Relative Feature Importance

ICPR 2020, 13th of January 2021

Gunnar König[1,2,*], **Christoph Molnar**[1], **Bernd Bischl**[1], **Moritz Grosse-Wentrup**[2]

[1]LMU Munich, [2]University of Vienna

[*]g.koenig.edu@pm.me

Background and Motivation

- Global, model-agnostic Feature Importance: How relevant is feature X_j for the model's performance?
  - based on comparing $\mathcal{R}(Y, f(X_R, \tilde{X}_j))$ (risk under perturbation) with $\mathcal{R}(Y, f(X_R, X_j))$ (risk on the test set)
- Under dependent features, two slightly different notions:
  - Which variables' information is being used by the model?
  - Via which features does useful information enter the model?

Background and Motivation

- emergence of several methods with different semantics
  - Permutation feature importance (PFI): regards X_j in *isolation* (Breiman, 2001; Fisher et al., 2019)
  - Conditional Feature Importance (CFI): regards X_j in relation to *all covariates* (Strobl et al., 2008; Fisher et al., 2019; Molnar et al., 2020)
  - Shapley Additive Global Explanations (SAGE): fair attribution, in relation to *all covariates* (Covert et al., 2020)
- often, importance in relation to a *specific* subset G of interest

Problem

Model Inference and Model audit

Nuru:
Importance of X_j that *cannot* be attributed to G?

(Relative Importance)

Claudio:
Importance of X_j that *can* be attributed to G?

(Indirect Influence)

Background

The need for Relative Feature Importance

- D: all features ($Y \notin D$),
- $R := D \backslash \{j\}$,
- G: arbitrary set,
- $\underline{R} := R \backslash G$

	PFI	CFI
perturbation	$\tilde{X}_j \sim P(X_j)$	$\tilde{X}_j^R \sim P(X_j \mid X_R)$
semantics	overall	unique
Nuru happy?	no	no
Claudio happy?	no	no

Background

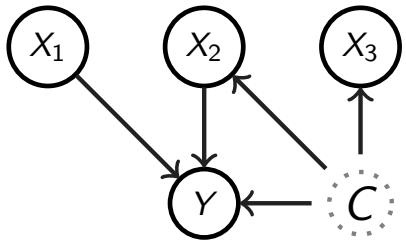
The need for Relative Feature Importance

- D: all features ($Y \notin D$),
- $R := D \backslash \{j\}$,
- G: arbitrary set,
- $\underline{R} := R \backslash G$

	PFI	CFI	RFI
perturbation	$\tilde{X}_j \sim P(X_j)$	$\tilde{X}_j^R \sim P(X_j \mid X_R)$	$\tilde{X}_j^G \sim P(X_j \mid X_G)$
semantics	overall	unique	relative to G
Nuru happy?	no	no	yes
Claudio happy?	no	no	yes

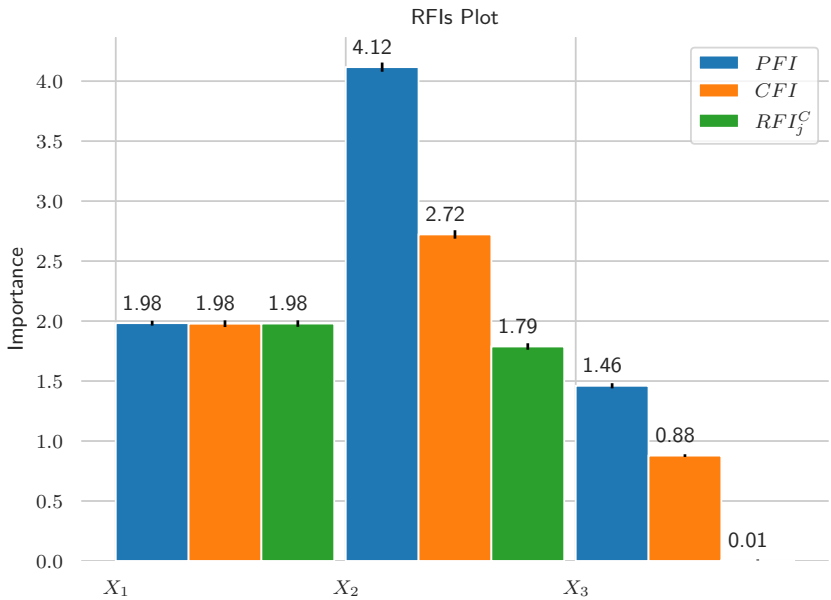
Relative Feature Importance: Interpretation

Example 1: Model Inference



linear gaussian data, OLS linear regr

$$f(x_1, x_2, x_3) = 1.0x_1 + 1.17x_2 + 0.67x_3$$



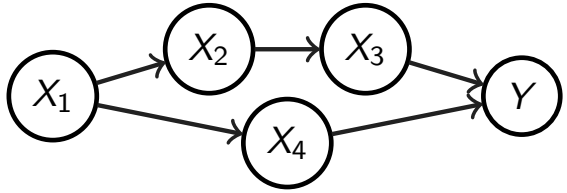
Nuru:

Importance of X_j that *cannot* be attributed to C?

