Fooling Partial Dependence via Data Poisoning

Hubert Baniecki, Wojciech Kretowicz, Przemyslaw Biecek

{hubert.baniecki.stud,wojciech.kretowicz.stud,przemyslaw.biecek}@pw.edu.pl

MI[2].AI, Warsaw University of Technology, Poland

ECML PKDD, Grenoble, France, September 19 – 23, 2022

Introduction: Why?

1. We highlight that Partial Dependence can be **manipulated** with adversarial data perturbations.
2. We introduce a novel concept of using a **genetic algorithm** for fooling model explanations of **any black-box**. We use a gradient algorithm to perform it efficiently for neural networks.
3. Experiments on various models and their sizes shows the **hidden debt of model complexity** related to explainable machine learning.

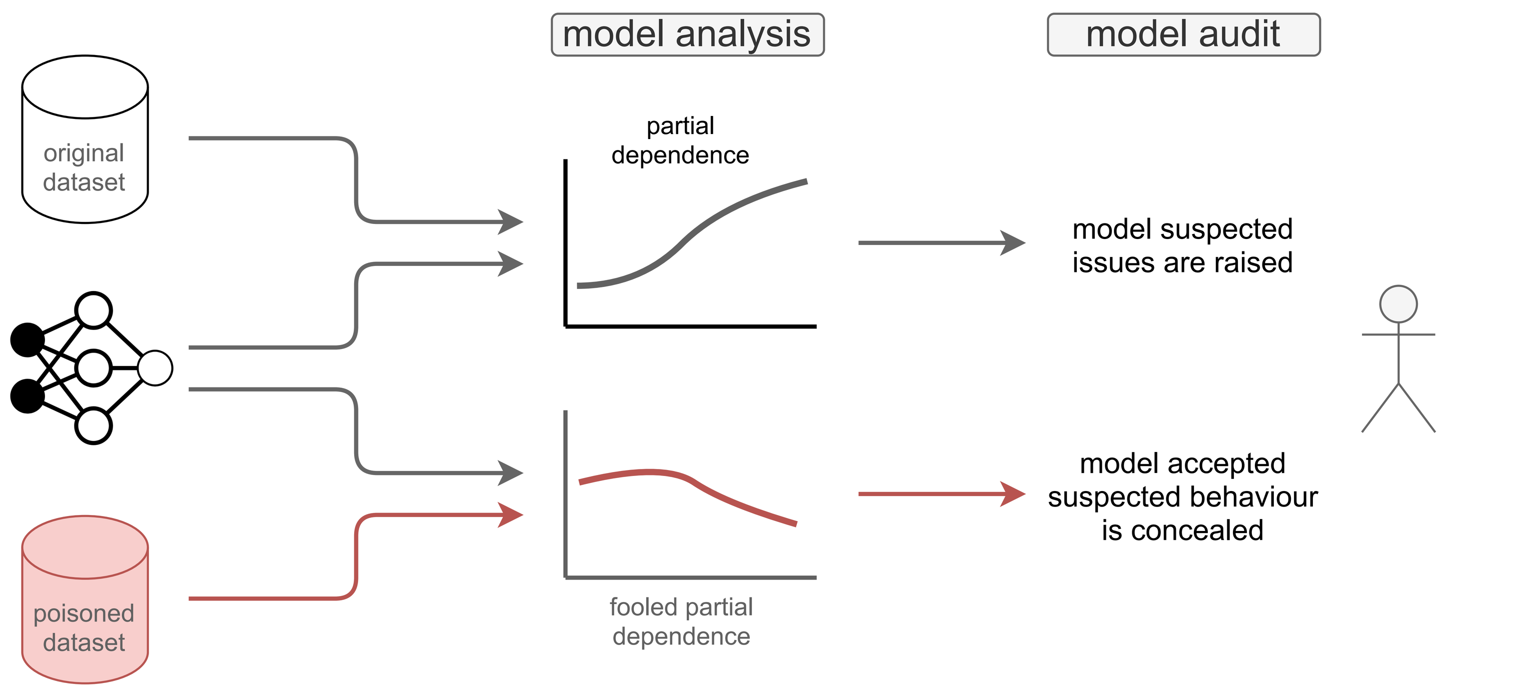

Figure 1: **Framework for fooling model explanations via data poisoning.** The **red** color indicates the adversarial route, a potential security breach, which an attacker may use to manipulate the explanation. Researchers could use this method to provide a misleading rationale for a given phenomenon, while auditors may purposely conceal the suspected, e.g. biased or irresponsible, reasoning of a black-box.

