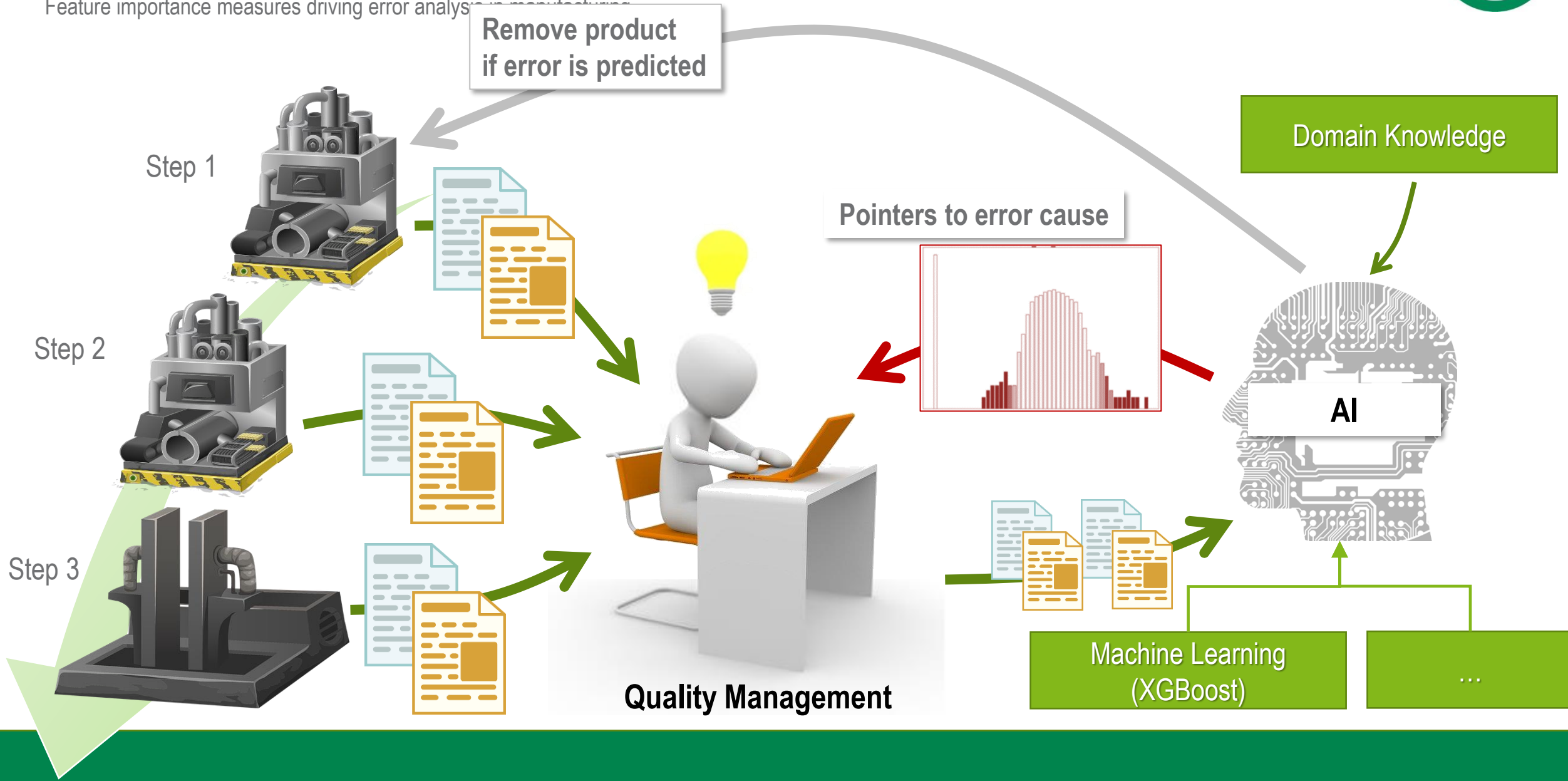
HOCHSCHULE FURTWANGEN UNIVERSITY | HFU

Shapley Values based Regional Feature Importance Measures driving Error Analysis in Manufacturing

