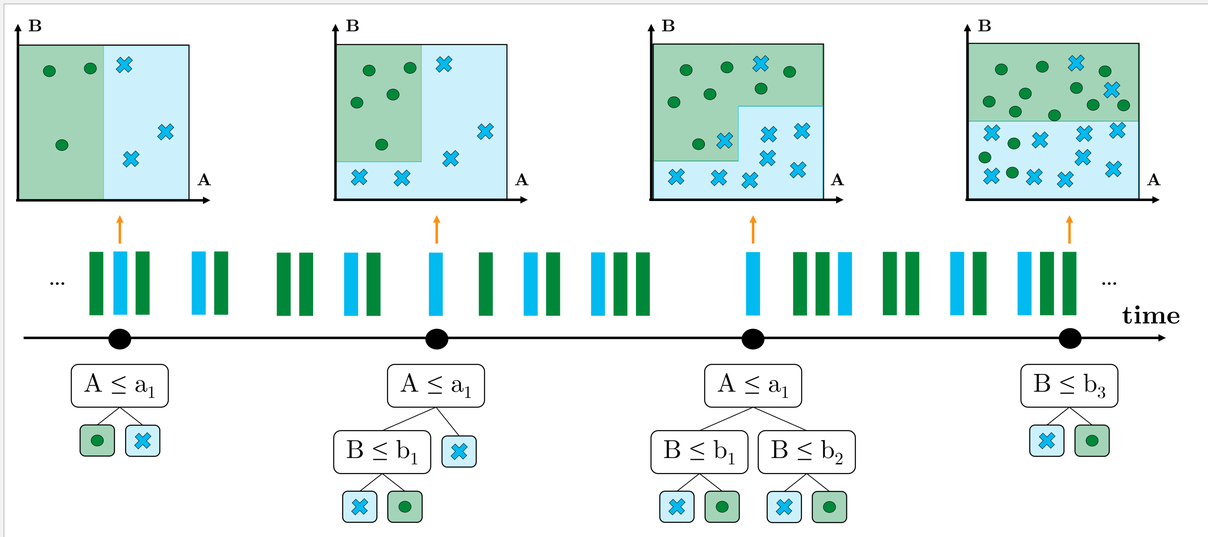
ESANN 2023

On Feature Removal for Explainability in Dynamic Environments

Fabian Fumagalli[1,*], Maximilian Muschalik[2,*], Eyke Hüllermeier[2], Barbara Hammer[1]
[1] Bielefeld University, Germany, [2] LMU Munich, Germany

TRR 318

The Problem: Changing Black Box Models



➥ Requires *efficient*, *any-time* ML models.

A Solution: Online Explanations

✓ **Methods:** First incremental, online XAI methods exist.

Global Feature Importance Methods in Dynamic Environments

Global Feature Effect Methods in Dynamic Environments

Incremental Permutation Feature Importance (iPFI) [2]

Computes global feature importance **incrementally** based on the well known **permutations tests**

Incremental Shapley Additive Global Explanation (iSAGE) [3]

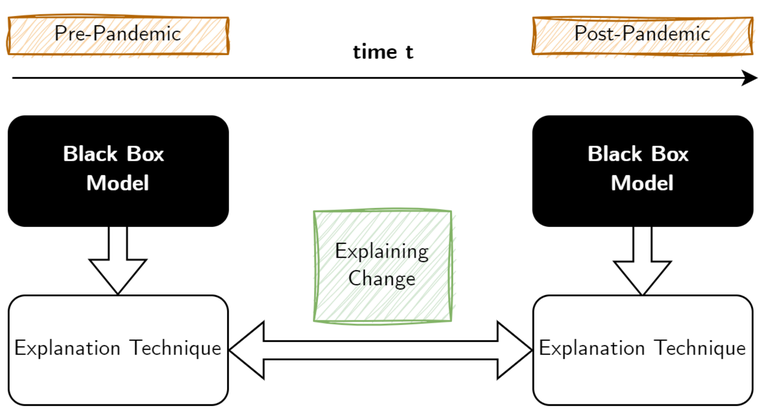
Computes global feature importance **incrementally** based on the **Shapley-based SAGE values**

Incremental Partial Dependence Plots (iPDPs) [4]

Computes global feature effects **incrementally** based on the established **PDPs** and **ICE curves**

➥ Requires *efficient*, *any-time* **feature removal** methods.

