Selected Challenges in Explaining Complex Models

Evan Krell

Based on work by Christoph Molnar [2]

innovation in COmputing REsearch lab (iCORE)

Papers & Resources

Paper [1] Murdoch, W. James, et al. "Definitions, methods, and applications in interpretable machine learning." Proceedings of the National Academy of Sciences 116.44 (2019): 22071-22080.

Paper [2] Molnar, Christoph, et al. "General pitfalls of model-agnostic interpretation methods for machine learning models." arXiv preprint arXiv:2007.04131 (2020).

Paper [3] McGovern, Amy, et al. "Making the black box more transparent: Understanding the physical implications of machine learning." Bulletin of the American Meteorological Society 100.11 (2019)

Video Peering Inside the Black Box if Machine Learning for Earth Science

- Part 1 [4] https://youtu.be/n8Lsz56_EDI (Amy McGovern)
- Part 2 [5] https://youtu.be/LTD0wLyMVE8 (Imme Ebert-Uphoff)

Course [6] Ryan Lagerquist & Imme Ebert-Uphoff. AI2ES/CIRA Short Course on Explainable Artificial Intelligence for Environmental Science
https://docs.google.com/document/d/1lqpABwDl3kPe6ThE-NIDR64PimnltJEuKNkysDZuWKQ

Book [7] Christoph Molnar: Interpretable Machine Learning
christophm.github.io/interpretable-ml-book/

Code [8] Evan Krell: PartitionShap demo notebooks
https://github.com/conrad-blucher-institute/partitionshap-multiband-demo

Pitfall [one-fits-all]: Confusing feature importance & feature effect

- Typical XAI output: each feature assigned score
- But scores mean different things → must interpret correctly
- Two major categories:
  1. **Feature importance:** how much did feature help reduce error?
  2. **Feature effect:** how much did feature influence prediction?
- **Q:** Model used the feature (effect) → because it was helpful? (importance)
- **A:** Not necessarily ...
  Consider: random weights → nothing important → still used *something*

FogNet example

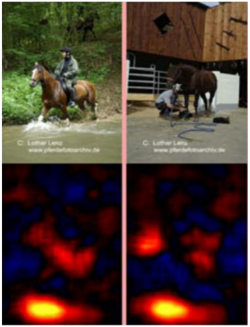
- Train fog prediction model
- Different types of fog (advection & radiation)
- Forecasters use different strategies for each
- Radiation fog less common → underrepresented in training data
- Overall, FogNet has high performance ... but radiation has misses
- Use SHAP (feature effect) to see what model uses for both cases
- SHAP → both cases heavily influenced by feature X
- Know that it is not reliable for predicting radiation fog
- Feature X has **high effect**, but hurts the accuracy
  → **low importance** for radiation cases

Pitfall: bad model generalization

- XAI has been used to:
  - Investigate the model (relationship between data & prediction)
  - Investigate the data (relationship between data & target)
- XAI directly lets us do (1),
  with accurate model can learn something about (2)
- If model underfits → wrong conclusions about data-generating process

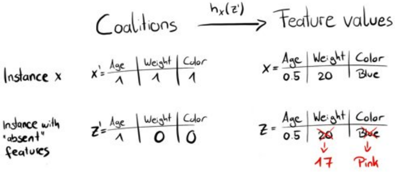
Molnar solution: ensure test set accuracy before attempting (2)
However: even if appears to generalize, should be cautious using XAI to say that a model has learned novel scientific knowledge → use to form hypothesis

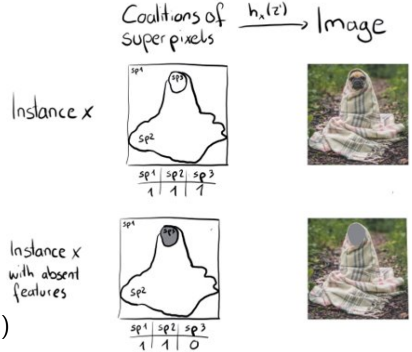
Example: horse



- Suppose watermark present in test data
- High test accuracy → *might say model generalizes*
- So that means the watermark is a newly discovered means for identifying horses in the real world?
- More challenging to decide when the XAI shows a failure case that is not comically obvious

