Fighting the Fair Fight: An Overview of the Black-Box Fairness Auditing Techniques

by Shubham Singh

Dhruv Shah

January 31, 2025

Table of contents

Preliminaries

Basic Definitions

1. Machine Learning Model - An algorithm or function that processes a dataset to output classifications, learning from a training set by adjusting weights based on a loss function to minimize errors.
2. Sensitive Attributes - Features such as race, gender, or religion that, if used in decision-making, could lead to discrimination or unfair treatment of individuals.
3. Proxy Attributes - Features that are not inherently sensitive but are correlated with sensitive attributes, potentially leading to indirect discrimination if utilized by the model.

Basic Definitions

4. Disparity - The unequal treatment or outcomes experienced by different groups within a population, often due to biases in data or model design. On a granualar level, it can be broken down into two parts.
   - 4.1 Disparate Treatment - Disparity occurs at the model or input stage of the decision making process.
   - 4.2 Disparate Impact - Disparity occurs at the outcome stage of the decision making process. Occurs when the relationship between sensitive and proxy variables is not explicitly observable.

Overview of Three Fairness Auditing Techniques

FairTest: Motivation

Motivation: Staples, a retail company, used an variable pricing algorithm for marketing purpose. Later, the algorithm was found to be discriminating against people living in lower-income neighborhoods by showing them higher prices.
This situation was termed as "unintended consequence" of their algorithm. Such problems are known as *bugs* or *unwarranted associations*

Unwarranted Associations

Unwarranted Association are formally defined as:

- strong associations between the algorithm output and the attributes of a protected user group.
- the associations arise in a meaningful subset of users.
- lack explanatory factors (eg. correlations unrelated to actual causal relationships)
- generalize well over different datasets (wide scope applicability)

Approach

To inspect an algorithm's output O, data attributes are categorized into three types:

- Protected Attributes (S) - Attributes where discrimination can occur (e.g., gender, race).
- Contextual Attributes (X) - Attributes that split the population to uncover hidden biases. (e.g., location or education level revealing race).
- Explanatory Attributes (E) - Justify seemingly discriminatory outcomes (e.g., experience justifying hiring decisions).

Example: In hiring, experience (E) may justify bias toward older candidates, but it can also act as a proxy (X) for age (S).

Note: The categorization of S, X, and E is task-dependent and subjective.

Metrics for Association Strength

Based on the values that O and S can take, the authors classify the choice of metrics that can be used to assess the strength of the association.

- Frequency Distribution: Measures disparity when both output (O) and sensitive attribute (S) are binary. Uses ratio $\frac{Pr(o_1|s_1)}{Pr(o_2|s_2)} - 1$ and difference $Pr(o_1|s_1) - Pr(o_2|s_2)$.
- Mutual Information (MI): Captures statistical dependence between O and S for non-binary attributes. Normalized MI (NMI) adjusts for entropy differences.
- Correlation: Uses Pearson's correlation to quantify linear dependence between O and S, helping in error profiling.
- Regression: Estimates strength of association when O has a large value space by analyzing regression coefficients.
- Conditional Metrics: Evaluates unwarranted associations given an explanatory attribute (E) using $E_E(M(S;O)|E)$.

Fair Test Design

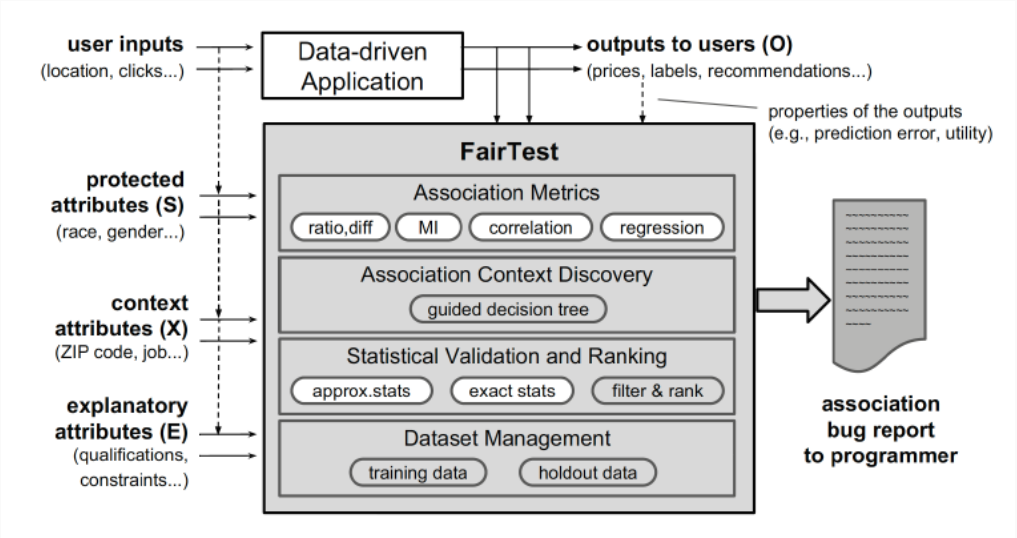

Figure 1: Architecture of FairTest[5]

Gradied Feature Auditing: Motivation

Adler et. al propose a Gradient Feature Auditing (GFA) algorithm to identify the *indirect* influence of proxy variables in the decision the outcome of a black-box model.

Idea: *The information content of a feature can be estimated by trying to predict it from the remaining features.* (Idea from [2])

How does this quantify the influence of a feature?
Minimally perturb/modify the data so that the feature can no longer be predicted from the remaining data. Hence, eliminating the influence of this feature both directly and indirectly.