Relative Feature Importance: Interpretation

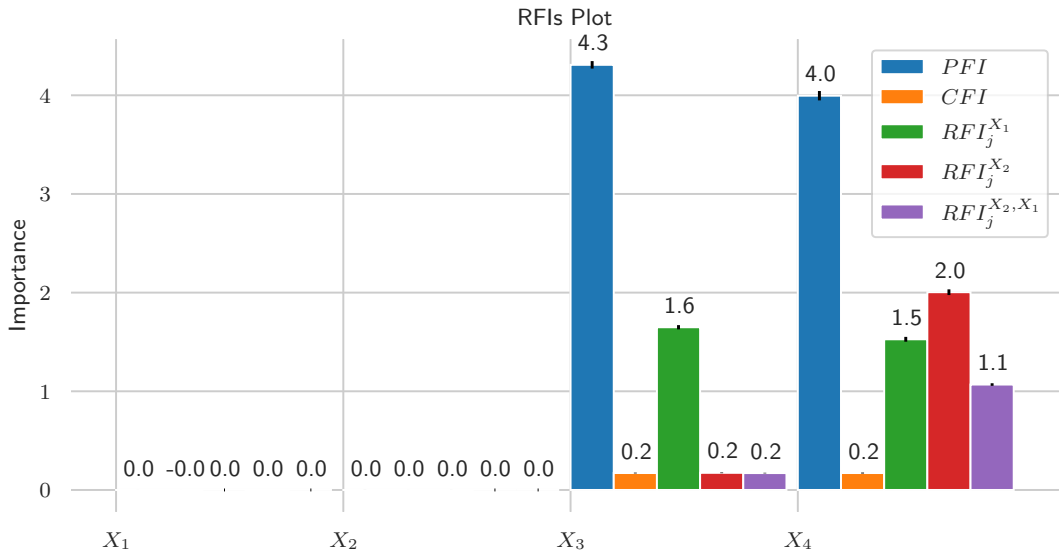
Example 2: Model Audit

Claudio:
Importance of X_j that *can* be attributed to X_1?



linear gaussian data, OLS regr

$$f(x_1, \ldots, x_4) = 0.01x_2 + 1.00x_3 + 1.00x_4$$



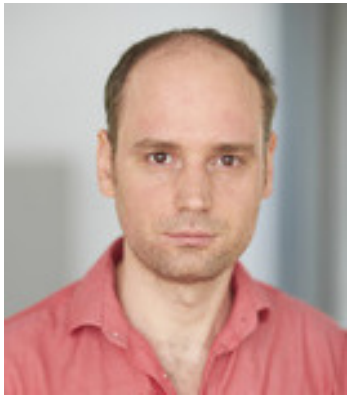
Summary

- Permutation Feature Importance (PFI) and Conditional Feature Importance (CFI) have extreme implicit definitions of relevance.
- We propose Relative Feature Importance (RFI), a generalization of PFI and CFI that provide more nuanced insight into model and data.
- Theoretical results on how to interpret RFI in our paper.
  - We characterize RFI by how the method behaves in its context. This context involves both model and data.
- A `python` package will soon be available on my Github page.

Team

Gunnar König [1,2]

Christoph Molnar [2]



Bernd Bischl [2]

Moritz Grosse-Wentrup [1]

University of Vienna[1]

LMU Munich[2]

References

Breiman, L. (2001). Random forests. *Machine Learning*, pages 1–122.

Covert, I., Lundberg, S., and Lee, S.-I. (2020). Understanding Global Feature Contributions Through Additive Importance Measures. *arXiv preprint arXiv:2004.00668*.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Molnar, C., König, G., Bischl, B., and Casalicchio, G. (2020). Model-agnostic feature importance and effects with dependent features-a conditional subgroup approach. *arXiv preprint arXiv:2006.04628*.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9:1–11.

Backup

Relative Feature Importance: Definition

Definition 1: *RFI w.r.t. G with $Y \notin G$ and a fixed model f is defined as*

$$RFI_j^G := \tilde{\mathcal{R}}^{j|G} - \mathcal{R},$$

with $\tilde{\mathcal{R}}^{j|G} := \mathcal{R}(Y, f(X_R, \tilde{X}_j^G))$ and $\mathcal{R} = \mathcal{R}(Y, f(X_j, X_R))$. The replacement variable has to satisfy

- $\tilde{X}_j^G \sim P(X_j|X_G)$ *and*
- $\tilde{X}_j^G \perp\!\!\!\perp (X_{\underline{R}}, Y)|X_G$.

Relative Feature Importance: Interpretation

Theorem 1 and 2

Theorem 1: *If* $RFI_j^G \neq 0$ *then*

- $X_j \not\perp\!\!\!\perp (Y, X_{\underline{R}}) | X_G$ *in the underlying distribution (data level)*
- $\tilde{X}_j \not\perp\!\!\!\perp \hat{Y} | X_R$ *w.r.t. the interventional distribution* $P(X_j|X_G)P(X_G, X_{\underline{R}}) > 0$ *(model level)*

Theorem 2: *If the difference* $\Delta RFI_j^{G \to G \cup N} = RFI_j^G$ - $RFI_j^{G \cup N} \neq 0$, *then* $X_j \not\perp\!\!\!\perp X_N | X_G$.