Introduction

Feature importance measures driving error analysis in manufacturing



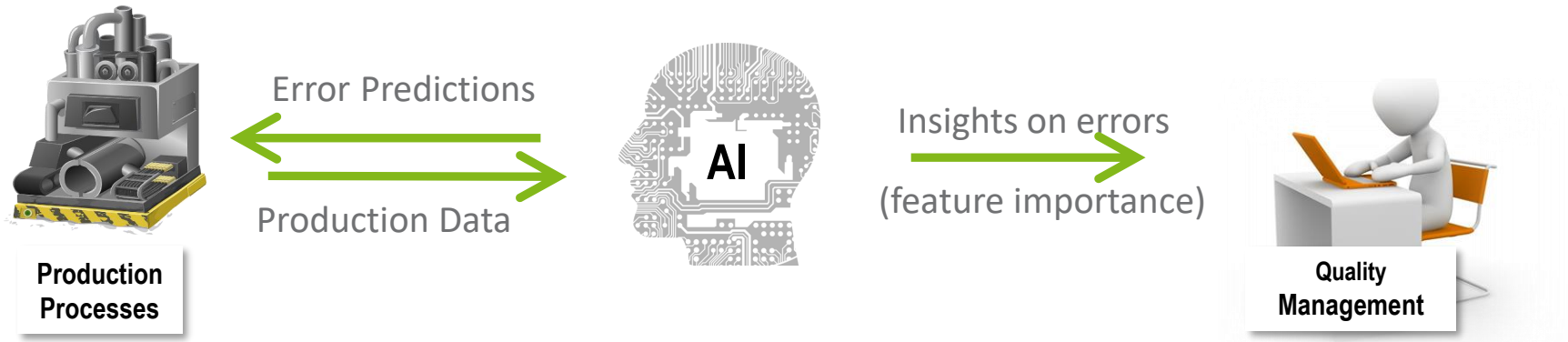
Introduction

Feature importance measures driving error analysis in manufacturing

- **Data driven error analysis can leverage predictive models for error cause analysis**
  - **Quality Management analyzes features pointing at “interesting” phenomena**
    - **Feature importance measures provide insights to the Human-In-the-Loop**
    - **State-of-the-Art feature importance measures are not tailored to this task**

Contribution:

- Define an applied notion of “interestingness” and proposed three approaches to determine this as feature importance measure
  - Evaluating and comparing with state-of-the-art importance measures on synthetic and real-world data



