Fairwashing: the risk of rationalization

Ulrich Aïvodji [1]   Hiromi Arai [2 3]   Olivier Fortineau [4]   Sébastien Gambs [1]   Satoshi Hara [5]   Alain Tapp [6 7]

Abstract

Black-box explanation is the problem of explaining how a machine learning model – whose internal logic is hidden to the auditor and generally complex – produces its outcomes. Current approaches for solving this problem include model explanation, outcome explanation as well as model inspection. While these techniques can be beneficial by providing interpretability, they can be used in a negative manner to perform fairwashing, which we define as promoting the false perception that a machine learning model respects some ethical values. In particular, we demonstrate that it is possible to systematically rationalize decisions taken by an unfair black-box model using the model explanation as well as the outcome explanation approaches with a given fairness metric. Our solution, LaundryML, is based on a regularized rule list enumeration algorithm whose objective is to search for fair rule lists approximating an unfair black-box model. We empirically evaluate our rationalization technique on black-box models trained on real-world datasets and show that one can obtain rule lists with high fidelity to the black-box model while being considerably less unfair at the same time.

1. Introduction

In recent years, the widespread use of machine learning models in high stakes decision-making systems (*e.g.*, credit scoring (Siddiqi, 2012), predictive justice (Kleinberg et al., 2017) or medical diagnosis (Caruana et al., 2015)) combined with proven risks of incorrect decisions – such as people being wrongly denied parole (Wexler, 2017a) – has considerably raised the public demand for being able to provide explanations to algorithmic decisions.

In particular, to ensure transparency and explainability in algorithmic decision processes, several initiatives have emerged for regulating the use of machine learning models. For instance, in Europe, the new General Data Protection Regulation has a provision requiring explanations for the decisions of machine learning models that have a significant impact on individuals (Goodman & Flaxman, 2017).

Existing methods to achieve explainability include transparent-box design and black-box explanation (also called post-hoc explanation) (Lipton, 2018; Lepri et al., 2017; Montavon et al., 2018; Guidotti et al., 2018). The former consists in building transparent models, which are inherently interpretable by design. This approach requires the cooperation of the entity responsible for the training and usage of the model. In contrast, the latter involves an adversarial setting in which the black-box model, whose internal logic is hidden to the auditor, is reversed-engineered to create an interpretable surrogate model.

While many recent works have focused on black-box explanations, their use can be problematic in high stakes decision-making systems (Reddix-Smalls, 2011; Wexler, 2017b) as explanations can be unreliable and misleading (Rudin, 2018; Melis & Jaakkola, 2018). Current techniques for providing black-box explanations include model explanation, outcome explanation and model inspection (Guidotti et al., 2018). Model explanation consists in building an interpretable model to explain the whole logic of the black box while outcome explanation only cares about providing a local explanation of a specific decision. Finally, model inspection consists of all the techniques that can help to understand (*e.g.*, through visualizations or quantitative arguments) the influence of the attributes of the input on the black-box decision.

Since the right to explanation as defined in current regulations (Goodman & Flaxman, 2017) does not give precise directives on what it means to provide a "valid explanation" (Wachter et al., 2017; Edwards & Veale, 2017), there is a legal loophole that can be used by dishonest companies to cover up the possible unfairness of their black-box models by providing misleading explanations. In particular, due to the growing importance of the concepts of fairness in machine learning (Barocas & Selbst, 2016), a company might be tempted to perform *fairwashing*, which we define

[1]Université du Québec à Montréal [2]RIKEN Center for Advanced Intelligence Project [3]JST PRESTO [4]ENSTA ParisTech [5]Osaka University [6]UdeM [7]MILA. Correspondence to: Ulrich Aïvodji <aivodji.ulrich@courrier.uqam.ca>.

as promoting the false perception that the learning models used by the company are fair while it might not be so.

In this paper, our main objective is to raise the awareness of this issue, which we believe to be a serious concern related to the use of black-box explanation. In addition, to provide concrete evidence on this possible risk, we show that it is possible to forge a fairer explanation from a truly unfair black box through a process that we coin as *rationalization*.

In particular, we propose a systematic rationalization technique that, given black-box access to a model f, produces an ensemble C of interpretable models $c \approx f$ that are fairer than the black-box according to a predefined fairness metric. From this set of plausible explanations, a dishonest entity can pick a model to achieve fairwashing. One of the strength of our approach is that it is agnostic to both the black-box model and the fairness metric. To demonstrate the genericity of our technique, we show that the rationalization can be used to explain globally the decision of a black-box model (*i.e.*, rationalization of model explanation) as well as its decision for a particular input (*i.e.*, rationalization of outcome explanation). While our approach is mainly presented in this paper as a proof-of-concept to raise awareness on this issue, we believe that the number of ways rationalization can be instantiated are endless.

