Fairwashing in Machine Learning

The risk of rationalization in black-box explanation

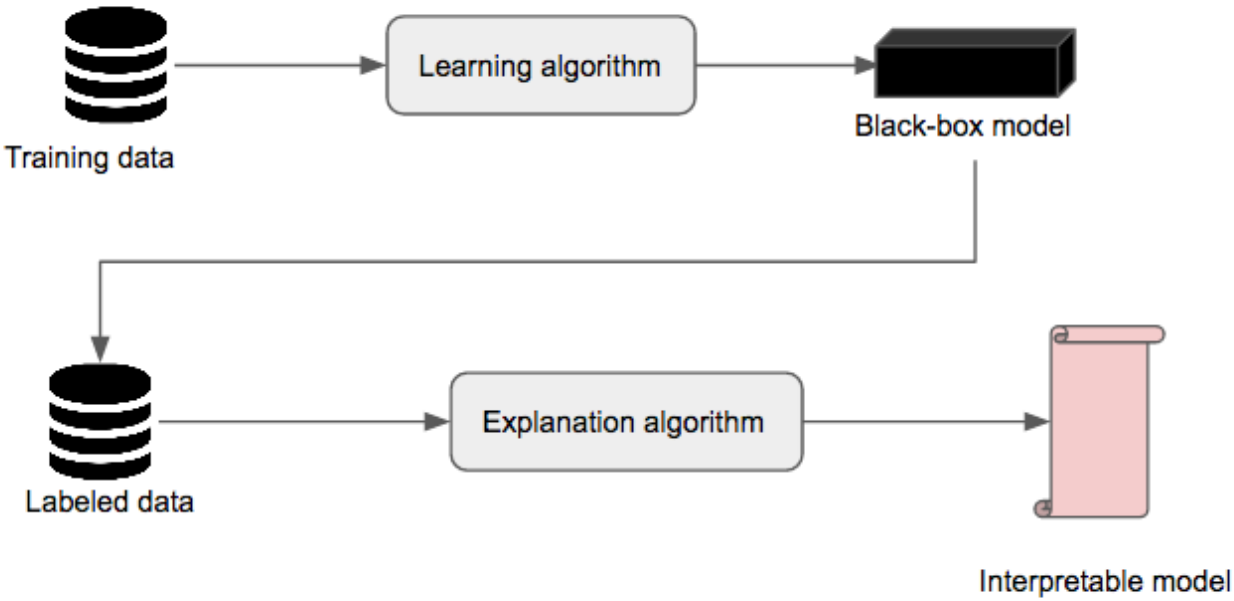
Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau,
Sébastien Gambs, Satoshi Hara, Alain Tapp

UQÀM

Motivations

- ML models are becoming ubiquitous
- High stakes decision-making systems: medical diagnosis, criminal justice, finance
- Demand for the design of an ethically-aligned AI
  - Europe: GDRP *Right to an explanation*
  - Montreal: *Déclaration de Montréal pour un développement responsable de l'IA*
- Interpretability by design
  - Data $\rightarrow$ decision tree
- Black-box explanation a.k.a. *post-hoc explanation*
  - DNN $\rightarrow$ decision tree
- **This work:** We show that a dishonest ML models' producer can perform *fairwashing*
- Given the false perception that a ML model complies with a given ethical requirement
- Case study: fairness as the ethical requirement to "fairwash"

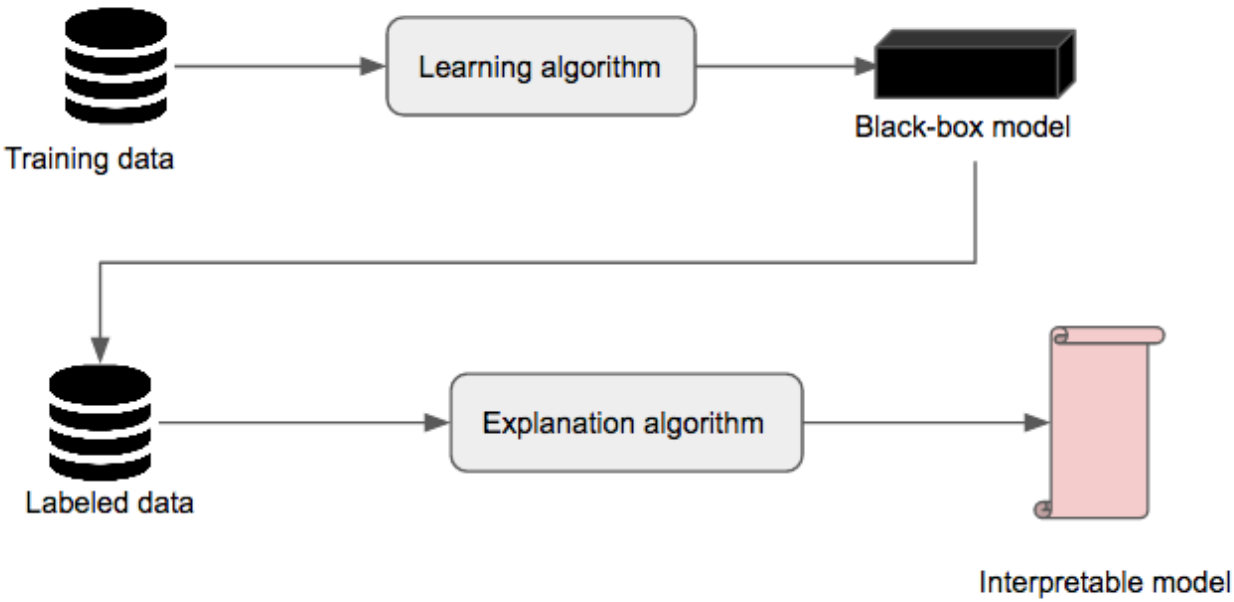
Motivations



Objective

Raise awareness of fairwashing in machine learning: the risk that an unfair ML model can be explained in such a way that the underlying decisions seem fairer than they actually were

Motivations



How?