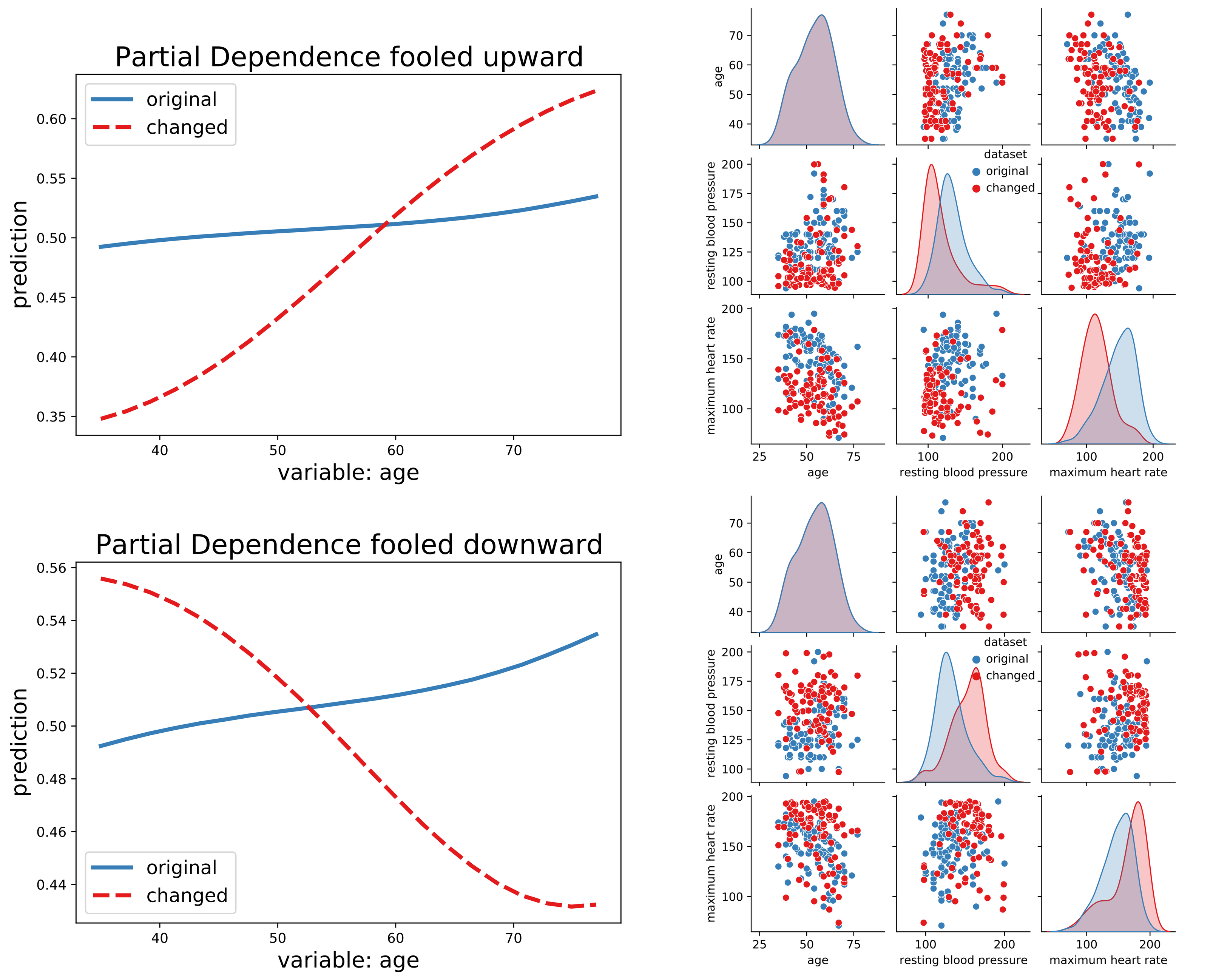

Figure 2: Partial Dependence of `age` in the SVM model prediction of a heart attack (class 0). **Left:** Two manipulated explanations suggest an increasing or decreasing relationship between `age` and the predicted outcome depending on a desired outcome. **Right:** Distribution of the explained variable `age` and the two poisoned variables from the data, in which the remaining ten variables attributing to the explanation remain unchanged. The mean of the variables' Jensen-Shannon distance equals only 0.027 in the upward scenario and 0.021 in the downward scenario.