The outline of the paper is as follows. First in Section 2, we review the related work on interpretability and fairness necessary to the understanding of our work. Afterwards in Section 3, we formalize the rationalization problem before introducing LaundryML, our algorithm for the enumeration of rule lists that can be used to perform fairwashing. Finally, in Section 4 we report on the evaluation obtained when applying on rationalization approach on two real datasets before concluding in Section 5.

2. Related work

2.1. Interpretability and explanation

In the context of machine learning, interpretability can be defined as the ability to explain or to provide the meaning in understandable terms to a human (Doshi-Velez & Kim, 2017). An explanation is an interface between humans and a decision process that is both an accurate proxy of the decision process and understandable by humans (Guidotti et al., 2018). Examples of such interfaces include linear models (Ribeiro et al., 2016), decision trees (Breiman, 2017), rule lists (Rivest, 1987; Letham et al., 2015; Angelino et al., 2018) and rule sets (Li et al., 2002). In this paper, we focus on two black-box explanation tasks, namely model explanation and outcome explanation (Guidotti et al., 2018).

Definition 1 (Model explanation). Given a black-box model b and a set of instances X, the model explanation problem consists in finding an explanation $E \in \mathcal{E}$ belonging to a human-interpretable domain $\mathcal{E}$, through an interpretable global predictor $c_g = f(b, X)$ derived from the black-box b and the instances X using some process $f(\cdot, \cdot)$.

For example, if we choose the domain $\mathcal{E}$ to be the set of decision trees, model explanation amounts to identifying a decision tree that approximates well the black-box model. The identified decision tree can be interpreted as an explanation of the black-box model (Craven & Shavlik, 1996).

Definition 2 (Outcome explanation). Given a black-box model b and an instance x, the outcome explanation problem consists in finding an explanation $e \in \mathcal{E}$, belonging to a human-interpretable domain $\mathcal{E}$, through an interpretable local predictor $c_l = f(b, x)$ derived from the black-box b and the instance x using some process $f(\cdot, \cdot)$.

For example, choosing the domain $\mathcal{E}$ to be the set of linear models, the outcome explanation amounts to identifying a linear model approximating well the black-box model in the neighbourhood of x. The identified linear model can be interpreted as a local explanation of the black-box on the instance x. Approaches such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) belong to this class.

2.2. Fairness in machine learning

To explain the fairness of machine models, various metrics have been proposed in recent years (Narayanan, 2018; Berk et al., 2018). Roughly there are two distinct families of fairness definitions: *group fairness* and *individual fairness*, which requires decisions to be consistent for individuals that are similar (Dwork et al., 2012). In this work, we focus on group fairness that requires the approximate equalization of a particular statistical property across groups defined according to the value of a sensitive attribute. In particular, if the sensitive attribute is binary, its value splits the population into a minority group and a majority group.

One of the common group fairness metrics is *demographic parity* (Calders et al., 2009), which is defined as the equality of distribution of decisions conditioned to the sensitive attribute. *Equalized odds* (Hardt et al., 2016) is another common criterion requiring the false positive and false negative rates be equal across the majority and minority groups. In a nutshell, demographic parity quantifies biases in both training data and learning, while equalized odds focus on the learning process. In this paper, we adopt demographic parity to evaluate the overall unfairness in algorithmic decision-making.

Most of the existing tools for quantifying the fairness of a machine learning model do not address the interpretability issue at the same time. Possible exceptions include AI Fairness 360 that provides visualizations of attribute bias and feature bias localization (Bellamy et al., 2018), DCUBE-GUI (Berendt & Preibusch, 2012) that has an in-

terface for visualizing the unfairness scores of association rules resulted from data mining as well as `FairML` (Adebayo, 2016) that computes the relative importance of each attribute into the prediction outcome. While these tools actually provide a form of model inspection, they cannot be used directly for model or outcome explanations.

2.3. Rule Lists

A rule list (Rivest, 1987; Letham et al., 2015; Angelino et al., 2018) $d = (d_p, \delta_p, q_0, K)$ of length $K \geq 0$ is a $(K + 1)-$tuple consisting of K distinct association rules $r_k = p_k \rightarrow q_k$, in which $p_k \in d_p$ is the antecedent of the association rule and $q_k \in \delta_p$ its corresponding consequent, followed by a default prediction q_0. An equivalent way to represent a rule list is in the form of a decision tree whose shape is a comb. Making a prediction using a rule list means that its rules are applied sequentially until one rule triggers, in which case the associated outcome is reported. If no rules are trigerred, then the default prediction is reported.