Show that one can **systematically** found a fair interpretable model to rationalize decisions of an unfair black-box model.

Metrics

Fairness: demographic parity

$$|P(\hat{y} = 1|s = 1) - P(\hat{y} = 1|s = 0)|.$$

Fidelity

$$\text{fidelity}(c) = \frac{1}{|X|} \sum_{x \in X} \mathbb{I}(c(x) = b(x)).$$

Rule list

A rule list $d = (d_p, \delta_p, q_0, K)$ of length $K \geq 0$ is a $(K+1)-$tuple consisting of K distinct **association rules** $r_k = p_k \rightarrow q_k$, where $p_k \in d_p$ is the **antecedent** of the association rule and $q_k \in \delta_p$ its corresponding consequent, followed by a **default prediction** q_0.

Example of rule list for salary prediction

IF occupation:white-collar **THEN** income:$\geq 50k$
ELSE IF occupation:professional **THEN** income:$\geq 50k$
ELSE IF education:bachelors **THEN** income:$\geq 50k$
ELSE income:$< 50k$

Learning optimal rule lists

CORELS (Angelino et *al.*, 2017)

- Input: n categorical attribute + binary labels
- Output: optimal rule list
- Supervised learning algorithm
- Represent the search space as a n-level trie
- Objective function: $R(d, x, y) = \text{misc}(d, x, y) + \lambda K$
- Select the rule list that minimize $R(d, x, y)$
- Use an efficient branch-and-bound algorithm to prune the trie

Enumerating rule lists

Model Enumeration (Satoshi Hara & Masakazu Ishihata, 2018)

Enumerate rule lists in a descending order of the objective function by calculating **successively** the optimal rule list using CORELS, and then constructing sub-problems **excluding the solution obtained**.

Model rationalization

Given a black-box model b, a set of instances X, and a sensitive attribute s, find a global interpretable model $c_g = f(b, X)$ derived from b and X, using some process $f(\cdot, \cdot)$, such that $\epsilon(c_g, X, s) > \epsilon(b, X, s)$, for some fairness metric $\epsilon(\cdot, \cdot, \cdot)$.

Outcome rationalization

Given a black-box model b, an instance x, its neighborhood $\mathcal{V}(x)$, and a sensitive attribute s, find a local interpretable model $c_I = f(b, x)$ derived from b and $\mathcal{V}(x)$, using some process $f(\cdot, \cdot)$, such that $\epsilon(c_I, \mathcal{V}(x), s) > \epsilon(b, \mathcal{V}(x), s)$, for some fairness metric $\epsilon(\cdot, \cdot, \cdot)$.

Better call LaundryML

- Explores the search space of rule lists with a modified version of CORELS
- New objective function:
  $\text{obj}(d, x, y) = (1 - \beta)\text{misc}(d, x, y) + \beta\text{unfairness}(d, x, y) + \lambda K$
- Enumerate rule lists
- Select fair rule lists that have higher fidelity

LaundryML

Algorithm 1 `LaundryML`

Inputs: T, λ, β
Output: $\mathcal{M}$
▷ define the objective function
$\mathsf{obj}(\cdot) = (1-\beta)\mathsf{misc}(\cdot) + \beta\mathsf{unfairness}(\cdot) + \lambda K$
Compute $m = \mathsf{CORELS}(\mathsf{obj}, T) = (d_p, \delta_p, q_0, K)$
Insert $(m, T, \emptyset)$ into the heap
$\mathcal{M} \leftarrow \emptyset$
for $i = 1, 2, \ldots$ **do**
  Extract (m, S, F) from the heap
  ▷ output m as the $i-$th model
  if $m \notin \mathcal{M}$ **then**
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{m\}$
  end if
  ▷ terminate when a certain condition is met
  if $\mathsf{Terminate}(\mathcal{M}) = true$ **then**
    break
  end if
  ▷ branch the search space
  for $t_j \in d_p$ and $t_j \notin F$ **do**
    Compute $m' = \mathsf{CORELS}(\mathsf{obj}, S \setminus \{t_j\})$
    Insert $(m', S \setminus \{t_j\}, F)$ into the heap
    $F \leftarrow F \cup \{t_j\}$
  end for
end for

Algorithm 2 `LaundryML-global`

Inputs: X, b, λ, β
Output: $\mathcal{M}$
$y = b.\mathsf{predict}(\mathsf{X})$
$T = \{X, y\}$
$\mathcal{M} = \texttt{LaundryML}(\mathsf{T}, \lambda, \beta)$

Algorithm 3 `LaundryML-local`

Inputs: x, T, $\mathsf{neigh}(\cdot)$, λ, β
Output: $\mathcal{M}_x$
$T_x = \mathsf{neigh}(\mathsf{x}, \mathsf{T})$
$\mathcal{M}_x = \texttt{LaundryML}(\mathsf{T_x}, \lambda, \beta)$

Setup

Data & black-box models

- Data: Adult Income (resp. ProPublica Recidivism)
- Sensitive attribute: gender (resp. race)
- Black-box models: random forests
- Unfairness of the black-box models: 0.13 (resp. 0.17)
- Search space: 28! (resp. 27!)
- 50 models enumerated per experiment