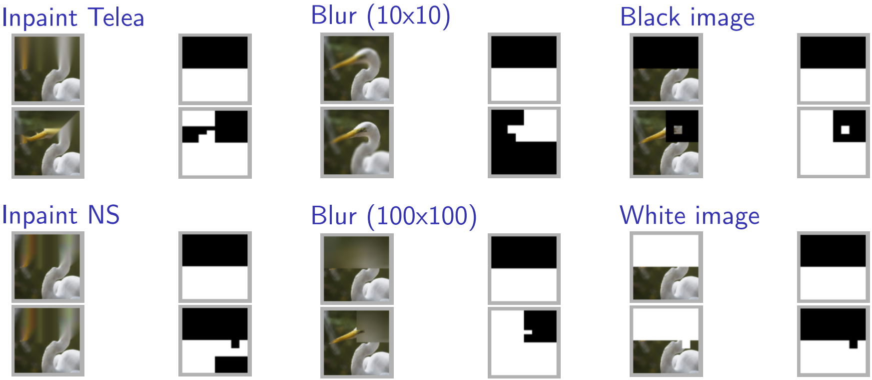
- Perturbation-based $\rightarrow$ test model by replacing values with **something**
- Examples: random values, a constant, other dataset values





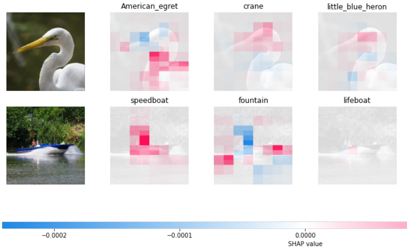
- Model not trained on this (unrealistic)
- Output maybe sensitive to these out-of-sample instances
- Predicts in response to **weird input** instead of what you want: prediction without that feature
- **Molnar solution:** grouping dependent features might help

- Not just a theoretical concern
- Example: SHAP for images: many feature masking options
- But which to choose?

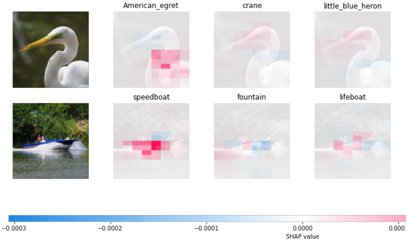


- Now let's see how the choice influences the explanation ...

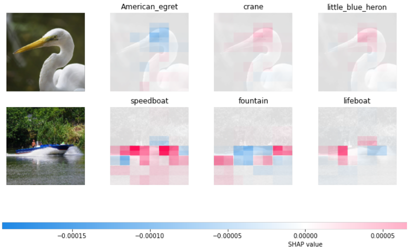
Inpaint Telea



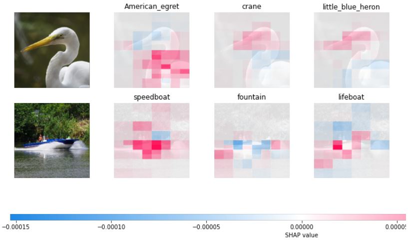
Gray image



Blur (10x10)

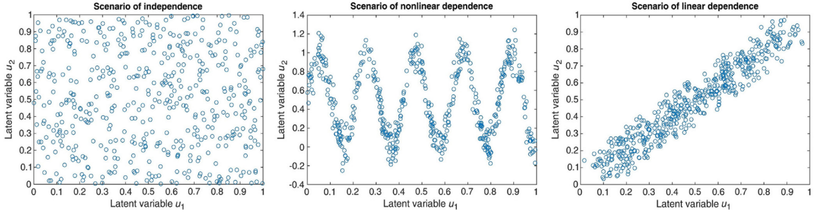


Blur (100x100)



Pitfall [Dependence]: Confusing linear correlation with general dependence

- Dependencies → potentially misleading XAI output
- Pearson Correlation Coefficient: linear correlation
  - ∼ 0 → feature may still be dependent



Hu, Wenxing, et al. "Distance canonical correlation analysis with application to an imaging-genetic study."

Journal of Medical Imaging 6.2 (2019): 026501.

- **Molnar solution:** apply methods for detecting non-linear dependency

  kernel-based methods, HSIC, information theory-based measures
- **My Q:** how well do these scale to huge input data?

What about raster data?

- Substantial spatial autocorrelation
- Often: correlation across channels

  wind components, atmospheric profiles, . . .
- Apply spatial statistics?

hyperspectral

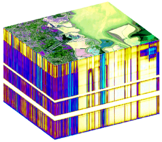
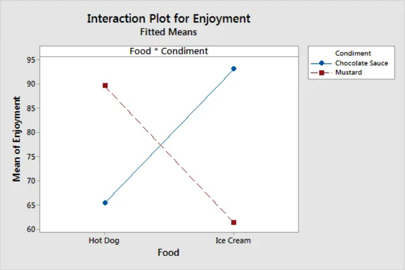


image: RD world

Pitfall [Interactions]: Misleading feature effects due to aggregation



- Feature interaction: X_i effect on prediction is influenced by X_j
- Global explanation $\rightarrow$ **effects averaged out**

Image: statisticsbyjim.com/regression/interaction-effects

- Expect effect of feature X_i to vary across samples
- Want to detect overall patterns like "strong X_i $\rightarrow$ predict tornado
- XAI method:*permute X_i values $\rightarrow$ monitor change in y*
- But when X_i's impact on y is because of another feature X_j, hard to detect
- Could try: *permute X_i & X_j $\rightarrow$ monitory change in y*
- But what if X_i interacts with X_j that interacts with X_z that interacts ...