2.4. Optimal rule list and enumeration of rule lists

CORELS (Angelino et al., 2018) is a supervised machine learning algorithm which takes as input a training set with n predictor variables, all assumed to be binary. First, it represents the search space of the rule lists as a n-level trie whose root node has n children, formed by the n predictor variables, each of which has $n - 1$ children, composed of all the predictor variables except the parent, and so on. Afterwards, it considers for a rule list $d = (d_p, \delta_p, q_0, K)$, an objective function defined as a regularized empirical risk:

$$R(\cdot) = \mathsf{misc}(\cdot) + \lambda K,$$

in which $misc(\cdot)$ is the misclassification error and $\lambda \geq 0$ the regularization parameter used to penalize longer rule lists. Finally, CORELS selects the rule list that minimizes the objective function. To realize this, it uses an efficient branch-and-bound algorithm to prune the trie. In addition, on high-dimensional datasets (*i.e.*, with a large number of predictor variables), CORELS uses both the number of nodes to be evaluated in the trie and the upper bound on the remaining search space size as a stopping criterion to identify suboptimal rule lists.

In (Hara & Ishihata, 2018), the authors propose an algorithm based on the Lawler's framework (Lawler, 1972), which allows to successively compute the optimal solution and then construct sub-problems excluding the solution obtained. In particular, the authors have combined this framework with CORELS (Angelino et al., 2018) to compute the exact enumeration for rule lists. In a nutshell, this algorithm takes as input the set T of all antecedents. It maintains a heap $\mathcal{H}$, whose priority is the objective function value of the rule lists and a list $\mathcal{M}$ for the enumer-

ated models. First, it starts by computing the optimal rule $m = \mathsf{CORELS}(T) = (d_p, \delta_p, q_0, K)$ for T using the CORELS algorithm. Afterward, it inserts the tuple $(m, T, \emptyset)$ into the heap. Finally, the algorithm repeats the following steps until the stopping criterion is met:

- (Step 1) Extract the tuple (m, S, F) with the maximum priority value from $\mathcal{H}$.

- (Step 2) Output m as the $i-$th model if $m \notin \mathcal{M}$.

- (Step 3) Branch the search space: compute and insert $m' = \mathsf{CORELS}(S \setminus \{t_j\})$ into $\mathcal{H}$ for all $t_j \in \delta_p$.

3. Rationalization

In this section, we first formalize the problem of the rationalization before introducing `LaundryML`, our regularized enumeration algorithm that can be used to perform fairwashing with respect to black-box machine learning models.

3.1. Problem formulation

A typical scenario that we envision is the situation in which a company wants to perform fairwashing because it is afraid of a possible audit evaluating the fairness of the machine learning it uses to personalize the services provided to its users. In this context, rationalization consists in finding an interpretable surrogate model c approximating a black-box model b, such that c is fairer than b. To achieve fairwashing, the surrogate model obtained through rationalization could be shown to the auditor (*e.g.*, an external dedicated entity or the users themselves) to convince him that the company is "clean". We distinguish two types of rationalization problems, namely the model rationalization problem and the outcome rationalization problem, which we define hereafter.

Definition 3 (Model rationalization). Given a black-box model b, a set of instances X and a sensitive attribute s, the model rationalization problem consists in finding an explanation $E \in \mathcal{E}$ belonging to a human-interpretable domain $\mathcal{E}$, through an interpretable global predictor $c_g = f(b, X)$ derived from the black-box b and the instances X using some process $f(\cdot, \cdot)$, such that $\epsilon(c_g, X, s) > \epsilon(b, X, s)$, for some fairness evaluation metric $\epsilon(\cdot, \cdot, \cdot)$.

Definition 4 (Outcome rationalization). Given a black-box model b, an instance x, a neighborhood $\mathcal{V}(x)$ of x and a sensitive attribute s, the outcome rationalization problem consists in finding an explanation $e \in \mathcal{E}$, belonging to a human-interpretable domain $\mathcal{E}$, through an interpretable local predictor $c_l = f(b, x)$ derived from the black-box b and the instance x using some process $f(\cdot, \cdot)$, such that $\epsilon(c_l, \mathcal{V}(x), s) > \epsilon(b, \mathcal{V}(x), s)$, for some fairness evaluation metric $\epsilon(\cdot, \cdot, \cdot)$.

In this paper, we will refer to the set of instances for which

the rationalization is done as the suing group. For the sake of clarity, we also restrict the context to binary attributes and to binary classification. However, our approach could also be used to multi-valued attributes. For instance, multi-valued attributes can be converted to binary ones using a standard approach such as a one-hot encoding. For simplicity, we also assume that there is only one sensitive attribute, though our work could also be extended to multiple sensitive attributes.

3.2. Performance metrics

We evaluate the performance of the surrogate by taking into account (1) its fidelity with respect to the original black-box model and (2) its unfairness with respect to a predefined fairness metric.

For the model rationalization problem, the *fidelity* of the surrogate c_g is defined as its accuracy relative to b on X (Craven & Shavlik, 1996), which is

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

For the outcome rationalization problem, the fidelity of the surrogate c_l is 1 if $c_l(x) = b(x)$ and 0 otherwise.