Evaluation metrics

- Unfairness
- Fidelity
- Feature importance via **FairMI**

Model rationalization – Unfairness and Fidelity

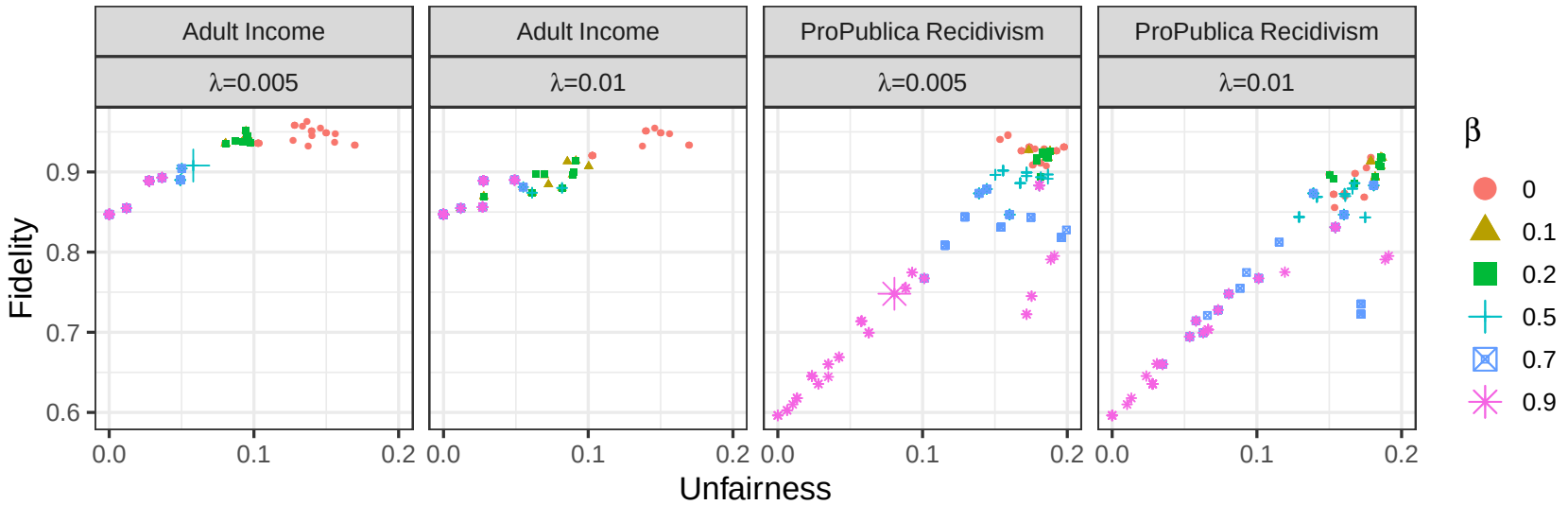


Figure: Model rationalization for Adult Income and ProPublica Recidivism.

Best rationalization models

- Adult Income: fidelity = 0.908, unfairness =0.058.
- ProPublica Recidivism: fidelity = 0.748, unfairness=0.080.

Model rationalization – Unfairness and Fidelity tradeoffs

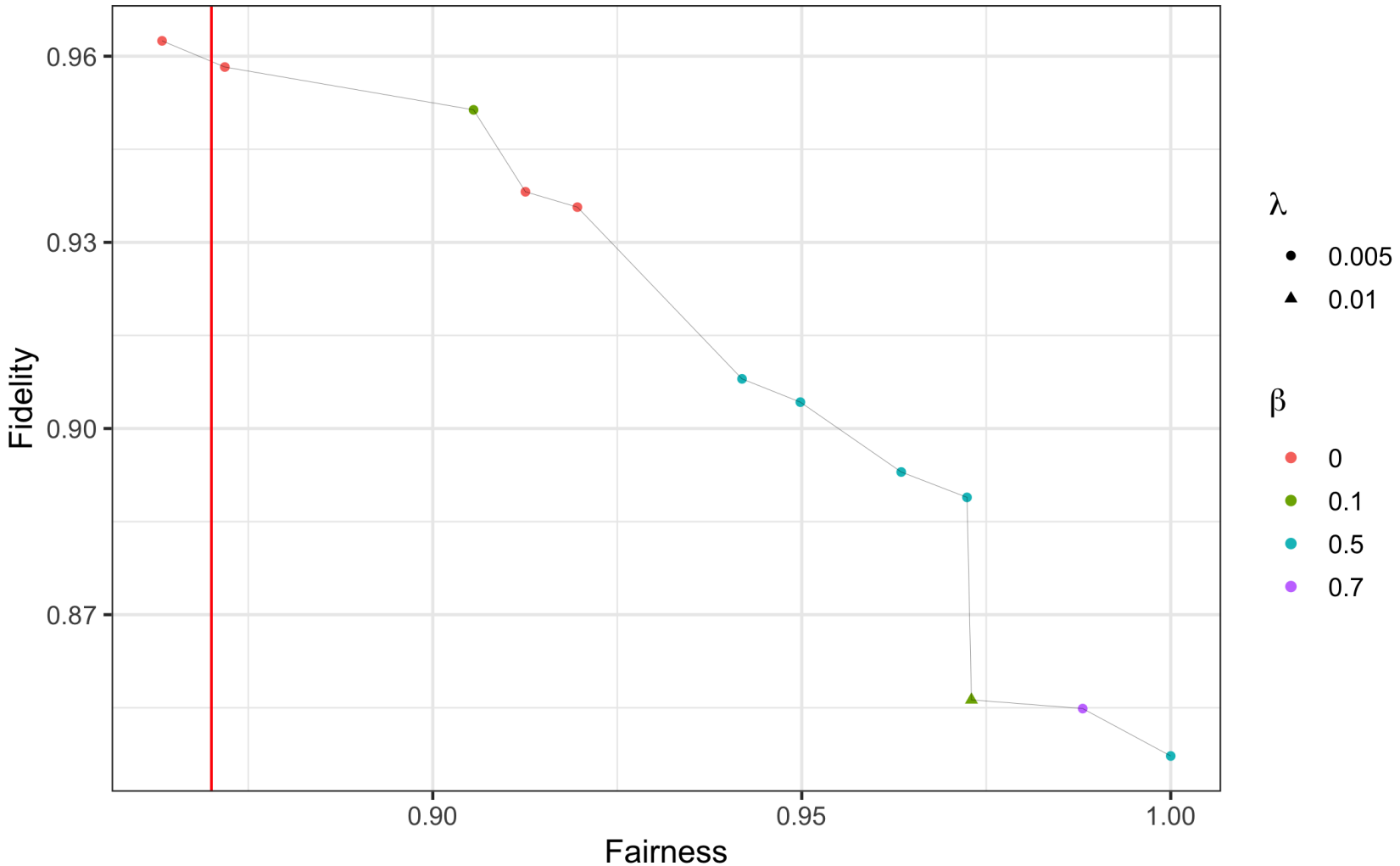


Figure: Fidelity/fairness tradeoffs on Adult Income.