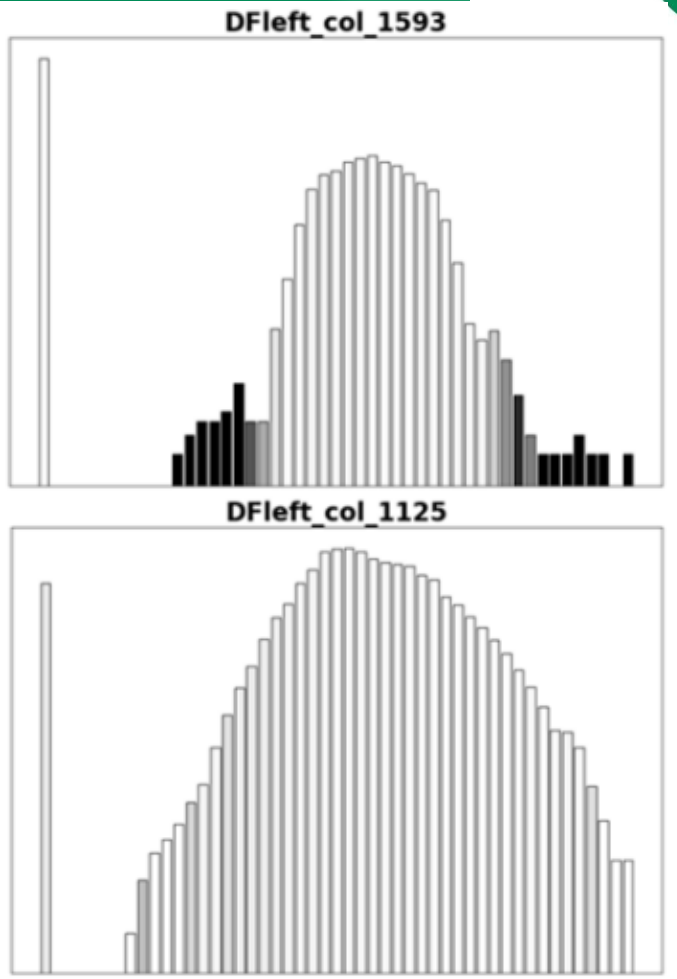
Agenda

- Motivation and Use Case
- Challenges with feature importance in error analysis
- Proposed SHAP-based approaches to determine regional feature importance
- Experiments and results
- Conclusion

Motivation and Use Case

Explaining errors in manufacturing data driven quality processes

- Proactive error prevention for rare but costly errors in production [1]
- Data driven error analysis approach to determine error causes
- Domain experts evaluate production data and take corrective steps in production
- Traditionally, cause analysis is performed by evaluating input features with correlated errors
- Increasing number of features requires automated feature ranking methods
- **Which measures of feature importance suits the task for cause analysis?**



Histograms of features with highlighted errors traditionally used for cause analysis [1]. Color coded error percentage: white=0%, black=100%

Challenges with feature importance

Global importance

Local importance

Feature Importance

Challenges with feature importance in cause analysis

Global

Local

Regional

Holistic view:

- Importance based on global “averages”
- Missing local phenomena

- Examples: Gain, Weight/Frequency, Gini, Abs. Avg. SHAP, …

Feature Importance

Challenges with feature importance in cause analysis

Global

Local

Regional

Instance based:

- Importance based on single samples
- Isolated samples provide little general insights
- Missing context e.g., relations to other samples

Feature Importance

Challenges with feature importance in cause analysis

Global

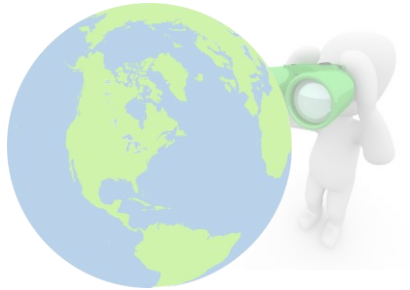
Local

Regional

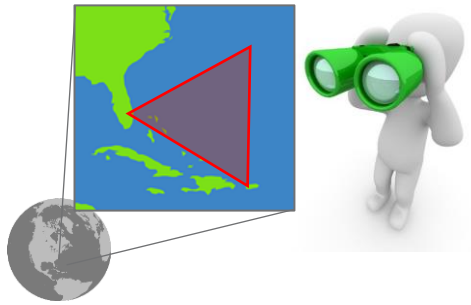
What is needed for feature importance in cause analysis?

A feature should have high importance if it **contributes to interesting predictions** [1]:

- *it provide at least sometimes strong hints for errors*
- *it is relevant in at least some cases*
- *it allows to draw conclusions upon inspection*



Global and generalizing



Holistic but targeted to interesting aspects

Case specific

Shapley values as basis for regional feature importance

Shapley values as basis for importance measures

Feature importance for cause analysis

"The Shapley value is the average marginal contribution of a feature value across all possible coalitions." [3]

SHAPley Values

Tree Explainer

Global Importance

- Desirable properties: Accuracy, consistency, missingness
- Shapley values can be computed for each prediction individually
- SHAP (SHapley Additive exPlanation)

Shapley values as basis for importance measures

Feature importance for cause analysis

SHAPley Values

Tree Explainer

Global Importance

The contribution $\phi_f \in \mathbb{R}$ of feature f on model m is attributed using Shapley values [3]:

$$\phi_f = \sum_{S \subseteq M \setminus \{f\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [m_x(S \cup \{f\}) - m_x(S)]$$

where M is the number of all features and S the set of input values.

Shapley values as basis for importance measures

Feature importance for cause analysis

The contribution $\phi_f \in \mathbb{R}$ of feature f on model m is attributed using Shapley values [3]:

$$\phi_f = \sum_{S \subseteq M \setminus \{f\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [m_x(S \cup \{f\}) - m_x(S)]$$

SHAPley Values

Tree Explainer

Global Importance

where M is the number of all features and S the set of input values.

