Fool SHAP with Stealthily Biased Sampling

[1]Gabriel Laberge [2]Ulrich Aïvodji [3]Satoshi Hara
[4]Mario Marchand [1]Foutse Khomh

[1]Polytechnique Montréal

[2]École de Technologie Supérieure

[3]Osaka University

[4]Université Laval à Québec

6 avril 2023

DEpendable & Explainable Learning

I have a private dataset $D = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$ and a black-box $f : \mathcal{X} \to [0, 1]$ to deploy. The feature $x_s \in \{\text{woman}, \text{man}\}$ is sensitive.

Company

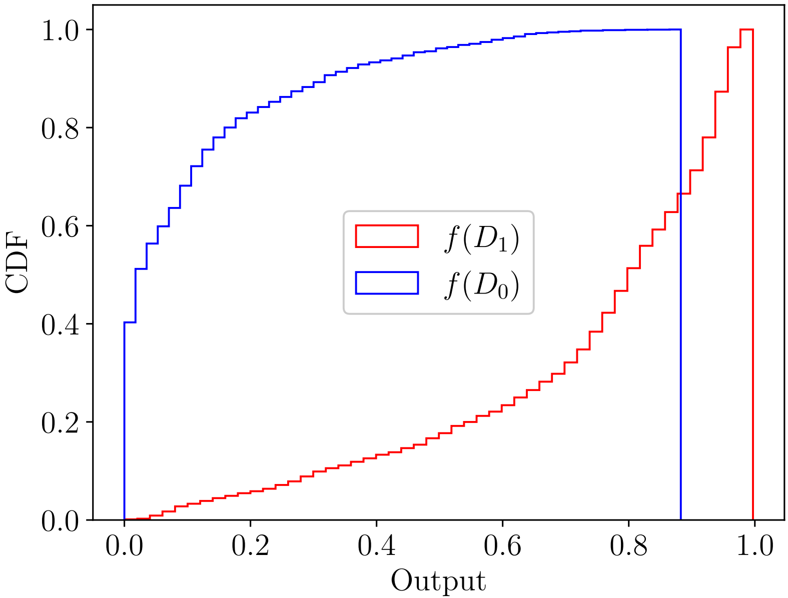
Auditor

To verify the model, we need to measure its **fairness** metrics. Can you provide access to collection of outputs $f(D_{\text{woman}}), f(D_{\text{man}})$?

$$D_{\text{woman}} = \{\boldsymbol{x}^{(i)} : x_s^{(i)} = \text{woman}\},$$
$$D_{\text{man}} = \{\boldsymbol{x}^{(i)} : x_s^{(i)} = \text{man}\}$$

Company

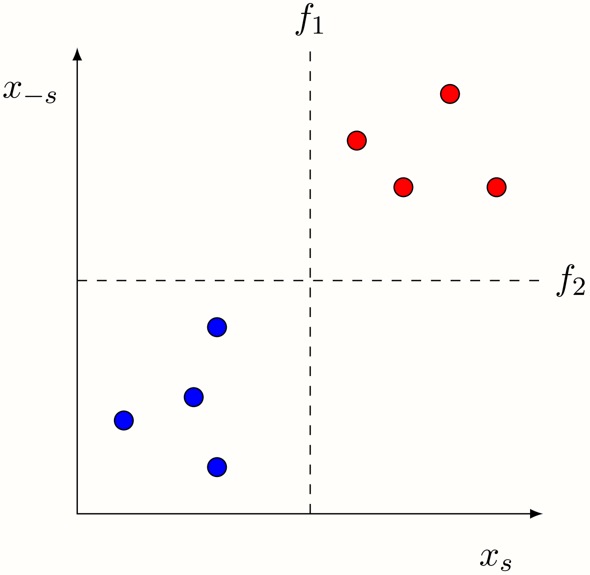
Auditor


1.0
0.8
0.6
0.4
0.2
0.0
CDF
0.0
0.2
0.4
0.6
0.8
1.0
Output
$f(D_1)$
$f(D_0)$

There may be a disparity in model outcomes but that does not means that the model is relying on the sensitive feature.
The model may rely on **meritocratic** features correlated with x_s.