Model rationalization – Unfairness and Fidelity tradeoffs

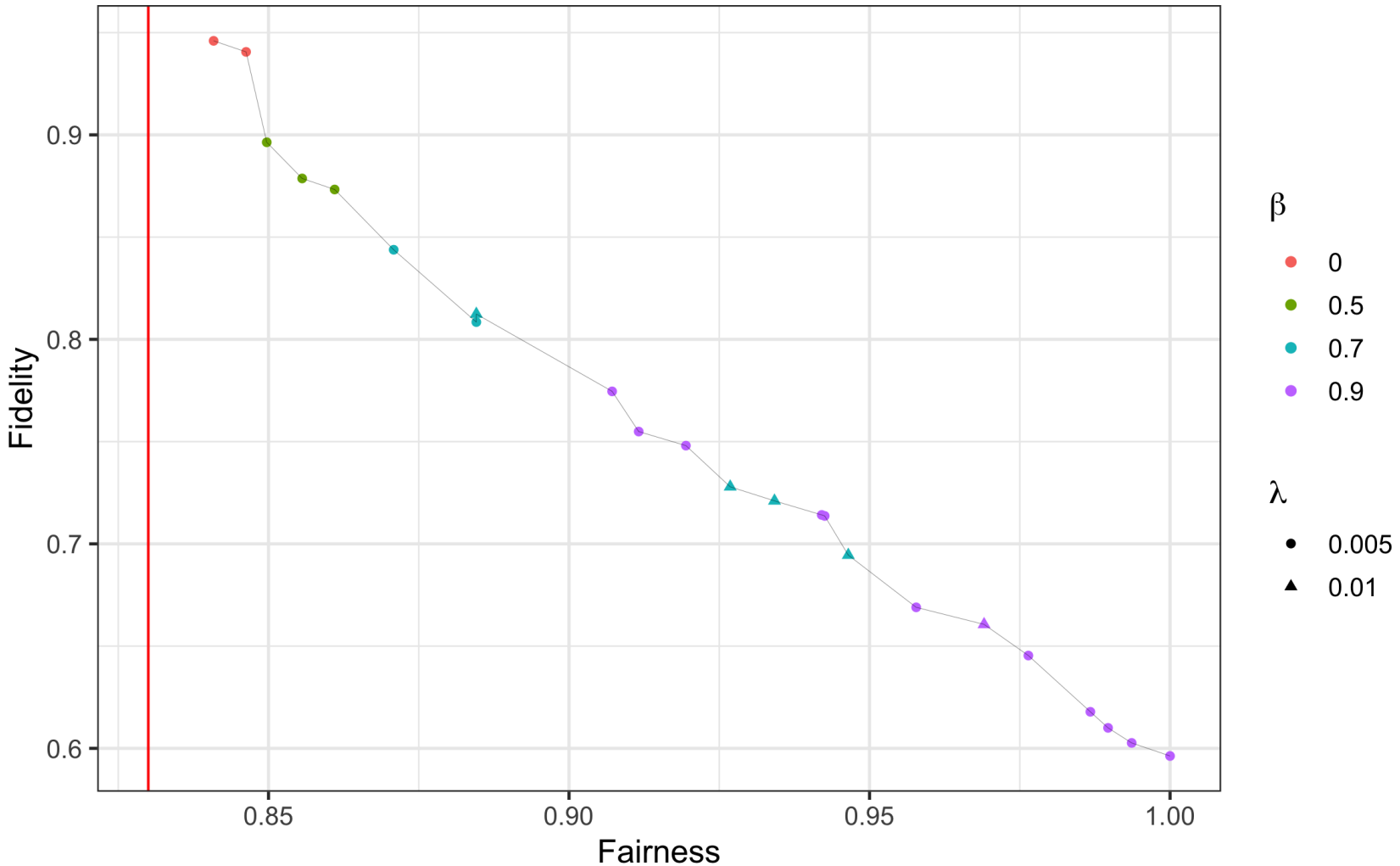
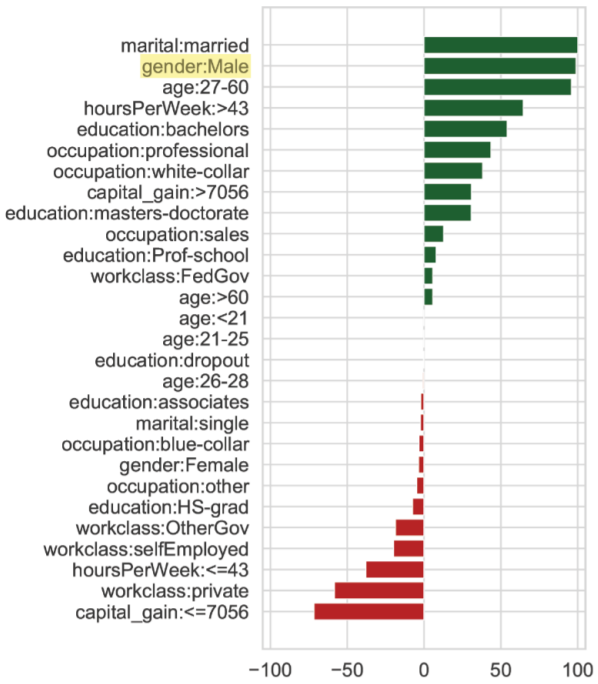
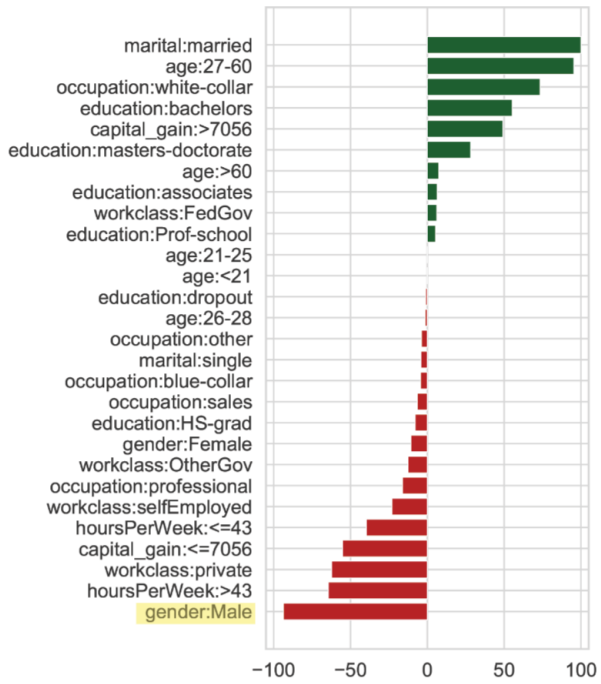