With respect to fairness, among the numerous existing definitions (Narayanan, 2018), we rely on the *demographic parity* (Calders et al., 2009) as the fairness metric. Demographic parity requires the prediction to be independent of the sensitive attribute, which is

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

in which $\hat{y}$ is the predicted value of c_g for $x \in X$ for the rationalization problem (respectively, the predicted value of c_l for $x \in \mathcal{V}(x)$ for the outcome problem. Therefore, *unfairness* is defined as

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

We consider that rationalization leads to *fairwashing* when the fidelity of the interpretable surrogate model is high while at the same time its level of unfairness is significantly lower than the original black-box algorithm.

3.3. `LaundryML`: a regularized model enumeration algorithm

To rationalize unfair decisions, we need to find a good surrogate model that has high fidelity and low unfairness at the same time. However, finding a single optimal model is difficult in general; indeed it is rarely the case that a single model achieves optimality with respect to two different criteria at the same time. To bypass this difficulty, we adopt a model enumeration approach. In a nutshell, our approach

works by first enumerating several models with sufficiently high fidelity to the original black-box model and low unfairness. Afterwards, the approach picks up the model that is convenient for rationalization based on some other criteria or through human inspection.

In this section, we introduce `LaundryML`, a regularized model enumeration algorithm that enumerates rule lists to achieve rationalization. More precisely, we propose two versions of `LaundryML`, one that solves the model rationalization problem and the other that solves the outcome rationalization.

`LaundryML` is a modified version of the algorithm presented in Section 2.4, which considers both the fairness and fidelity constraints in its search. The algorithm is summarized in Algorithm 1.

Algorithm 1 `LaundryML`

1: Inputs: T, λ, β
2: Output: $\mathcal{M}$
3: $\text{obj}(\cdot) = (1 - \beta)\text{misc}(\cdot) + \beta\text{unfairness}(\cdot) + \lambda K$
4: Compute $m = \text{CORELS}(\text{obj}, T) = (d_p, \delta_p, q_0, K)$
5: Insert $(m, T, \emptyset)$ into the heap
6: $\mathcal{M} \leftarrow \emptyset$
7: **for** $i = 1, 2, \dots$ **do**
8:     Extract (m, S, F) from the heap
9:     **if** $m \notin \mathcal{M}$ **then**
10:        $\mathcal{M} \leftarrow \mathcal{M} \cup \{m\}$
11:     **end if**
12:     **if** $\text{Terminate}(\mathcal{M}) = true$ **then**
13:        **break**
14:     **end if**
15:     **for** $t_j \in d_p$ and $t_j \notin F$ **do**
16:        Compute $m' = \text{CORELS}(\text{obj}, S \setminus \{t_j\})$
17:        Insert $(m', S \setminus \{t_j\}, F)$ into the heap
18:        $F \leftarrow F \cup \{t_j\}$
19:     **end for**
20: **end for**

Overall, `LaundryML` takes as inputs the set of antecedents T and the regularization parameters λ and β, respectively for the rule lists length and the unfairness. First, it redefines the objective function for CORELS to penalize both the rule list's length and unfairness (line 3):

$$\text{obj}(\cdot) = (1 - \beta)\text{misc}(\cdot) + \beta\text{unfairness}(\cdot) + \lambda K,$$

in which $\beta \in [0, 1]$ and $\lambda \geq 0$. Afterwards, the algorithm runs the enumeration routine described in Section 2.4 using the new objective function (lines 4–20). Using `LaundryML` to solve the model rationalization problem as well as the outcome rationalization problem is straightforward as described hereafter.

The model rationalization algorithm depicted in Algorithm 2 takes as input a black-box access to the black box model b,

the suing group X as well as the regularization parameters λ and β. First, the members of the suing group are labeled using the predictions of b. Afterwards, the labeled dataset T obtained is passed to the `LaundryML` algorithm.

Algorithm 2 `LaundryML-global`

1: Inputs: X, b, λ, β
2: Output: $\mathcal{M}$
3: $y = b(X)$
4: $T = \{X, y\}$
5: $\mathcal{M} = \texttt{LaundryML}(\mathsf{T}, \lambda, \beta)$

The rationalized outcome explanation presented in Algorithm 3 is similar to the rationalized model explanation, but instead of using the labelled dataset, it directly computes for the considered subject x its neighbourhood $T_x = \mathsf{neigh}(x, T_X)$ and uses it for the model enumeration.

Algorithm 3 `LaundryML-local`

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

In the next section, we report on the results that we have obtained with these two approaches on real-world datasets.