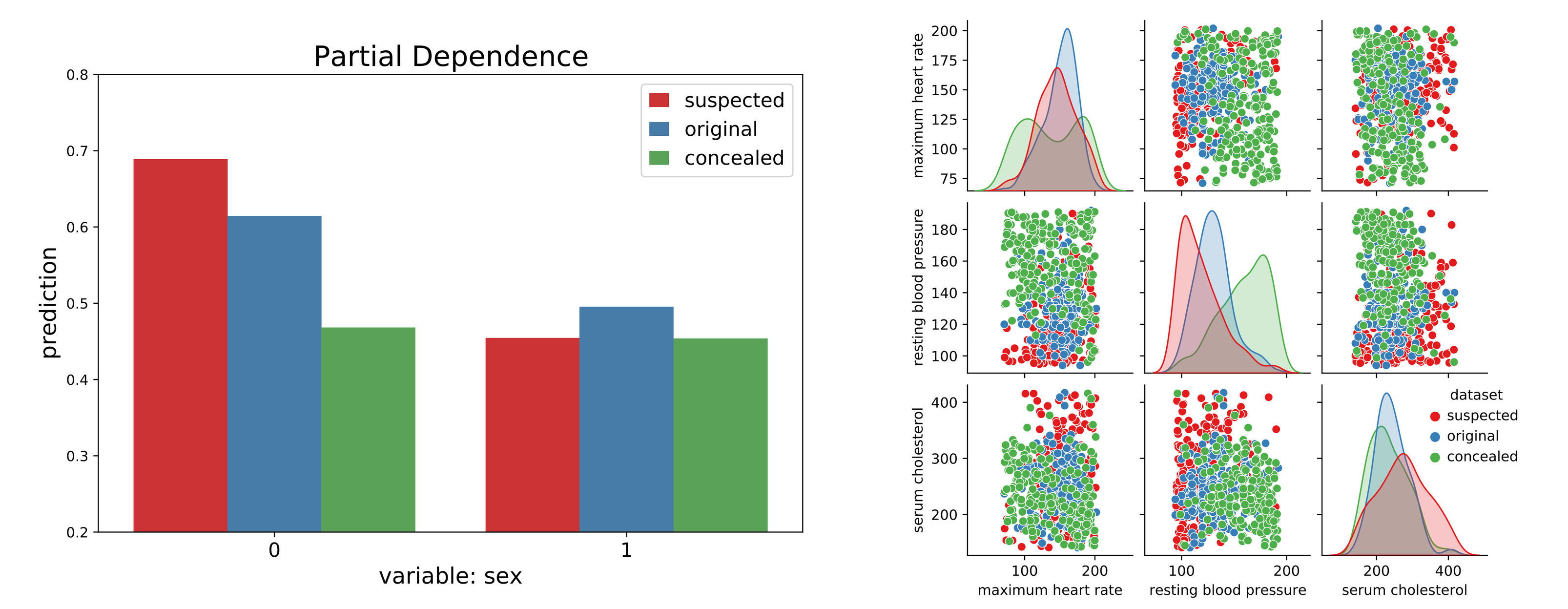

Figure 3: Partial Dependence of `sex` in the SVM model prediction of a heart attack (class 0). **Left:** Two manipulated explanations present a suspected or concealed variable contribution into the predicted outcome. **Right:** Distribution of the three poisoned variables from the data, in which `sex` and the remaining nine variables attributing to the explanation remain unchanged. The mean of the variables' J-S distance equals only 0.023 in the suspected scenario and 0.026 in the concealed scenario.