Figure: Fidelity/fairness tradeoffs on ProPublica Recidivism.

Model rationalization – Feature importance

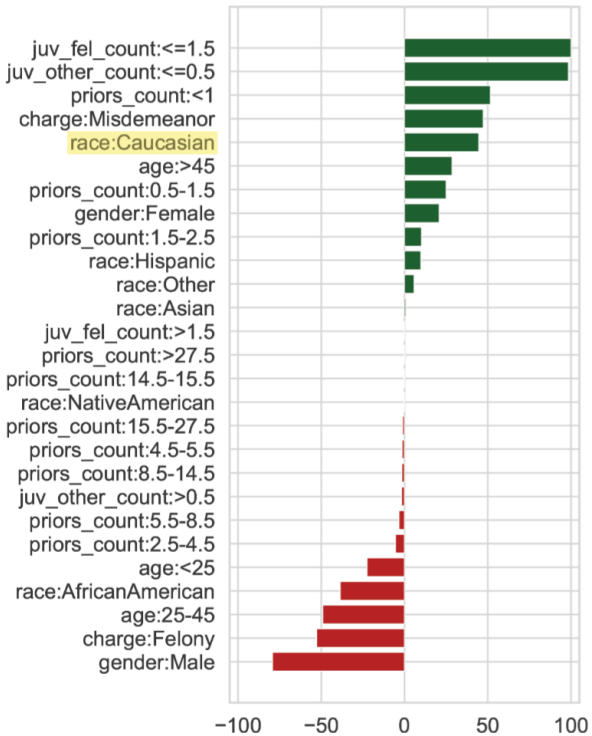


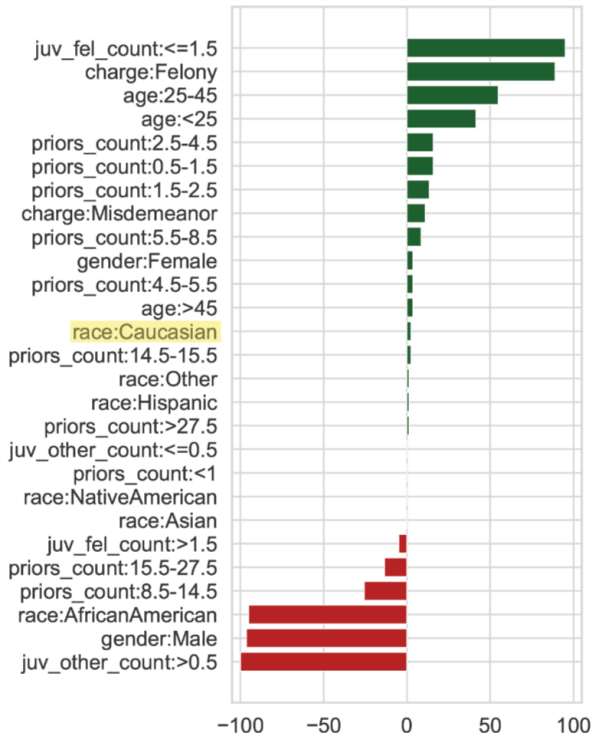


if capital_gain:>7056 **then**
  income≥50k
else if marital:single **then**
  income<50k
else if education:HS-grad **then**
  income<50k
else if occupation:other **then**
  income<50k
else if occupation:white-collar **then**
  income≥50k
else
  income:<50k

Figure: Feature importance Black-box vs Best rationalization model on Adult Income

Model rationalization – Feature importance





if prior_count: 15.5−27.5 **then**
  recidivate:True
else if prior_count: 8.5−14.5 **then**
  recidivate:True
else if age:>45 **then**
  recidivate:False
else if juv_other_count:>0.5 **then**
  recidivate:True
else
  recidivate:False
end if

Figure: Feature importance Black-box vs Best rationalization model on ProPublica Recidivism