Prediction of Hospital Admission



Training of Hospital admission model on patient data.

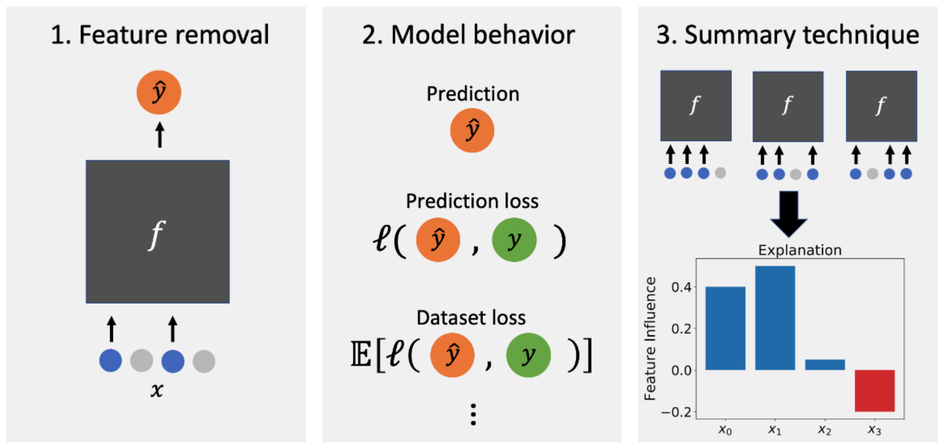
Concept Drift: Covid pandemic leads to changes in the model.

Background: Explaining by Removing [1]

Explaining by removing [1]

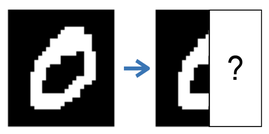
Given a model $f : \mathcal{X} \to \mathcal{Y}$ on a feature space $\mathcal{X}$ with features $D = \{1, \dots, d\}$, we define

- **Feature Removal:** A restricted model $F : \mathcal{X} \times \mathcal{P}(D) \to \mathcal{Y}$, where features in $\bar{S} := D \setminus S$ are removed.
- **Model Behavior:** A model property $\nu : \mathcal{P}(D) \to \mathcal{Y}$ that is based on the restricted model.
- **Summary Technique:** An aggregation method for different evaluations of ν.



Online Feature Removal Mechanisms

Original Data



Interventional Removal

Observational Removal

Images in taken with permission from Kristin Blesh's XAI'23 presentation on "Unfooling SHAP and SAGE: Knockoff Imputation for Shapley Values"

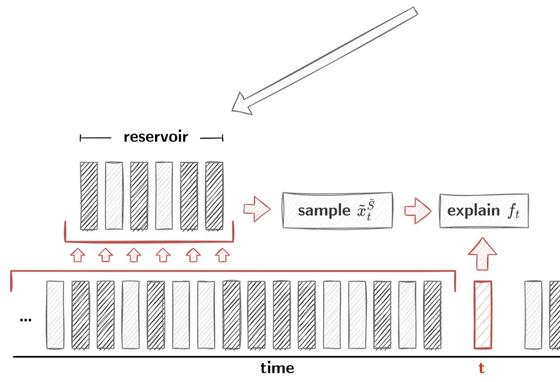
Interventional

restricted model

$$f^{\text{obs}}(x, S) := \mathbb{E}\left[f(x^{(S)}, X^{(\bar{S})})\right]$$

approximation

$$\hat{f}(x, S) := \frac{1}{M} \sum_{m=1}^{M} f(x^{(S)}, \tilde{x}_m^{(\bar{S})})$$