Molnar: plot the interaction effects, but be aware:

- Fails to reveal the type of underlying interaction
- Fail to reveal higher-order interactions

Pitfall [Interactions]: Failing to separate main from interaction effects

- This pitfall is subtly different from the last
- Last slide: effect averaged out by interactions
- Here: effected reported by XAI method includes **aggregated effects**
- Some of contribution should really be assigned to the interacting feature

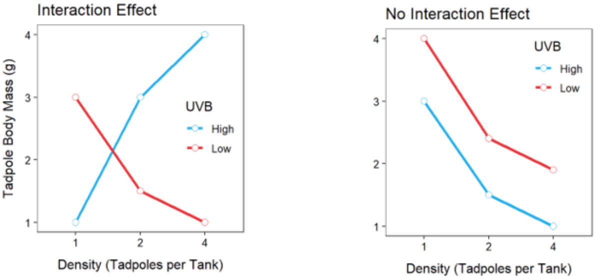


Image: http://derekogle.com/Book207/ANOVA2Foundations1.html

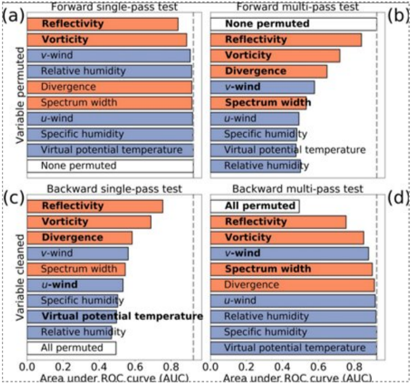
Molnar solution (?)

- Functional ANOVA $\rightarrow$ decompose into main & interaction effects
- XAI method: SHAP interaction values
  https://christophm.github.io/interpretable-ml-book/shap.html#shap-interaction-values
- Struggles: **higher order interactions**, **interactions of dependent features**

Pitfall: Ignoring model & approximation uncertainty

- XAI methods rarely report uncertainty
- Global methods often aggregate many local methods
- Global explanation could reflect outliers
- Feature scores, ranking may not be statically significant

Solution: significant testing



bold $\rightarrow$ predictor significantly more important than one below

- Ryan Lagerquist: boostrapping XAI [6]
- Feature X is top feature, but is the difference between top two significant?
- Video: https://drive.google.com/file/d/1aa9sOeRijH22p57bALL1tmoOKtvk5Akl/
- Notebook: https://colab.research.google.com/drive/1IaeeSE5J6r-eOOh_gmjNVqd3Bg_M6DSo

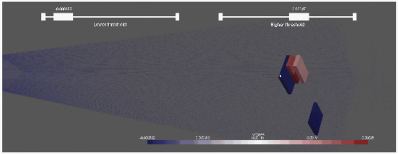
Pitfall: Interpreting high-dimensional XAI outputs

"Applying IML methods naively to high-dimensional datasets . . . overwhelming and high-dimensional IML output, which impedes human analysis." - [2]

- Molnar's general solution: reduce feature dimensionality → apply XAI
- Group features and perturb combined
- Issues:
  - "feature groups are not meaningful . . . interpretations . . . purposeless" - [2]
  - "'How a group of features influences . . . prediction' . . . unanswered" - [2]

For raster: maybe our 3D tool will help?

- FogNet's 384 channels
- Could plot each channel as own explanation heatmap
- But might miss across-channel patterns
- Interactive 3D visualization tool
  Video: https://youtu.be/kNFY6ff996E



Are complex models too hard to explain?

- Molnar summary: know **everything** about your data to trust XAI results
- But path unclear, unfeasible (?) for highly complex models
- AI2ES: tackling complex, nonlinear models with dependencies, interactions
- Seems for every method, papers pointing out theoretical & practical flaws

My perspective on XAI feasibility

- XAI actually did **reveal strategies** for horse & wolf classification
- Nonlinear CNNs predict despite individual pixels meaningless
- **Critic:** given XAI problems → can't trust those strategies are real
- Do we accept that? Probably not. Why? → random chance seems absurd

Key to using XAI: (non) randomness

- Suppose that we:
  - Run multiple XAI methods → highlight watermark effect
  - Add watermark to non-horse images → now classified as horse
  - Remove watermark from horses → XAI now highlights *horse parts*
- Using XAI explanations & other verification → **build compelling evidence**
- Pitfall awareness can keep us from using XAI naively
- But careful use of XAI can provide strong evidence of model insights

  triangulation, multiple feature groupings, sanity checks, significance testing, . . .