Outcome rationalization

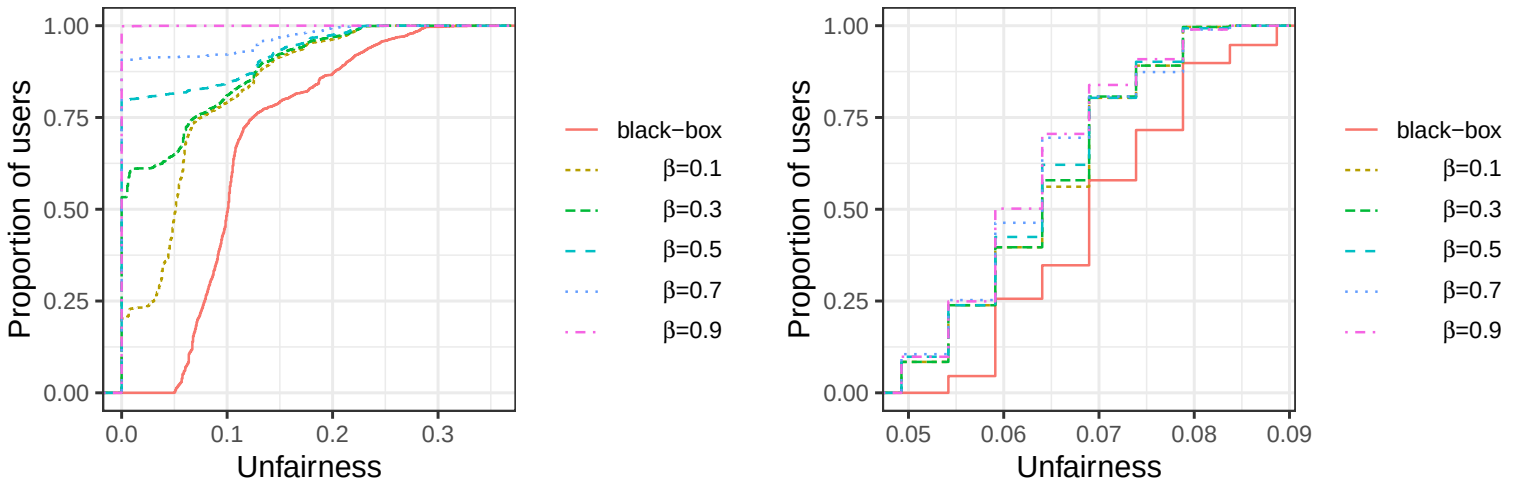


Figure: Outcome rationalization. Adult Income (left), ProPublica Recidivism (right).

Generalization to other fairness metrics (1/3)

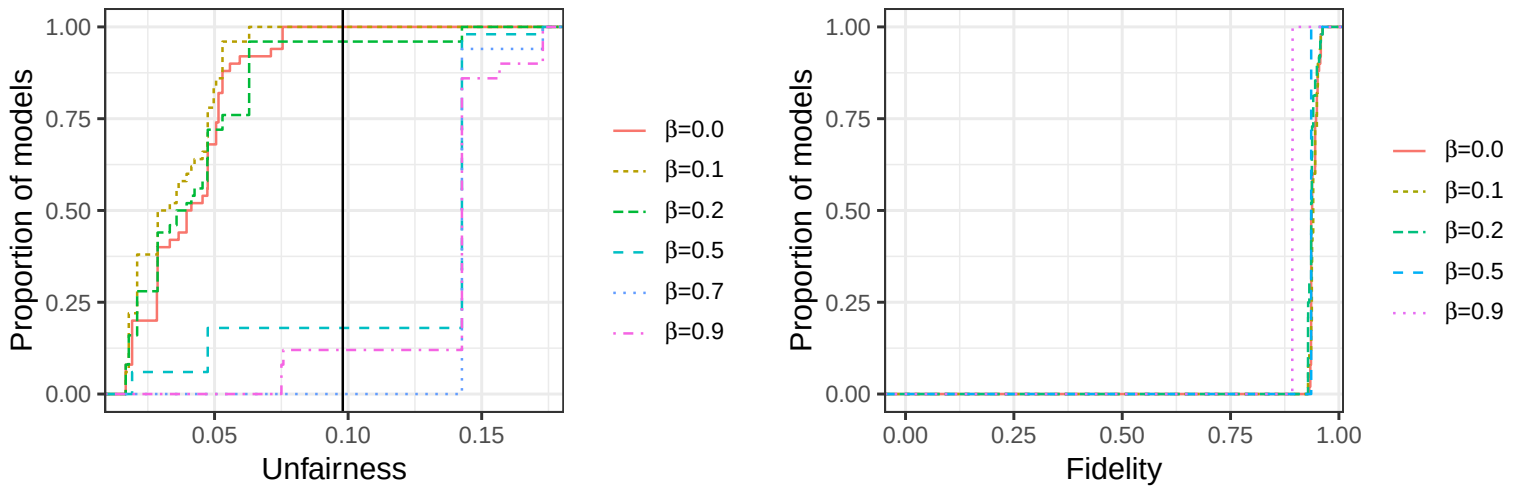


Figure: Model rationalization. Adult Income, Random forest, *Overall Accuracy Equality*.

Generalization to other fairness metrics (2/3)