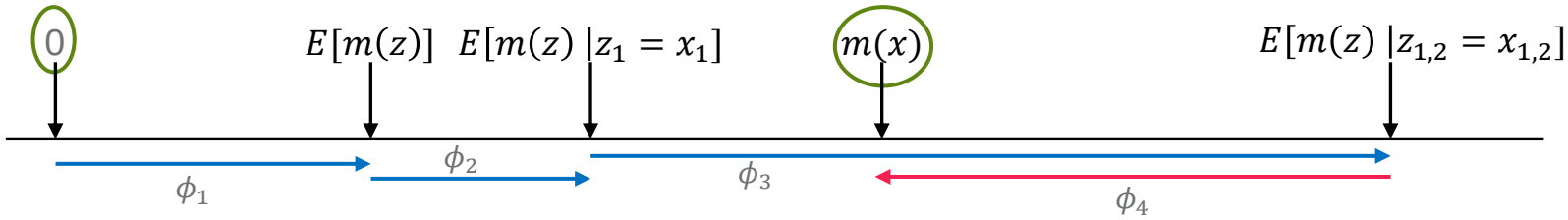


Figure – "SHAP feature attribution ": Contribution of each feature as change in the expected model prediction when conditioning on that feature (source [4]).

Shapley values as basis for importance measures

Feature importance for cause analysis

SHAPley Values

Tree Explainer

Global Importance

- Implementation for trees [4] (i.e., XGBoost)
- Computation of exact Shapley values in polynomial time
- $O(TLD^2)$, where T is the number of trees, L is the maximum number of leaves in any tree and D the maximal depth of any tree

Shapley values as basis for importance measures

Feature importance for cause analysis

SHAPley Values

Tree Explainer

Global Importance

Idea: Features with large absolute Shapley values are important

Average absolute Shapley values per feature across the data [4]:

$$I_f = \frac{1}{n} \sum_{i=1}^{n} |\phi_f^{(i)}|$$

Importance measures that reveal insights on errors

Previous Work

Importance measures that reveal insights on errors [1]

Fundamental idea:

- Aggregate Shapley values so that “interesting” features get a high score (i.e., not just averaging them for a global)

Name	Intuition	Formal Definition
Max Shap	Highest Shapley value across analysed dataset	$Max\ SHAP_f(m,S) = \max\{\phi_f(m,x) \mid x \in S\}$
Max Main	Like Max SHAP but without interaction effects	$Max\ Main\ Effect_f(m,S) = \max\{\phi_f(m,x) - \sum_{j \neq f} \phi_{f,j}(m,x) \mid \mid x \in S\}$
Range Shap	Range of Shapley values across analysed dataset	$Range\ SHAP_f(m,S) = \max\{\phi_f(m,x) \mid x \in S\} - \min\{\phi_f(m,x) \mid x \in S\}$

Proposed SHAP-based regional feature importance measures

Fundamental Idea

Proposed Regional Feature Importance Measures

Outlier-Approach

Micro-Average Approach

Slope-Approach

Fundamental idea:

- A scoring-function g: $g(f, X, \dots) \rightarrow \mathbb{R}$ that aggregates SHAP values and scores „interesting“ features high

f: target feature

X: dataset

Idea: A feature with abnormal Shapley values are potentially interesting

Outlier-Approach

Micro-Average Approach

Slope-Approach

Perform anomaly detection over the distribution of SHAP values:

$$g(f, X, \lambda) = \sum_{x' \in \mathrm{outl}(\lambda, X)} \phi_f(x')$$

$$\mathrm{outl}(\lambda, X) = \{ x \in X \mid \phi_f(x) \geq \bar{\phi}_f(X) + \lambda \, \sigma(\phi_f(X)) \}$$

f: target feature

X: dataset

σ: standard deviation

λ: multiplier of σ

$\bar{\phi}_f$: mean shap value

Outlier-Approach

Micro-Average Approach

Slope-Approach

Idea: A feature is of interest if it shows high Shapley values within a small feature value range