4. Experimental evaluation

In this section, we describe the experimental setting used to evaluate our rationalization algorithm as well as the results obtained. More precisely, we evaluate the unfairness and fidelity of models produced using both model rationalization and outcome rationalization. The results obtained confirm that it is possible to systematically produce models that are fairer than the black-box models they approximate while still being faithful to the original model.

4.1. Experimental setting

Datasets. We conduct our experiments on two real-world datasets that have been extensively used in the fairness literature due to their biased nature, namely *Adult Income* (Frank & Asuncion, 2010) and the *ProPublica Recidivism* (Angwin et al., 2016) datasets. These datasets have been chosen because they are widely used in the FAT community due to the gender bias (respectively racial bias) in the Adult (respectively the ProPublica Recidivism) dataset. The Adult Income dataset gathers information about individuals from the 1994 U.S. census, and the objective of the classification task is to predict whether an individual makes more or less than $50,000\$ per year. Overall, this dataset contains $32,561$ instances and 14 attributes. The ProPublica Recidivism dataset gathers criminal offenders records in Florida

during 2013 and 2014 with the classification task considered being to predict whether or not a subject will re-offend within two years after the screening. Overall, this dataset contains $6,167$ instances and 12 attributes.

Metrics. We use both fidelity and unfairness (as defined in Section 3.2) to evaluate the performance of `LaundryML-global` and `LaundryML-local`. In addition to these metrics, we also assess the performance of our approach by auditing both the black-box classifier and the best-enumerated models using FairML (Adebayo, 2016). In a nutshell, FairML is an automated tool requiring black-box access to a classifier to determine the relative importance of each attribute into the prediction outcome.

Learning the black-box models. We first split each dataset into three subsets, namely the training set, the suing group and the test set. After learning the black-box models on the training sets, we apply them to label the suing groups, which in turn are used to train the interpretable surrogate models. Finally, the test sets are used to evaluate the accuracy and the unfairness of the black-box models. In practice for both datasets, we rely on a model trained using *random forest* to play the role of the black-box models. Table 1 summarizes the performances of the black-box models that have been trained.

Scenarios investigated. As explained in Section 2, CORELS requires binary features. After the preprocessing, we obtain respectively 28 and 27 antecedents for the Adult Income and ProPublica Recidivism dataset. In the experiments, we considered the following two scenarios.

(S1) Responding to a suing group. Consider the situation in which a suing group complained for unfair decisions and its members request explanations on how the decisions affecting them were made. In this scenario, a single explanation needs to consistently explain the entire group as much as possible. Thus, we rationalize the decision using model rationalization based on Algorithm 2 (`LaundryML-global`).

(S2) Responding to individual claim. Imagine that a subject files a complaint following a decision that he deems unfair and that he requests an explanation on how the decision was made. In this scenario, we only need to respond to each subject independently. Thus, we rationalize the decision using outcome rationalization based on Algorithm 3 (`LaundryML-local`).

Experiments were conducted on an Intel Core i7 (2.90 GHz, 16GB of RAM). The modified version of CORELS is implemented in C++ and based on the original source

Table 1. Performances of the black-box models. Columns report the accuracy of the black-box models on their respective training set, suing group, and test set, as well as their unfairness on the suing groups.

Dataset	Accuracy			Unfairness	
	Training Set	*Suing Group*	*Test Set*	*Suing Group*	*Test Set*
Adult Income	85.22	84.31	84.9	0.13	0.18
ProPublica Recidism	71.42	65.86	67.61	0.17	0.13

code of CORELS[1]. Algorithm 1, Algorithm 2 as well as Algorithm 3 were all implemented in Python. To implement Algorithm 1, we modified the Lasso enumeration algorithm[2] of (Hara & Maehara, 2017). All our experiments can be reproduced using the code provided in https://github.com/aivodji/LaundryML.

For the scenario (S1), we use regularization parameters with values within the following ranges $\lambda = \{0.005, 0.01\}$ and $\beta = \{0.0, 0.1, 0.2, 0.5, 0.7, 0.9\}$ for both datasets, yielding 12 experiments per dataset. For each of these experiments, we enumerate 50 models.

For the scenario (S2), we use the regularization parameters $\lambda = 0.005$ and $\beta = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ for both datasets. For each dataset, we produce a rationalized outcome explanation of each rejected subject belonging to the minority groups (*i.e.*, female subjects of Adult Income's suing group and black subjects of ProPublica Recidivism's suing group), and for whom the unfairness as measured in their neighbourhood is greater than 0.05. Overall, they are 2,634 (respectively 285) such subjects in the Adult Income's suing group (respectively ProPublica Recidivism's suing group), yielding a total of 13,170 (respectively 1,425) experiments for the two datasets. For each of these experiments, we also enumerate 50 models and select the one (if it exists) that has the same outcome as the black-box model and the lowest unfairness. To compute the neighbourhood neigh($\cdot$) of each subject, we apply the k-nearest neighbour algorithm with k set to 10% of the size of the suing group. Although it would be interesting to see the effect of the size of the neighbourhood on the performance of the algorithm, we leave this aspect to be explored in future works.