Company

Auditor


f_1
x_{-s}
f_2
x_s

To validate your argument we could compute the **Shapley Values** $\boldsymbol{\Phi}$ and see which features contribute the most to the disparity.

Company

Auditor

Shapley Values to Explain Fairness

$$\sum_{i=1}^{d} \Phi_i(f, D_{\text{woman}}, D_{\text{man}}) = \mathbb{E}[f(\boldsymbol{x})|x_s = \text{woman}] - \mathbb{E}[f(\boldsymbol{x})|x_s = \text{man}]. \quad (1)$$

Shapley Values to Explain Fairness

$$\sum_{i=1}^{d} \Phi_i(f, D_{\text{woman}}, D_{\text{man}}) = \mathbb{E}[f(\boldsymbol{x})|x_s = \text{woman}] - \mathbb{E}[f(\boldsymbol{x})|x_s = \text{man}]. \quad (1)$$

Constraint

In practice, a Monte-Carlo estimate $\widehat{\boldsymbol{\Phi}}(f, S_{\text{woman}}, S_{\text{man}})$ is used with two subsets $S_{\text{woman}} \subset D_{\text{woman}}$ and $S_{\text{man}} \subset D_{\text{man}}$ sampled uniformly at random.

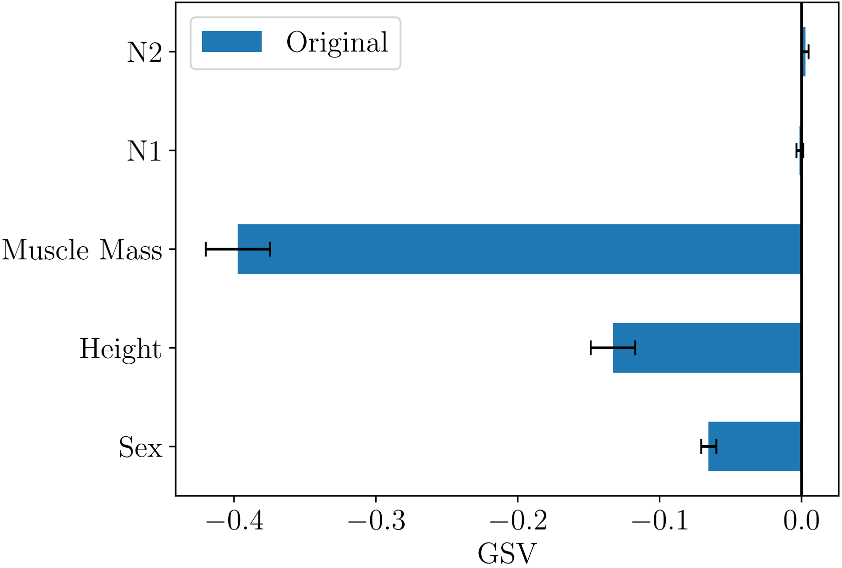
I will share with you two subsets $S_{\text{woman}} \subset D_{\text{woman}}$ and $S_{\text{man}} \subset D_{\text{man}}$ of size M so you can run SHAP on our model and get $\widehat{\mathbf{\Phi}}(f, S_{\text{woman}}, S_{\text{man}})$.

Company

Auditor

Ok let's run SHAP on our own and see what we get.

Company


Original
N2
N1
Muscle Mass
Height
Sex
−0.4
−0.3
−0.2
−0.1
0.0
GSV

Ouch ! Is there a way to cherry-pick the subsets $S'_{\text{woman}}, S'_{\text{man}}$ so that $|\widehat{\Phi}_s(f, S'_{\text{woman}}, S'_{\text{man}})|$ is small and the auditor cannot detect the manipulation ?