Partition the feature and determine the average SHAP values over equally sized intervals:

$$g(f, X, n) = \max\{\bar{\phi}_f(X_i) \mid i = 0, \dots, n-1\}$$

$$X_i = \{x \in X | (i * d) \leq x < (i * d) + d\}$$

f: target feature

X: dataset

n: number of intervals

$\bar{\phi}_f$: mean shap value

X_i: feature interval

Idea: A feature with rapid changes in Shapley values is of interest

Outlier-Approach

Micro-Average Approach

Slope-Approach

Partition the feature and determine regression slopes over the means of SHAP values over the intervals:

$$g(f, X, n, w) = max\{|slope(w_j)|\,|\, j = 0, \dots, n-1\}$$

$$w_j = \{\bar{\phi}_f(\mathrm{X_j}), \dots, \bar{\phi}_f(\mathrm{X_{j+w}})\}$$

$$\bar{\phi}_f(X_i) = i * \beta + \epsilon \mid \forall \bar{\phi}_f(X_i) \in w_j$$

f: target feature

X: dataset

n: number of intervals

w: window size

$\bar{\phi}_f$: mean shap value

X_i: feature interval

HOCHSCHULE FURTWANGEN UNIVERSITY HFU

Experiments

Experimental setup

HOCHSCHULE FURTWANGEN UNIVERSITY HFU

Setup 1: "Synthetic Dataset"

Experiments with synthetic data

- Creation of simple data sets with known "***interestingness***" of features
- Comparison of proposed importance measures with established importance measures against know ground truth

Setup 2: "Real-World Dataset"

Experiments with real-world manufacturing data

- Scoring of feature with proposed and established importance measures
- Manual inspection of features regarding ***"interestingness"***

- Binary classification with XGBoost with training score f1=1.0
- Computation of importance score on the training set

Experiments synthetic data

Example experiment with synthetic data

- Features A, B with e.g., uniform distribution 0 to 1 and one feature noise
- 3% of data points have errors

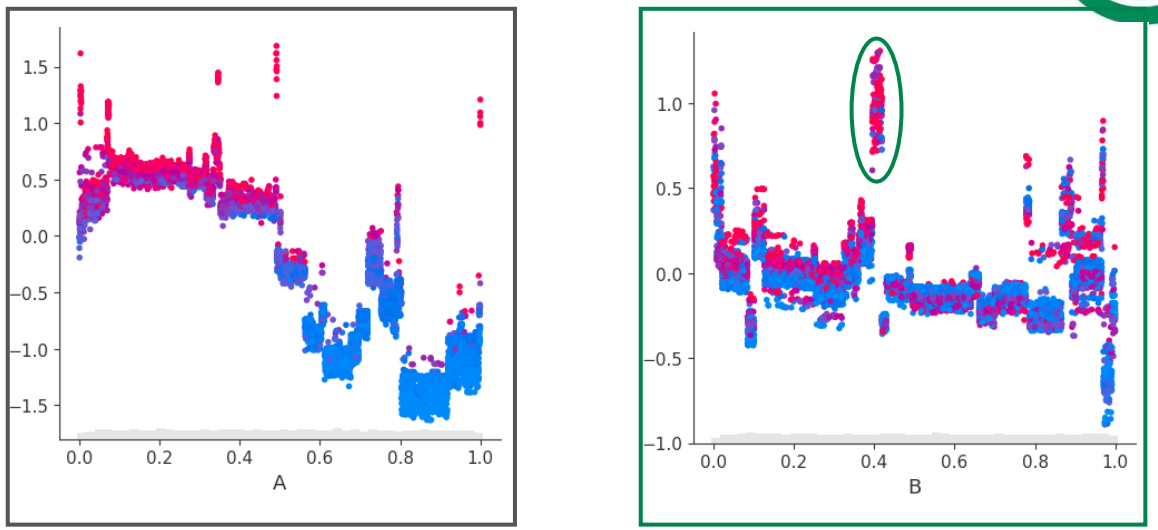


Figure “Shap plots of feature A and B”: Not interesting feature A (left) and interesting feature B (right). Color coded – red: error, blue: non error, green marker: area of interest

