Fooling Partial Dependence via Data Poisoning

Hubert Baniecki, Wojciech Kretowicz, Przemyslaw Biecek

MI^2.AI, Warsaw University of Technology, Poland

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Grenoble, France, September 19–23, 2022

Introduction: Why?

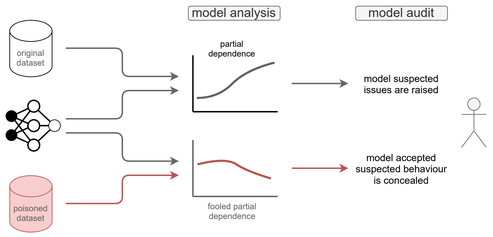

Figure 1: **Framework for fooling model explanations via data poisoning.** The red color indicates the adversarial route, a potential security breach, which an attacker may use to manipulate the explanation. Researchers could use this method to provide a misleading rationale for a given phenomenon, while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

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

Partial Dependence plot [4], profile [1], PDP [5] for model f and variable c in a random vector $\mathcal{X}$:

$$\mathcal{PD}_c(\mathcal{X}, z) \coloneqq E_{\mathcal{X}_{-c}}\left[f(\mathcal{X}^{c|=z})\right],$$

where $\mathcal{X}^{c|=z}$ stands for random vector $\mathcal{X}$, where c-th variable is replaced by value z. By $\mathcal{X}_{-c}$, we denote distribution of random vector $\mathcal{X}$ where c-th variable is set to a constant. The standard estimator of PD for data X is given as

$$\widehat{\mathcal{PD}}_c(X, z) \coloneqq \frac{1}{N}\sum_{i=1}^{N} f\left(X_i^{c|=z}\right),$$

where X_i is the i-th row of the matrix X.

Examples

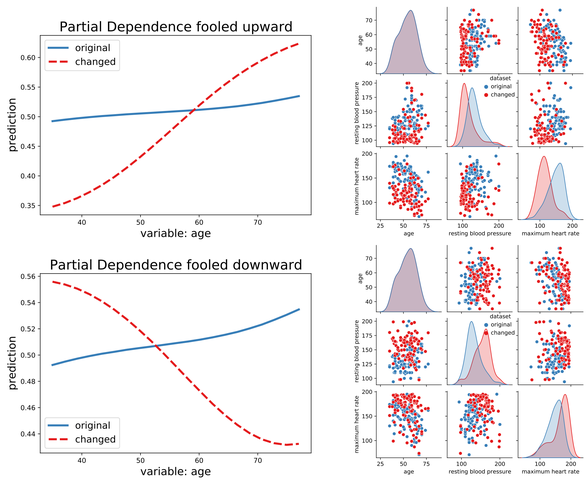

Figure 2: Partial Dependence of age in the SVM model prediction of a heart attack (class 0). **Left:** Two manipulated explanations suggest an increasing or decreasing relationship between age and the predicted outcome depending on a desired outcome. **Right:** Distribution of the explained variable 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, which might seem like an insignificant change of the data distribution.

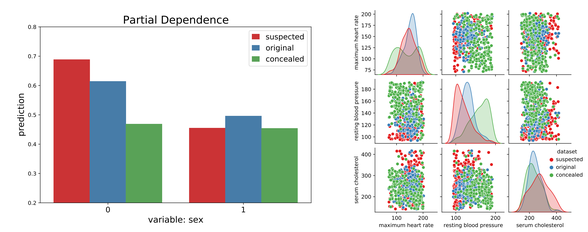

Figure 3: Partial Dependence of sex in the SVM model prediction of a heart attack (class 0). **Left:** Two manipulated explanations present a suspected or concealed variable contribution into the predicted outcome. **Right:** Distribution of the three poisoned variables from the data, in which sex and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