4.2. Experimental results

Summary of the results. Across the experiments, we observed the fidelity-fairness trade-off by performing model enumeration LaundryML-global. By visualizing this trade-off, we can select the threshold of the fairness condition easily. In some cases, the best models we found out significantly decrease the unfairness while retaining a high fidelity with the black-box model. We have also checked that the sensitive attributes have a low impact on the surro-

gate models' predictions compared to the original black-box ones. This confirms the effectiveness of the model enumeration, in the sense that we can identify a model convenient for rationalization out of the enumerated models. In addition, by using LaundryML-local, we can obtain models explaining the outcomes for users in the minority groups while being fairer than the black-box model.

(S1) Responding to a suing group. Figure 1 represents the results obtained on the scenario (S1) based on LaundryML-global, by depicting the changes of fidelity and unfairness of the models accompanied by the fairness regulation parameter β. From these results, we observe that as β increases, both the unfairness and the fidelity of the enumerated models decrease. In addition, as the enumerated models get more complex (*i.e.*, λ decreases), their fidelity increases. On both datasets, we selected the best model using the following strategy. First, we select models whose unfairnesses are at least two times less than the original black-box model and then among those models, and we select the one with the highest fidelity as the *best model*. On Adult Income, the best model has a fidelity of 0.908 and an unfairness of 0.058. On ProPublica Recidivism, the best model has a fidelity of 0.748 and an unfairness of 0.080.

In addition, we compared the obtained best models with the corresponding black-box models using FairML. With FairML, we ranked the importance of each feature to the model's decision process. Figure 3 shows that the best obtained surrogate models were found to be fair by FairML: in particular the rankings of the sensitive attributes (*i.e.*, gender:Male for Adult Income, and race:Caucasian for ProPublica Recidivism) got significantly lower in the surrogate models. In addition, we observe a shift from the 2nd position to the last one on Adult Income as well as a change from the 5th rank to the 13th rank on ProPublica Recidivism.

(S2) Responding to individual claim. Table 2 as well as Figure 2 demonstrate the results on (S2) by using LaundryML-local. From these results, we observe that as the fairness regulation parameter β increases, the unfairness of the enumerated models decreases. In particular, with $\beta = 0.9$, a completely fair model (unfairness = 0.0) was found to explain the outcome for each rejected user of the
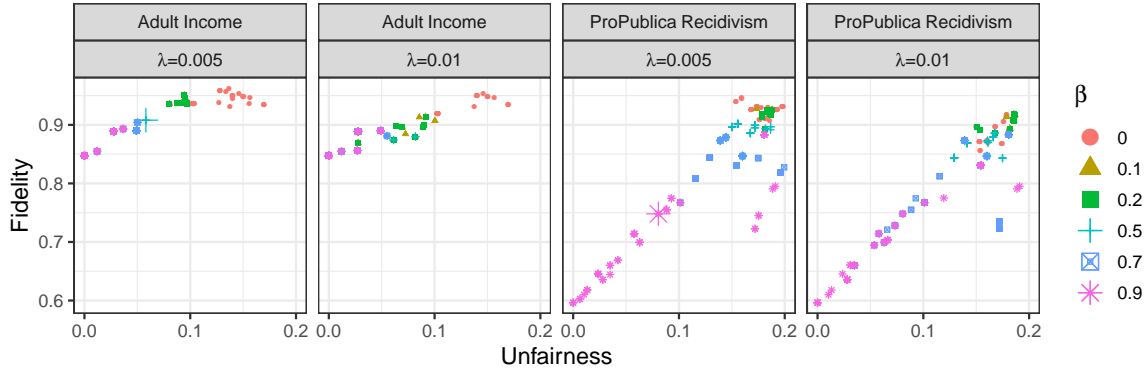
---

[1]https://goo.gl/qTeUAu
[2]https://goo.gl/CyEU4C

Figure 1. Fidelity and unfairness of rationalized explanation models produced by `LaundryML-global` on the suing groups of Adult Income and ProPublica Recidivism datasets.
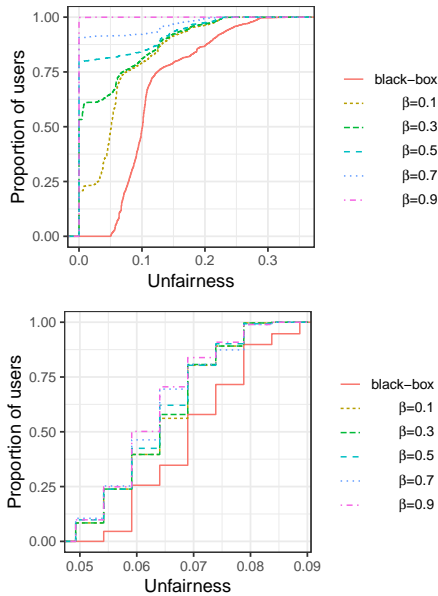

Figure 2. CDFs of the unfairness of the best model found by `LaundryML-local` per user on Adult Income (up) and ProPublica Recidivism (down).