Methods

We iteratively change X with either:

- **Genetic-based** model-agnostic algorithm that does not make any assumption about the structure of model and explanation.
- **Gradient-based** algorithm designed for models with differentiable outputs, e.g. neural networks [2, 4].

There are two possible fooling strategies:

- **Targeted attack** changes the dataset to achieve the closest explanation result to the predefined desired function [2, 4]

$$\mathcal{L}^{\mathcal{PD},\,t}(X) = \|\mathcal{PD}_c(X) - T\|.$$

- **Robustness check** aims for the most distant model explanation from the original one X'

$$\mathcal{L}^{\mathcal{PD},\,r}(X) = -\|\mathcal{PD}_c(X) - \mathcal{PD}_c(X')\|.$$

Benchmark results

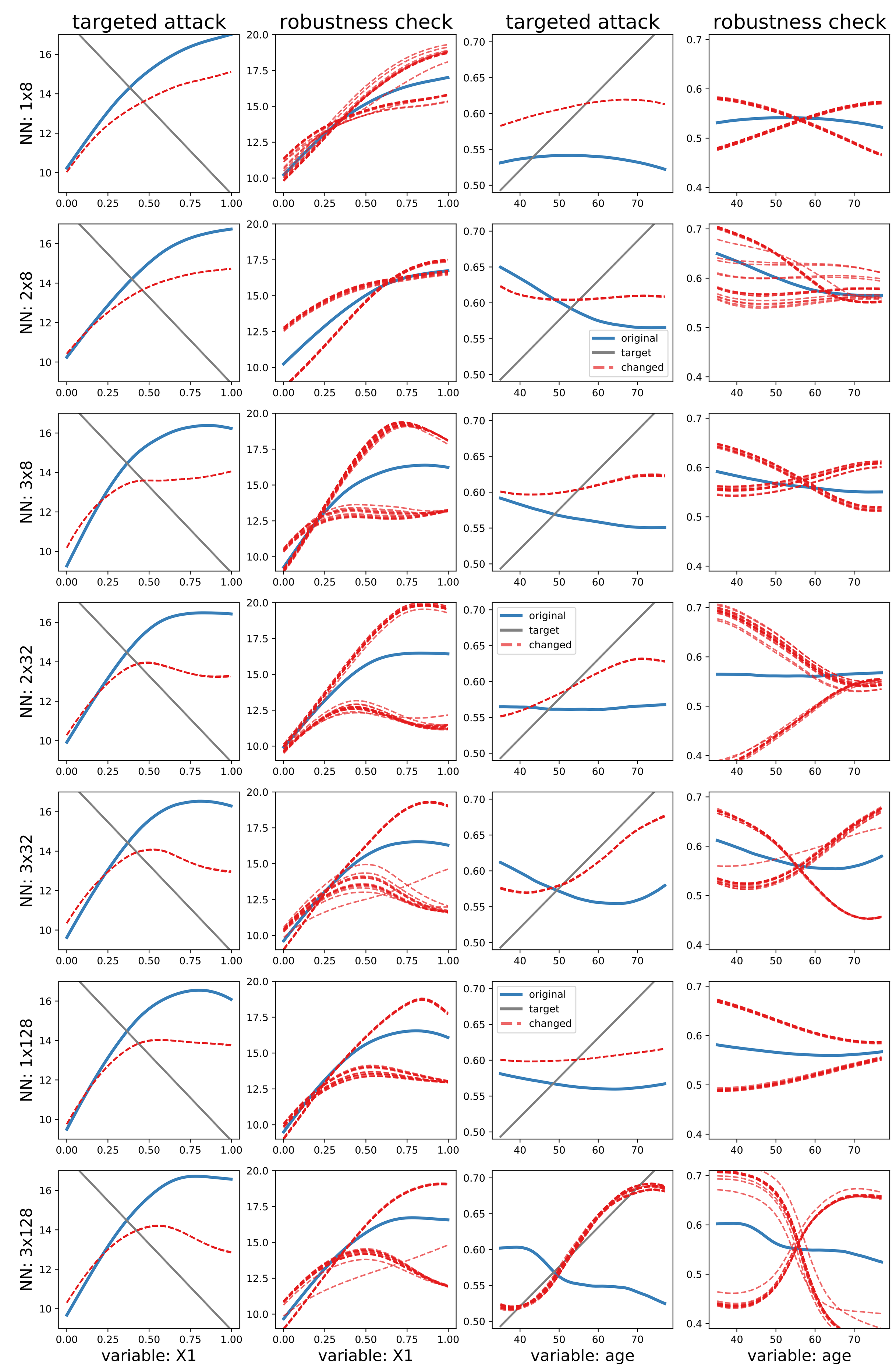

Figure 4: Fooling Partial Dependence of neural network models (**rows**) fitted to the `friedman` and `heart` datasets (**columns**). We performed multiple randomly initiated gradient-based fooling algorithms on the explanations of variables X_1 and `age` respectively. The **blue** line denotes the original explanation, the **red** lines are the fooled explanations, and in the targeted attack, the **grey** line denotes the desired target. **We observe that the explanations' vulnerability greatly increases with model complexity.** Interestingly, the algorithm seems to converge to two contrarary optima when no target is provided.

Task \ Model	LM	RF	GBM	DT	KNN	NN	SVM
`friedman`	$0_{\pm 0}$	$152_{\pm 76}$	$127_{\pm 71}$	$332_{\pm 172}$	$164_{\pm 61}$	$269_{\pm 189}$	$576_{\pm 580}$
`heart`	$2_{\pm 3}$	$20_{\pm 5}$	$77_{\pm 28}$	$798_{\pm 192}$	$133_{\pm 21}$	$501_{\pm 52}$	$451_{\pm 25}$

Task	Model \ Trees	10	20	40	80	160	320
`friedman`	GBM	$57_{\pm 12}$	$114_{\pm 20}$	$157_{\pm 37}$	$176_{\pm 20}$	$189_{\pm 8}$	$210_{\pm 9}$