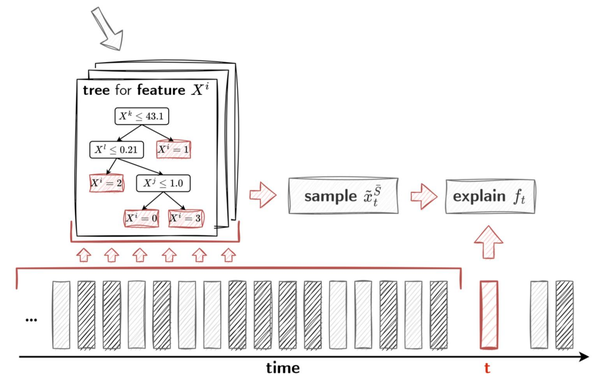


Reservoir Sampling

Observational

restricted model

$$f^{\text{obs}}(x, S) := \mathbb{E}\left[f(x^{(S)}, X^{(\bar{S})}) \mid X^{(S)} = x^{(S)}\right]$$



Incremental Partition Trees

Empirical Comparison

Agrawal Stream

Salary

$X^{\text{sal.}} \sim \text{unif}(20k, 150k)$

Age

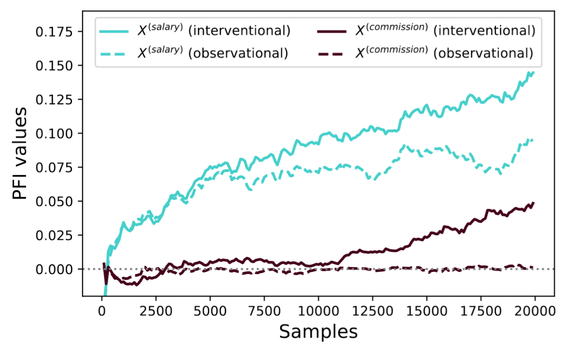
$X^{\text{age}} \sim \text{unif}(20, 80)$

Commission

$X^{\text{com.}} = \mathbf{1}(X^{\text{sal.}} \leq 75k) \cdot Z$ with $Z \sim \text{unif}(10k, 75k)$

Target

Grp A: ((**age** < 40) $\wedge$ (((50$K \leq$ **salary** $\leq$ 100K) ? (100$K \leq$ **loan** $\leq$ 300K) : (200$K \leq$ **loan** $\leq$ 400K)))) $\vee$
((40 $\leq$ **age** < 60) $\wedge$ (((75$K \leq$ **salary** $\leq$ 125K) ? (200$K \leq$ **loan** $\leq$ 400K) : (300$K \leq$ **loan** $\leq$ 500K)))) $\vee$
((**age** ≥ 60) $\wedge$ (((25$K \leq$ **salary** $\leq$ 75K) ? (300$K \leq$ **loan** $\leq$ 500K) : (100$K \leq$ **loan** $\leq$ 300K))))

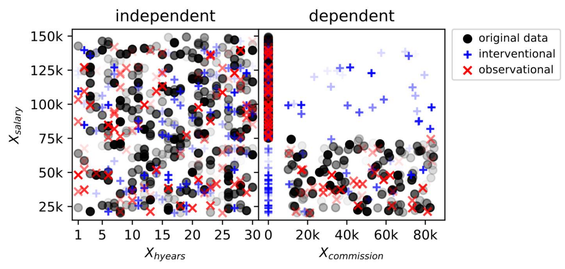


Observational Feature Removal (whole stream)

The **commission** feature receives **negligible importance,** as the information can be **inferred from salary**

Interventional Feature Removal (whole stream)

The **commission** feature receives **positive importance**, as the model **learned** to classify based **on salary and commission**



Gaussian

Multivariate Gaussian ($d = 4$)