Company

Detection

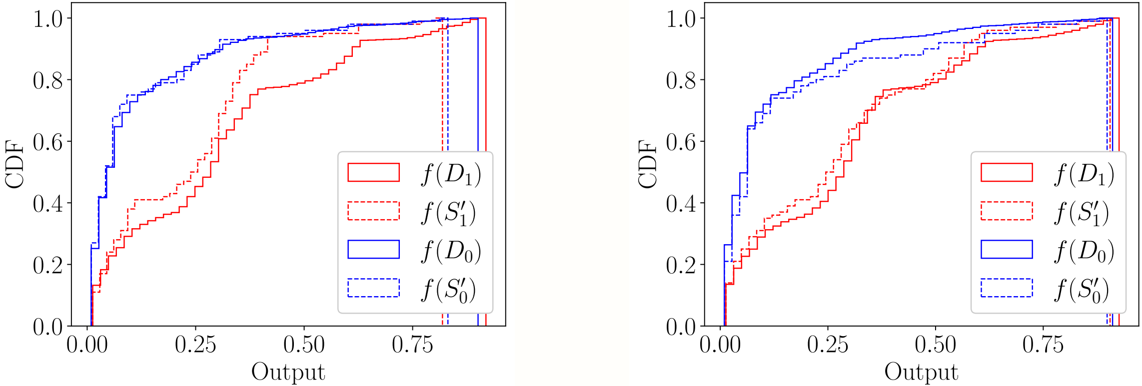
The audit already has access to $f(D_{\text{woman}}), f(D_{\text{man}})$. Hence they can detect the manipulation with a statistical test

$$\texttt{Detect\_fraud}(f(D_{\text{woman}}), f(D_{\text{man}}), f(S'_{\text{woman}}), f(S'_{\text{man}}))$$

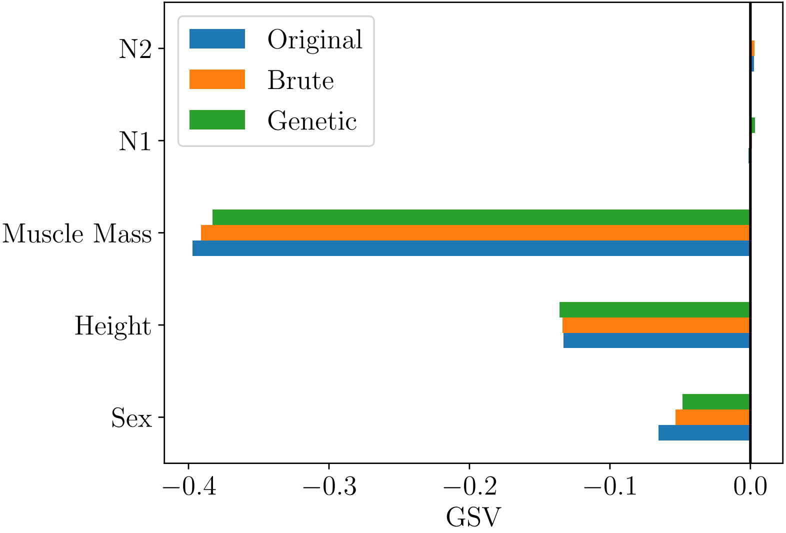
Detection

The audit already has access to $f(D_{\text{woman}}), f(D_{\text{man}})$. Hence they can detect the manipulation with a statistical test

$$\texttt{Detect\_fraud}(f(D_{\text{woman}}), f(D_{\text{man}}), f(S'_{\text{woman}}), f(S'_{\text{man}}))$$



Baselines



Original
Brute
Genetic
N2
N1
Muscle Mass
Height
Sex
−0.4
−0.3
−0.2
−0.1
0.0
GSV

Method 3 : Fool SHAP

Issues with Genetic Algorithm

1. Feature correlations are ignored in cross-over operation.
2. There is no notion of proximity to the original data.

Method 3 : Fool SHAP

Issues with Genetic Algorithm

1. Feature correlations are ignored in cross-over operation.
2. There is no notion of proximity to the original data.

Solution : Fool SHAP

1. Sample S'_{woman} uniformly at random.

Method 3 : Fool SHAP

Issues with Genetic Algorithm

1. Feature correlations are ignored in cross-over operation.
2. There is no notion of proximity to the original data.

Solution : Fool SHAP

1. Sample S'_{woman} uniformly at random.
2. Define $\mathcal{B} = \frac{1}{N_{\text{man}}} \sum_{\boldsymbol{x}^{(i)} \in D_{\text{man}}} \delta(\boldsymbol{x}^{(i)})$ and $\mathcal{B}' = \sum_{\boldsymbol{x}^{(i)} \in D_{\text{man}}} \omega_i \delta(\boldsymbol{x}^{(i)})$