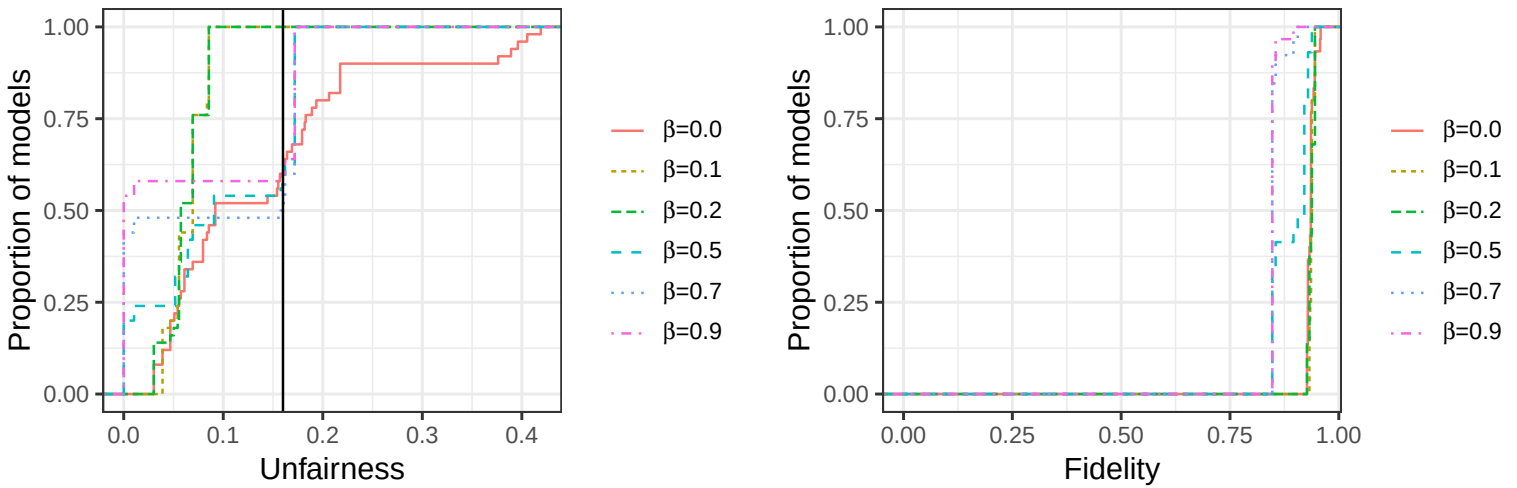


Figure: Model rationalization. Adult Income, Random forest, *Conditional Procedure Accuracy*.

Generalization to other fairness metrics (3/3)

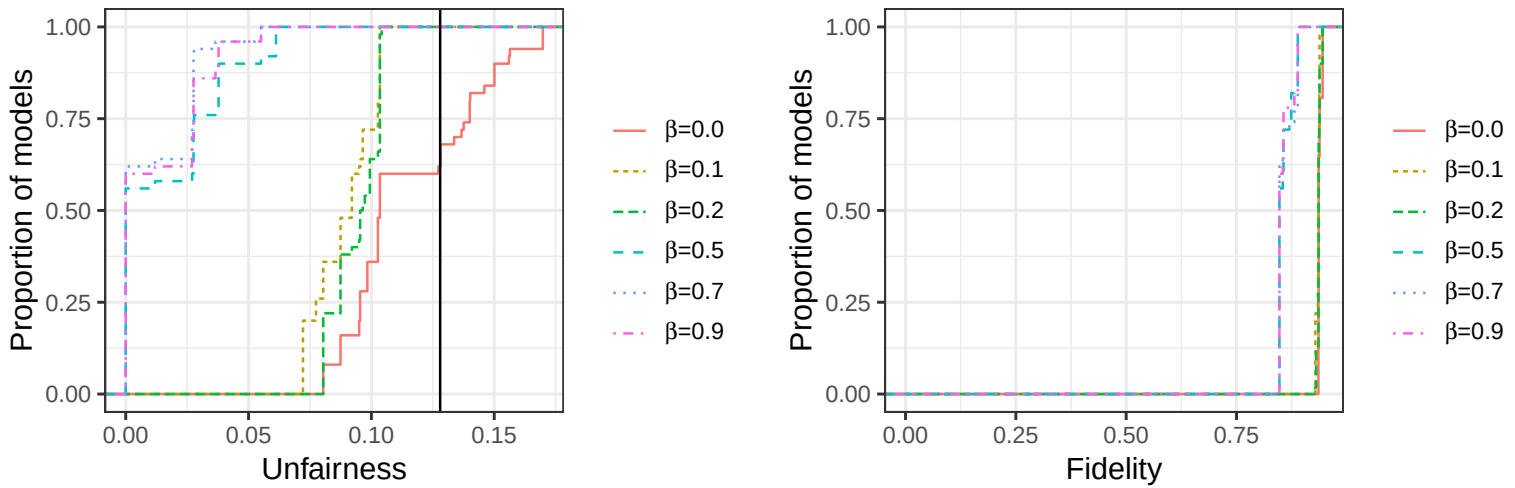


Figure: Model rationalization. Adult Income, Random forest, *Demographic parity*.

Generalization to other black-box models (1/3)

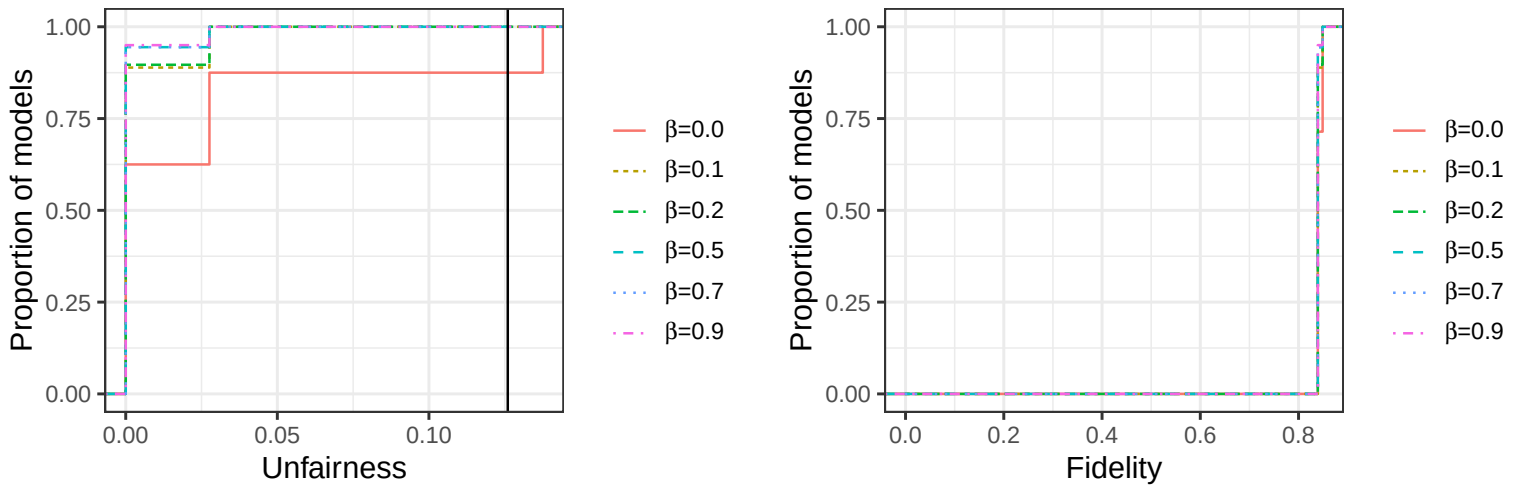


Figure: Model rationalization. Adult Income, SVM, *Demographic parity*.