Table 2. Proportion of users of the minority group for whom there exists at least one models that agrees with the black-box models.

	Adult Income	**ProPublica Recidism**
$\beta = 0.1$	92.90%	100%
$\beta = 0.3$	94.95%	100%
$\beta = 0.5$	97.72%	100%
$\beta = 0.7$	99.16%	100%
$\beta = 0.9$	100%	100%

minority group of Adult Income. For ProPublica Recidivism, the enumerated model found for each user is at least twice less unfair than the black-box models.

Remark. Our preliminary experiments in (S1) show that the fidelity of the model rationalization on the test set tends to be slightly lower than the one on the suing group, which means that the explanation is customized specifically for the suing group. Thus, if the suing group gather additional members on which the existing explanation is applied, the explanation may not be able to rationalize as well those additional members. This opens the way for future research directions such as developing methods for detecting that an explanation is actually a rationalization or conversely obtaining rationalizations that are difficult to be detected.

5. Conclusion

In this paper, we have introduced the rationalization problem and the associated risk of fairwashing in the machine learning context and shown how it can be achieved through model explanation as well as outcome explanation. We have also proposed a novel algorithm called `LaundryML` for regularized model enumeration of rules lists, which incorporate fairness as a constraint along with the fidelity. Experimental results obtained using real-world datasets demonstrate the feasibility of the approach. Overall, we hope that our work will raise the awareness of the machine learning community and inspire future research towards the ethical issues raised by the possibility of rationalizing, in particular with respect to the risk of performing fairwashing in this context.

Our framework can be extended to other interpretable models or fairness metrics as we show in the additional experiments provided in appendices. As future works, along with ethicists and law researchers, we want to investigate the general social implications of fairwashing in order to develop a comprehensive framework to reason on this issue. Finally, as mentioned previously we also aim at developing approaches that can estimate whether or not an explanation is likely to be a rationalization in order to be able to detect fairwashing.
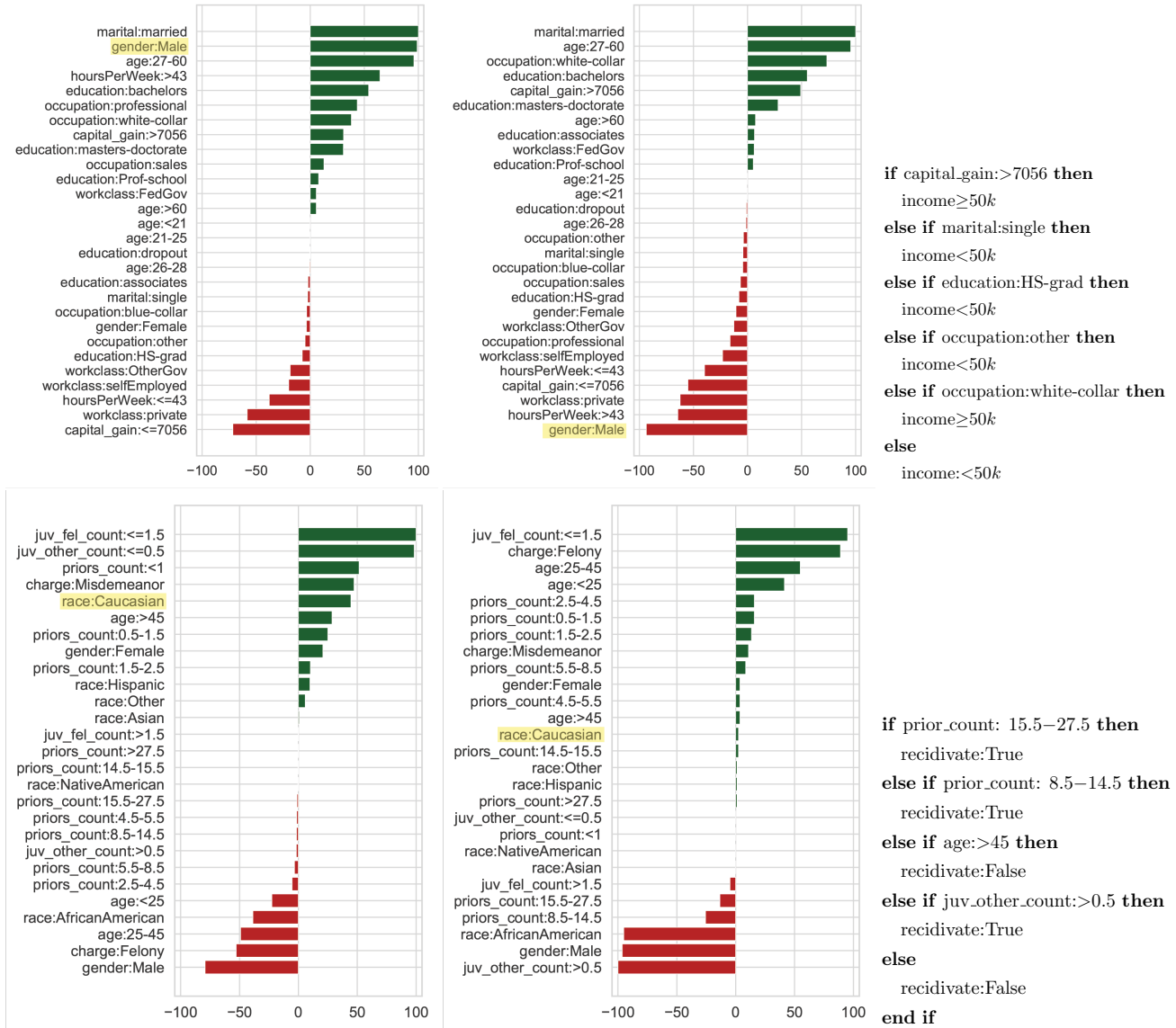
Figure 3. Relative feature dependence ranking obtained using FairML to audit models trained on the Adult Income dataset (upper panel) and the ProPublica Recidivism dataset (lower panel). Green indicates that the feature highly contributes to a high salary rating on Adult (respectively, a low recidivism rating on ProPublica Recidivism). Features that characterize the majority groups are highlighted in yellow. Blackbox model (left) vs. LaundryML model (middle) and its description (right).