Benchmarks

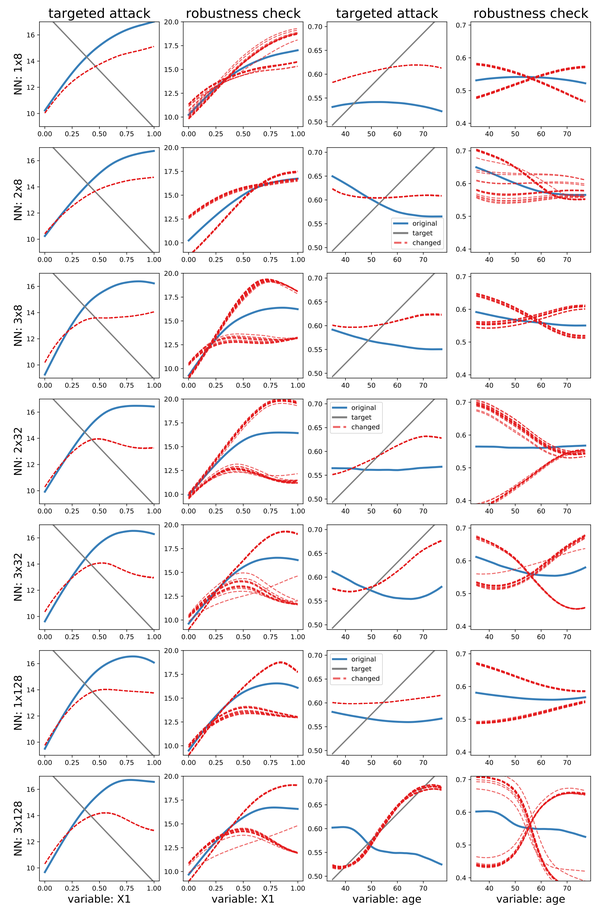

Figure 4: Fooling Partial Dependence of neural network models (**rows**) fitted to the `friedman` and `heart` datasets (**columns**). We performed multiple randomly initiated gradient-based fooling algorithms on the explanations of variables X_1 and age respectively. The blue line denotes the original 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.

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, 3].

There are two possible fooling strategies:

- **Targeted attack** changes the dataset to achieve the closest explanation result to the predefined desired function [3, 6]

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

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

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

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.**

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 2: Scaled attack loss values of the robustness checks for PD of various ReLU neural networks. We add additional noise variables to the data before model fitting, e.g. `friedman`+2 denotes the referenced dataset with 2 additional variables sampled from the normal distribution. **We observe that the explanations' vulnerability greatly increases with task complexity.**

Task \ NN	1×8	2×8	3×8	2×32	3×32	1×128	3×128
`friedman`	$25_{\pm 3}$	$33_{\pm 0}$	$75_{\pm 24}$	$100_{\pm 32}$	$98_{\pm 42}$	$54_{\pm 15}$	$97_{\pm 50}$
`friedman`+1	$31_{\pm 2}$	$40_{\pm 4}$	$50_{\pm 9}$	$106_{\pm 40}$	$115_{\pm 44}$	$57_{\pm 15}$	$114_{\pm 55}$
`friedman`+2	$34_{\pm 1}$	$40_{\pm 10}$	$50_{\pm 22}$	$106_{\pm 52}$	$115_{\pm 50}$	$50_{\pm 15}$	$137_{\pm 66}$
`friedman`+4	$46_{\pm 6}$	$33_{\pm 0}$	$83_{\pm 8}$	$145_{\pm 31}$	$163_{\pm 27}$	$40_{\pm 5}$	$140_{\pm 58}$
`friedman`+8	$71_{\pm 9}$	$47_{\pm 3}$	$89_{\pm 15}$	$204_{\pm 25}$	$176_{\pm 25}$	$39_{\pm 6}$	$156_{\pm 34}$

References

[1] Biecek, P., Burzykowski, T.: Explanatory Model Analysis. Chapman and Hall/CRC (2021)
[2] Dimanov, B., Bhatt, U., Jamnik, M., Weller, A.: You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods. In: AAAI SafeAI (2020)
[3] Dombrowski, A.K., Alber, M., Anders, C., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame. In: NeurIPS (2019)
[4] Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics **29**(5), 1189–1232 (2001)
[5] Greenwell, B.M.: pdp: An R Package for Constructing Partial Dependence Plots. The R Journal **9**(1), 421–436 (2017)
[6] Heo, J., Joo, S., Moon, T.: Fooling Neural Network Interpretations via Adversarial Model Manipulation. In: NeurIPS (2019)

Check out the paper on arXiv!

arXiv:2105.12837

hbaniecki.com

MI2DataLab/fooling-partial-dependence

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.