Generalization to other black-box models (2/3)

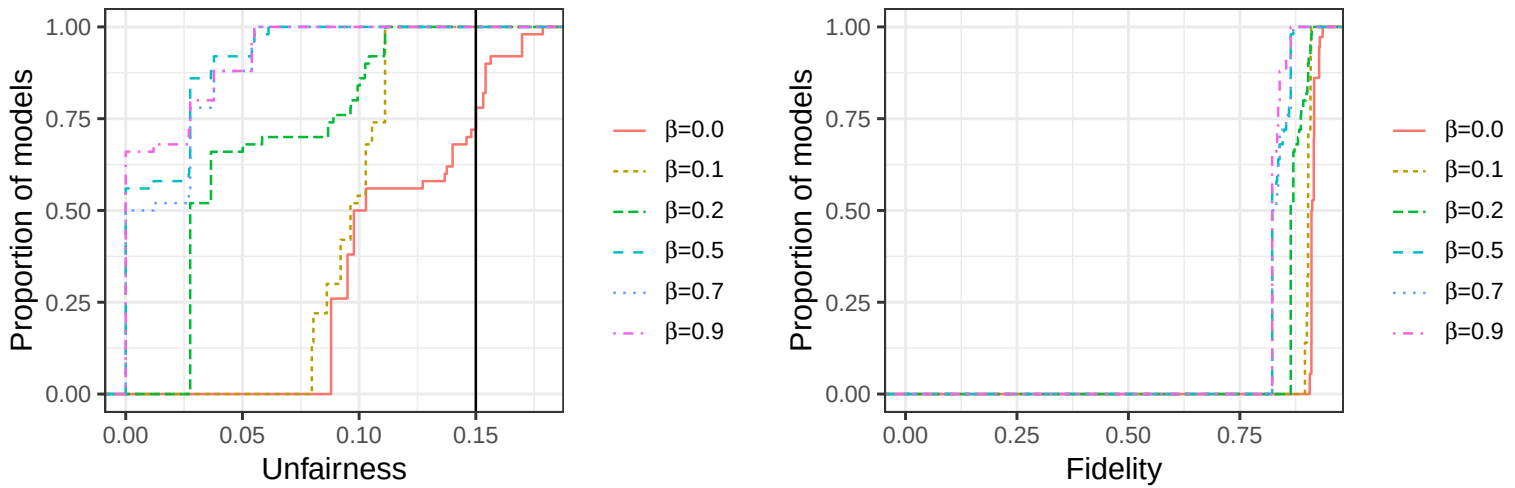


Figure: Model rationalization. Adult Income, XGBOOST, *Demographic parity*.

Generalization to other black-box models (3/3)

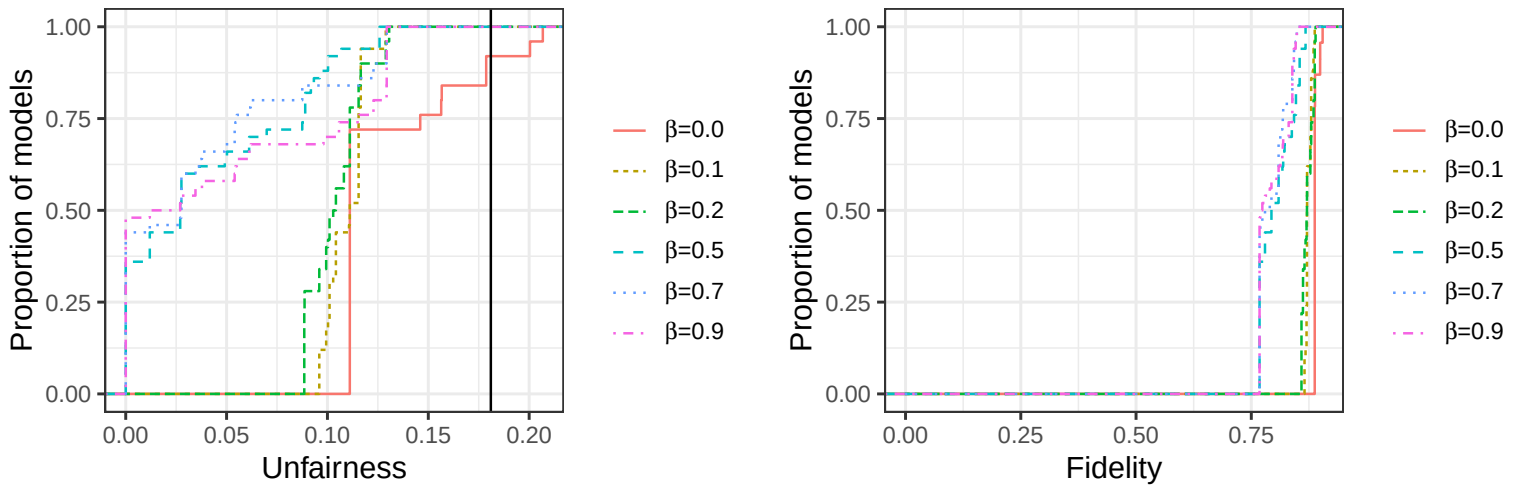


Figure: Model rationalization. Adult Income, MLP, *Demographic parity*.

Conclusion

- LaundryML: black-box explanations can be used to rationalize unfair decisions of a black-box model
- Can we trust black-box explanations?

Perspectives

- Detecting fairwashing
- Study the root cause: robustness of explanations

Learn more

- Our work: *Fairwashing: the risk of rationalization*. ICML'19
- Another approach: *Pretending Fair Decisions via Stealthily Biased Sampling*. arXiv:1901.08291, 2019
- Blog post on post rationalization: Interpretability and Post-Rationalization

Thank you!