	RF	$233_{\pm 22}$	$219_{\pm 25}$	$219_{\pm 9}$	$201_{\pm 23}$	$216_{\pm 13}$	$209_{\pm 15}$
`heart`	GBM	$1_{\pm 0}$	$3_{\pm 1}$	$29_{\pm 4}$	$70_{\pm 24}$	$152_{\pm 56}$	$321_{\pm 95}$
	RF	$62_{\pm 7}$	$55_{\pm 3}$	$29_{\pm 9}$	$21_{\pm 6}$	$14_{\pm 5}$	$13_{\pm 2}$

Table 1: Scaled attack loss values of the robustness checks for PD of various machine learning models (**top**), and complexity levels of tree-ensembles (**bottom**). We perform the fooling 6 times and report the mean ± sd. **We observe that the explanations' vulnerability increases with GBM complexity.**

Partial Dependence (plot, profile, PDP) [1, 3] for model f and variable c in a random vector $\mathcal{X}$ is defined as $\mathcal{PD}_c(\mathcal{X},z) := E_{\mathcal{X}_{-c}}\left[f(\mathcal{X}^{c|=z})\right]$, where $\mathcal{X}^{c|=z}$ is $\mathcal{X}$ with the c-th variable replaced by z. $\mathcal{X}_{-c}$ is the distribution of $\mathcal{X}$ with the c-th variable set to a constant. **PD estimator** for dataset X and variable c is given by $\widehat{\mathcal{PD}_c}(X,z) := \frac{1}{N}\sum_{i=1}^{N} f\left(x_i^{c|=z}\right)$.

References

[1] Biecek, P., Burzykowski, T.: Explanatory Model Analysis. Chapman and Hall/CRC (2021)
[2] Dombrowski, A.K., et al.: Explanations can be manipulated and geometry is to blame. In: NeurIPS (2019)
[3] Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics **29**(5), 1189–1232 (2001)
[4] Heo, J., Joo, S., Moon, T.: Fooling Neural Network Interpretations via Adversarial Model Manipulation. In: NeurIPS (2019)

Paper & Contact info

tinyurl.com/ECML22

hbaniecki.com

Acknowledgements

This work was financially supported by the NCN OPUS grant no. 2017/27/B/ST6/0130 and SONATA BIS grant no. 2019/34/E/ST6/00052.