- If 0.4 < B < 0.425, then 8% of data points have errors
- If A < 0.6, then 2.5% of data points have errors

Not interesting

Interesting

Rank	Classic Metrics					SHAP-based Metrics				Proposed Metrics			
										Slope Approaches		Micro-Avg Approach	Outlier Approach
	Weight	Gain	Cover	Total Gain	Total Cover	Average Abs SHAP	Max Main	Max Shap	Range SHAP	Max SHAP	IQR SHAP		
1	A	A	A	A	A	A	A	A	A	B	B	B	B
2	B	B	B	B	B	B	noise	noise	B	Noise	A	A	A
3	noise	noise	noise	noise	noise	noise	B	B	noise	A	noise	noise	noise

Experiment with real world-data

- Secom dataset originating from a semiconductor manufacturer [6]
- 591 features and 1667 instances with 106 error instances
- Manual assessment of the top five features of all importance measure
- Comparison of grouped metrics with proposed metrics

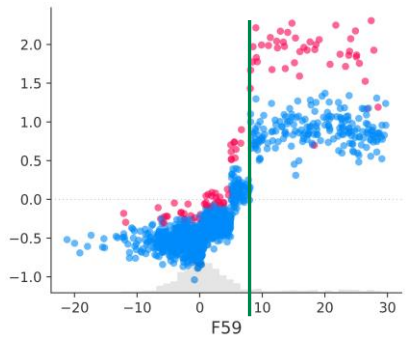
Rank	Classic Metrics					SHAP-based Metrics				Proposed Metrics			
										Slope Approaches		Micro-Avg Approach	Outlier Approach
	Weight	Gain	Cover	Total Gain	Total Cover	Average Abs Shap	Max Main	Max Shap	Range Value	Max Shap	IQR Shap		
1	F59	F210	F168	F59	F59	F59	F59	F59	F59	F59	F59	F64	F59
2	F333	F539	F429	F333	F64	F21	F64	F64	F64	F64	F64	F59	**F423**
3	F103	F29	F426	F64	F426	F333	F40	F426	F333	F429	F33	F103	F64
4	F2	F109	F100	F132	F121	F488	F426	F333	F103	F333	**F130**	F40	F333
5	F33	F304	F331	F33	F574	F103	F153	F40	F33	**F475**	F429	F121	F2