$X_t \sim \mathcal{N}(\mathbf{0}, \Sigma_t)$

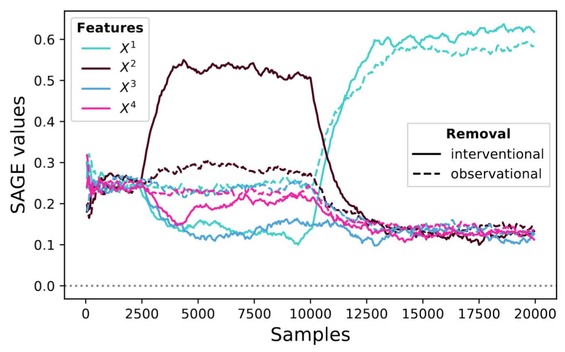
Target

$Y_t := \mathbf{1}(X_t^{(1)} > 0)$

Data Stream with Concept Drift

$$\Sigma_t = \begin{cases} 5 \cdot \mathbf{J} & \text{for } t < t_0 \quad \text{(features strongly correlated)} \\ \mathbf{I} & \text{for } t \geq t_0 \quad \text{(features uncorrelated)} \end{cases}$$

$$\mathbf{J} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



Observational Feature Removal (before concept drift)

All features receive **low importance**, as each feature can be **inferred from** the **remaining** ones

Interventional Feature Removal (before concept drift)

The **second feature** receives **high importance** as the model **learned** to classify based on the **second feature**

After Concept Drift (uncorrelated features)

The **int. and obs.** approach yield **similar results**, as feature are now uncorrelated. Both show that the **model learned** to classify based on the **first feature**

Open Source: iXAI

iXAI

docs passing | pypi v0.1.3 | status alpha | License MIT

Installation

```
pip install ixai
```

Quickstart

```
>>> for (n, (x, y)) in enumerate(stream, start=1)
...     accuracy.update(y, model.predict_one(x))   # inference
...     incremental_pfi.explain_one(x, y)          # explaining
...     model.learn_one(x, y)                      # learning
```

✓ iXAI currently contains four explanation methods: iPFI [2], iSAGE [3], iPDP [4], and MDI [7]
✓ iXAI allows for **interventional** and **observational** feature removal by means of different imputer systems.
❑ Looking for **collaborators**!

Conclusion

Observational Feature Removal

Includes the **dependencies** of features into the explanation. **Reveals** the **information** that a feature provides **to the model**. We **confirm** that this approach is **true to the data** [5].

Interventional Feature Removal

Breaks feature **dependencies** when computing the explanation. **Reveals** more accurately what the **model has learned**. We **confirm** that this approach is **true to the model** [5].

Future Work

➢ **Human-grounded evaluation**: Conduct user studies and investigate how to explain dynamic models **efficiently** and **effectively**.
➢ **Move to Higher Orders**: Move from feature importance to feature interactions.

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation
TRR 318 CONSTRUCTING EXPLAINABILITY
UNIVERSITÄT BIELEFELD
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
MCML Munich Center for Machine Learning

Contact

✉ maximilian.muschalik@lmu.de
✉ ffumagalli@techfak.uni-bielefeld.de

References

[1] Covert, I., Lundberg, S. M., & Lee, S.-I. (2021). Explaining by Removing: A Unified Framework for Model Explanation. Journal of Machine Learning Research, 22, 209:1-209:90. http://jmlr.org/papers/v22/20-1316.html

[2] Fumagalli, F., Muschalik, M., Hüllermeier, E., & Hammer, B. (2023). Incremental permutation feature importance (iPFI): Towards online explanations on data streams. Machine Learning. https://doi.org/10.1007/s10994-023-06385-y

[3] Muschalik, M., Fumagalli, F., Hammer, B., & Hüllermeier, E. (2023). iSAGE: An Incremental Version of SAGE for Online Explanation on Data Streams. In Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Proceedings, Part III (pp. 428-445). Springer. https://doi.org/10.1007/978-3-031-43418-1_26

[4] Muschalik, M., Fumagalli, F, Jagrani, R., Hammer, B., & Hüllermeier, E. (2023). iPDP: On Partial Dependence Plots in Dynamic Modeling Scenarios. In Explainable Artificial Intelligence - First World Conference, xAI 2023, Proceedings, Part I. Springer Cham.

[5] Chen, H., Janizek, J. D., Lundberg, S. M., & Lee, S.-I. (2020). True to the Model or True to the Data? CoRR, abs/2006.16234. [Preprint]. arXiv:2006.16234