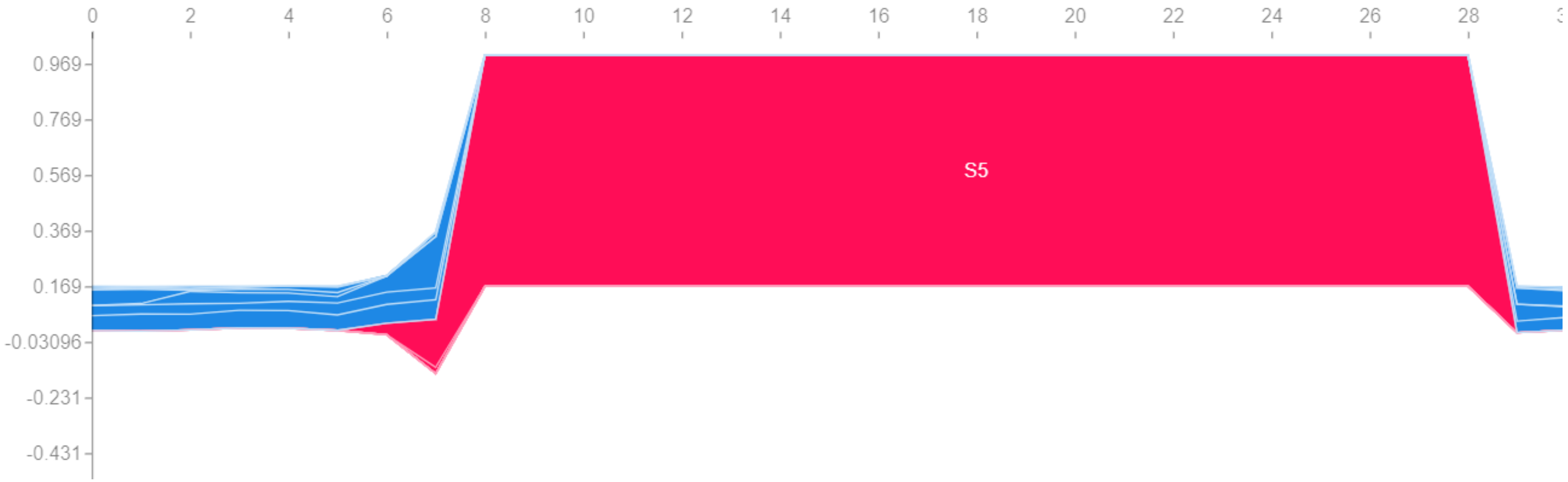
Day: 2, Model prediction: 1, Reality: 1



Day: 3, Model prediction: 1, Reality: 1



Interval of one month



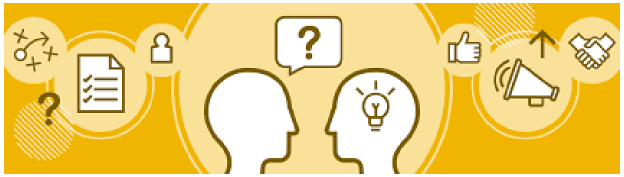
Conclusion: It is possible to identify the affected valve pair(s) by SHAP

Agenda

- Introduction
- Industrial Hydrogen Compressor Dataset
- Explainable AI
- Learning Points

Learning Points

- Hands-on approach to a real-world data analysis task
- No data analysis task is the same
- Many innovative ideas are needed
- Even Neural Networks are interpretable & verifiable
- Project group work can be challenging
- A happy client is very rewarding



Thank you for your attention!

References

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). *“Why should i trust you” Explaining the predictions of any classifier.* Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144)

[2] Hannes Knobloch, Adem Frenk, and Wendy Chang. *Predicting Battery Lifetime with CNNs.* Available at https://towardsdatascience.com/predicting-battery-lifetime-with-cnns-c5e1faeecc8f. [Accessed November 2019]. Sept. 2019.

[3] Scott M Lundberg and Su-In Lee. *A unified approach to interpreting model predictions.* In: Advances in neural information processing systems. 2017, pp. 4765 - 4774.

[4] Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* https://christophm. github.io/interpretable-ml-book.

[5] Kristen A Severson et al. *124 Lithium-ion data set.* Available at https://data. matr.io/1/projects/5c48dd2bc625d700019f3204. [Accessed October 2019]. Mar. 2019.

[6] Kristen A Severson et al. *Data-driven prediction of battery cycle life before capacity degradation.* In: Nature Energy 4.5 (2019), p. 383.

[7] Lloyd S Shapley. *A value for n-person games.* In: Contributions to the Theory of Games 2.28 (1953), pp. 307 - 317.

[8] Bernard Widrow. *Adaptive Signal Processing.* Pearson, 1985, pp. 307 - 317