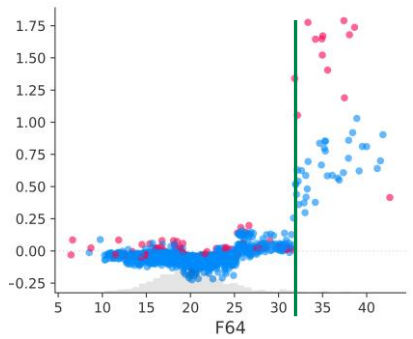
Experiment with real world-data

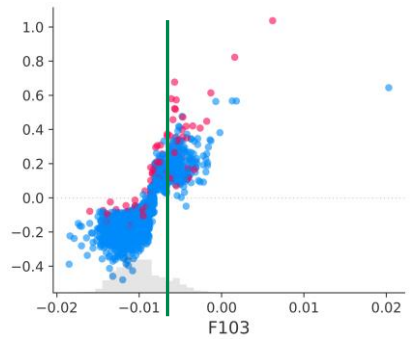
Agreements of importance

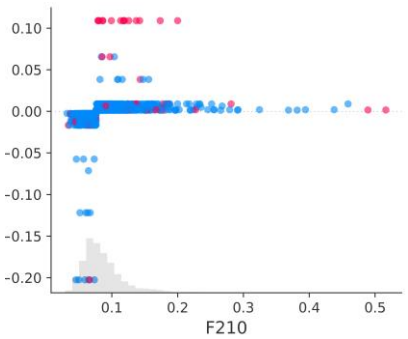
Disagreements of importance

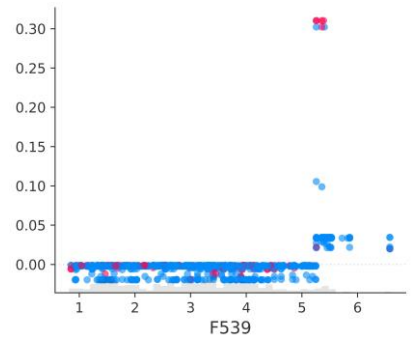
Highlights of proposed

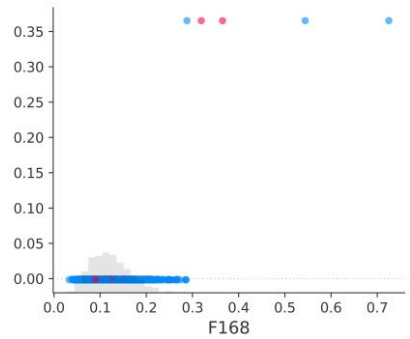


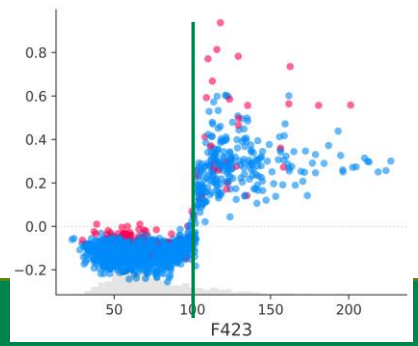


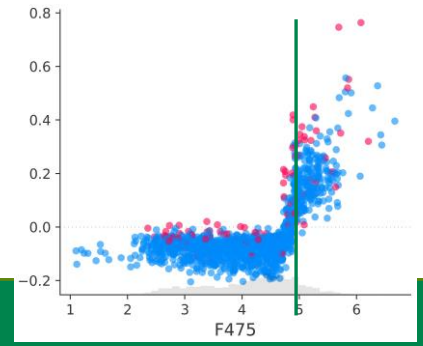


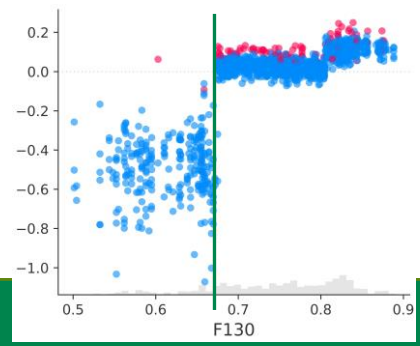












Conclusion and future work

Conclusion:

- Shapley values have potential as the basis for regional importance measures
- Regional feature importance measures can pinpoint interesting features for cause analysis in manufacturing
- Experiments show the usefulness of proposed SHAP-based approaches for cause analysis

Future Work:

- Evaluation on “interestingness” needs more objective measure
- Evaluation of importance measures with domain experts


Thank
You!

REFERENCES

[1] H. Ziekow, U. Schreier, A. Gerling, and A. Saleh, "Interpretable Machine Learning for Quality Engineering in Manufacturing - Importance Measures that Reveal Insights on Errors", *The Upper-Rhine Artificial Intelligence Symposium, UR-AI 2021, Artificial Intelligence - Application in Life Sciences and Beyond*, Germany, Kaiserslautern: Hochschule Kaiserslautern, University of Applied Sciences, pp. 96–105, October 2021

[2] XGBoost Documentation, “Python API”, *Reference. xgboost developers*. [Online]. Available from: https://xgboost.readthedocs.io/en/latest/python/python_api.html, retrieved on 08/26/2022

[3] C. Molnar, “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable”, *2nd edn.*, 2022. [Online]. Available from: https://christophm.github.io/interpretable-ml-book, retrieved on 08/26/2022.

[4] S. M. Lundberg, and S. I. Lee, "A unified approach to interpreting model predictions", *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, NY, USA, pp. 4768–4777, 2017.

[5] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees", *Nature machine intelligence, 2(1)*, pp. 56–67, 2020.

[6] D. Dua and C. Graff, "UCI Machine Learning Repository", Irvine, CA: University of California, School of Information and Computer Science, 2019. [Online]. Available from: http://archive.ics.uci.edu/ml.