Method 3 : Fool SHAP

Issues with Genetic Algorithm

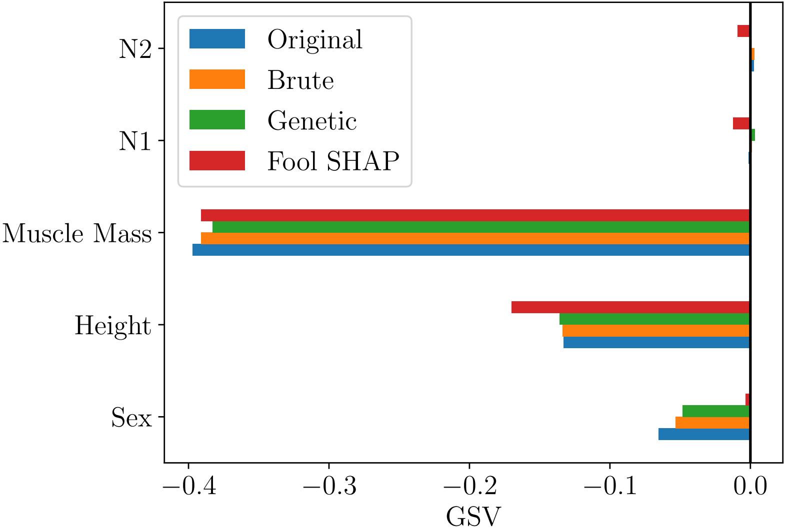
1. Feature correlations are ignored in cross-over operation.
2. There is no notion of proximity to the original data.

Solution : Fool SHAP

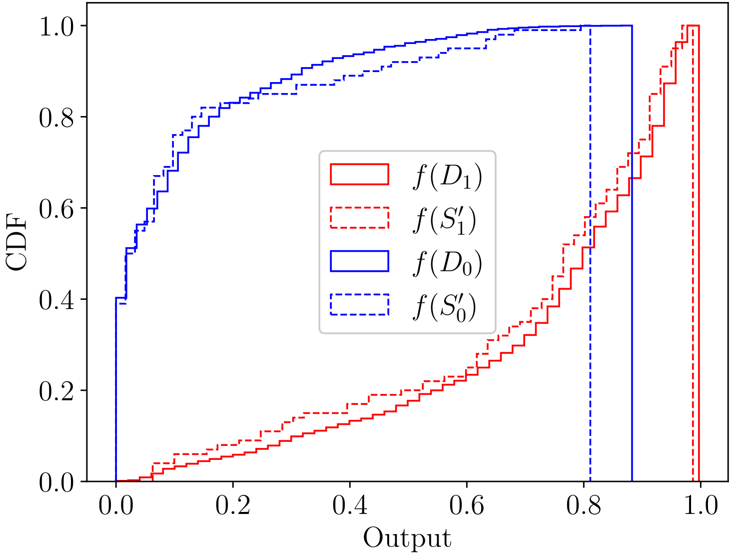
1. Sample S'_{woman} uniformly at random.
2. Define $\mathcal{B} = \frac{1}{N_{\text{man}}} \sum_{\boldsymbol{x}^{(i)} \in D_{\text{man}}} \delta(\boldsymbol{x}^{(i)})$ and $\mathcal{B}' = \sum_{\boldsymbol{x}^{(i)} \in D_{\text{man}}} \omega_i \delta(\boldsymbol{x}^{(i)})$
3. Optimize the weights $\boldsymbol{\omega}$ such that :
   - $|\widehat{\Phi}_s(f, S'_{\text{woman}}, S'_{\text{man}})|$ with $S'_{\text{man}} \sim \mathcal{B}'^M$ is small.
   - $\mathcal{B}'$ is close to $\mathcal{B}$ w.r.t the Wasserstein Distance.

Solved with a Minimum Cost Flow (MCF) Linear Program.