Acknowledgements

References

Adebayo, J. A. Fairml: Toolbox for diagnosing bias in predictive modeling. Master's thesis, Massachusetts Institute of Technology, 2016.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18 (234):1–78, 2018.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica, May*, 23, 2016.

Barocas, S. and Selbst, A. D. Big data's disparate impact. *California Law Review*, 104:671, 2016.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Berendt, B. and Preibusch, S. Exploring discrimination: A user-centric evaluation of discrimination-aware data mining. In *Proceedings of the IEEE 12th International Conference on Data Mining Workshops (ICDMW'12)*, pp. 344–351. IEEE, 2012.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

Breiman, L. *Classification and regression trees*. Routledge, 2017.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW'09)*, pp. 13–18. IEEE, 2009.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, pp. 1721–1730. ACM, 2015.

Craven, M. and Shavlik, J. W. Extracting tree-structured representations of trained networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS'96)*, pp. 24–30, 1996.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Conference on Innovations in Theoretical Computer Science*, pp. 214–226. ACM, 2012.

Edwards, L. and Veale, M. Slave to the algorithm: why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16:18, 2017.

Frank, A. and Asuncion, A. Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california. *School of information and computer science*, 213:2–2, 2010.

Goodman, B. and Flaxman, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):93, 2018.

Hara, S. and Ishihata, M. Approximate and exact enumeration of rule models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI'18)*, pp. 3157–3164, 2018.

Hara, S. and Maehara, T. Enumerate Lasso solutions for feature selection. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 1985–1991, 2017.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS'16)*, pp. 3315–3323, 2016.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293, 2017.

Lawler, E. L. A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem. *Management science*, 18(7): 401–405, 1972.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., and Vinck, P. Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, pp. 1–17, 2017.

Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

Li, J., Shen, H., and Topor, R. Mining the optimal class association rule set. *Knowledge-Based Systems*, 15(7): 399–405, 2002.

Lipton, Z. C. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018. doi: 10.1145/3233231. URL https://doi.org/10.1145/3233231.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS'17)*, pp. 4765–4774, 2017.

Melis, D. A. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pp. 7775–7784, 2018.

Montavon, G., Samek, W., and Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

Narayanan, A. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference on Fairness Accountability Transparency (FAT*)*, 2018.

Reddix-Smalls, B. Credit scoring and trade secrecy: An algorithmic quagmire or how the lack of transparency in complex financial models scuttled the finance market. *UC Davis Bus. LJ*, 12:87, 2011.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pp. 1135–1144. ACM, 2016.

Rivest, R. L. Learning decision lists. *Machine learning*, 2 (3):229–246, 1987.

Rudin, C. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*, 2018.

Siddiqi, N. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons, 2012.

Wachter, S., Mittelstadt, B., and Floridi, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

Wexler, R. When a computer program keeps you in jail: How computers are harming criminal justice. *New York Times*, 2017a.

Wexler, R. Code of Silence. *Washington Monthly*, 2017b.