Method 3 Fool SHAP



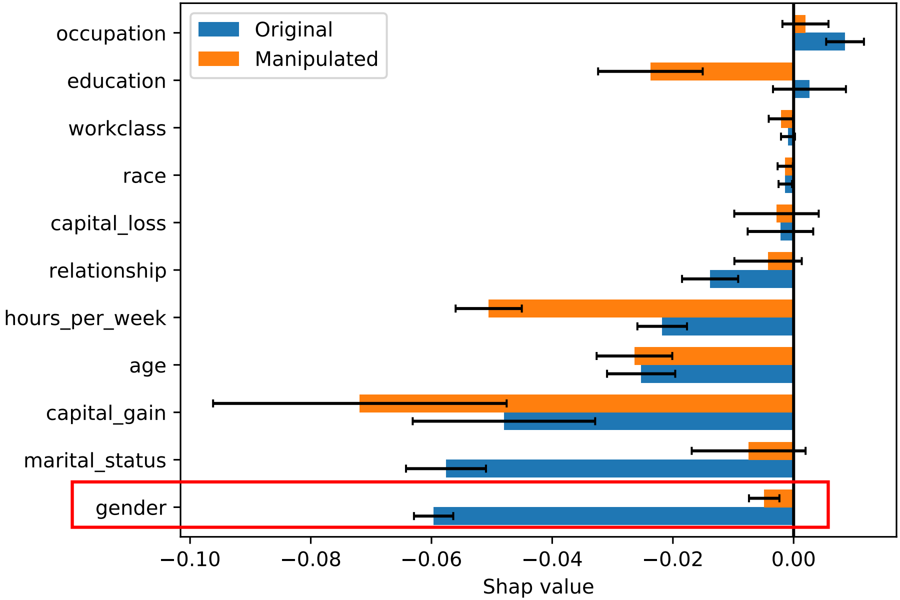
Method 3 Fool SHAP

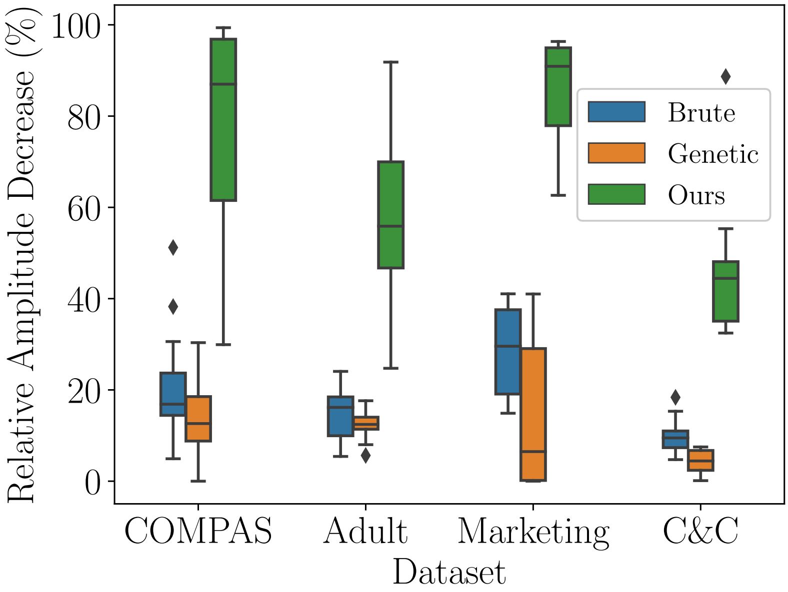


Here are the subsets $S'_{\text{woman}}, S'_{\text{man}}$ requested.

Company

Auditor


Original
Manipulated
occupation
education
workclass
race
capital_loss
relationship
hours_per_week
age
capital_gain
marital_status
gender
−0.10
−0.08
−0.06
−0.04
−0.02
0.00
Shap value


Relative Amplitude Decrease (%)
100
80
60
40
20
0
COMPAS
Adult
Marketing
C&C
Dataset
Brute
Genetic
Ours

Conclusion

Contributions

- A new and effective attack on SHAP.
- Said attacks are hard to detect by an external auditor.
- An auditor would need some access to the input features of the private data to circumvent the attack.

Conclusion

Contributions

- A new and effective attack on `SHAP`.
- Said attacks are hard to detect by an external auditor.
- An auditor would need some access to the input features of the private data to circumvent the attack.

Future Work

- Allow the audit to query more information about the private dataset.
- Cherry-pick S'_{woman} et S'_{man} simultaneously (Bilinear Problem).
- Apply to other measures of fairness.