Contributions

Authors create a modified dataset with minimal indirect influence and observing the models' performance. The contributions of this method is:

- An algorithm to construct a modified dataset with obscured indirect influence, with theoretical support.
- Formal definition of indirect influence in terms of black-box model outcomes.
- Evaluation of the approach on multiple publicly available datasets.

Note: Here the model cannot be retrained, and the focus is on *quantifying* the influence of a feature.

Approach

Indirect Influence: Termed as the influence of a non-sensitive attribute on the model outcome that is not classified as sensitive alone, but has a veiled relationship to the sensitive attribute.

- Let $f : X \to Y$ be a black-box classification function.
- X is a d-dimensional feature space, and Y is the domain of outcomes (e.g., $\{-1, 1\}$ for binary classification).
- Dataset: $(X, Y) = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$.
- Accuracy of f:

$$\text{acc}(X, Y, f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{[y_i \neq f(x_i)]}.$$

- Indirect influence of feature j: The change in accuracy caused by removing or obscuring X_j.

Typical approach is to perturb the j^{th} feature (usually done by random perturbation).

Challenges with Random Perturbations:

- Randomly perturbing features disrupts indirect influence and removes useful task-related information.
- Proxy features can degrade classification quality, making it hard to isolate the effect of X_j.

Proposed Approach

Proposed Approach:

- Perturb X deterministically to eliminate direct and indirect influence of j.
- Organized around the question: *Can we predict the value of feature j from the remaining features?*
- If X_j cannot be predicted, its influence is considered eliminated.
- Modify the dataset minimally to obscure X_j.

Definitions

Balanced Error Rate (BER):

$$BER(X, Y, f) = \frac{1}{|\text{supp}(Y)|} \sum_{j \in \text{supp}(Y)} \frac{\sum_{y_i = j} \mathbb{I}(f(X_i) \neq j)}{|\{i | y_i = j\}|}$$

Measures robustness to class imbalance by averaging error rates per class.

ϵ-Obscure Feature:

$$BER(X \setminus X^{(i)}, X^{(i)}, f) > \epsilon$$

A feature $X^{(i)}$ is ϵ-obscure if it cannot be predicted using the remaining features in X.

Indirect Influence (II):

$$II(i) = \text{acc}(X, Y, f) - \text{acc}(X \setminus X^{(i)}, Y, f)$$

Quantifies the change in accuracy when feature i is removed.

GFA Algorithm: Computing the influence

Consider feature $O = X_i$ to be removed is categorical and $W = X_j$ to be obscured is numerical. Let $W_x = Pr(W|O = x)$ and the cummulative distribution be $F_x(w) = Pr(W \geq w|O = x)$.
Define a *median distribution* A. Its cumulative distribution F_A is given by $F_A^{-1} = median_{x \in O} F_x^{-1}(u)$. Now, the modification is achieved by changing the values of W to mimic the median distribution A, such that $\hat{W} = F_A^{-1}(F_x(w))$.
Authors refer [2], where it is shown that such a modification obscures O maximally, while keeping the change in W minimal.

FlipTest: Motivation

Classic fairness metrics like demographic parity and equalized odds fail to capture discriminatory behaviour towards individuals.

This technique is motivated by the question: *had an individual been of a different protected status, would the model have treated them differently?*

Some previous ([3]) work studies causal relationships with protected attribute to understand model discrimination. But then such causal graphs can be difficult to build. Also, we miss other features which may not be causal but are correlated.

Approach

Now, just flipping the protected attribute is not sufficient to assess the discriminatory behaviour because the model can learn this attribute from the proxy attributes.

Example: The authors motivate their approach using an example of a synthetic dataset created by *Lipton et. al* [4]. Features: Hair Length and Work Experience. Classifier: Decide whether a person should be hired. Our Task: Investigate the possibility of a gender bias in the model (Note: The model satisfies demographic parity fairness)

- Map set of women to their male correspondents.
- Analyze the cases where women are treated differently than their male counterparts.

Question: How to do the mapping?

For example, A man with 10 years of work experience cannot be mapped to a woman with no work experience. Therefore, just because two people are treated differently does not mean discrimination has occurred. *(disparate treatment might not always cause disparate impact.)*

This motivates the use of an optimal transport mapping. Intuitively, it is minimizing the sum of the distances between a woman and the man that she is mapped to (her counterpart), where the distance quantifies how different a pair of people are.

Then analyse the optimal transport mapping using *flipsets*

Optimal Transport Mapping

Let Consider the two distributions **S** and **S**$'$ for two classes over the feature space $\mathcal{X}$.

Let n be the number of samples drawn from these two distributions, such that the set $S = \{x_1, \ldots, x_n\}$ and set $S' = \{x'_1, \ldots, x'_n\}$, where $n = |S| = |S'|$. The cost function is defined as $c : \mathcal{X} \times \mathcal{X} \to [0, \infty]$, such that $c(x, x')$ denotes the cost of moving a point from S to S'. An optimal transport map accomplishes moving between two points by minimizing the cost function. It is given as a bijection $f : S \to S'$ such that the expected cost is minimized:

$$\mathbb{E}[c(x, f(x))] = \frac{1}{n} \sum_{i=1}^{n} c(x_i, f(x_i))$$

Flipsets

Flipset consist of all women whose outcomes were different from their counterparts'.
Definition: Let $h : X \to \{0, 1\}$ be a classifier and $G : S \to S'$ be an optimal transport mapping (or an approximation). The flipset $F(h, G)$ is the set of points in S whose mapping into S' under G changes classification.

$$F(h, G) = \{x \in S \mid h(x) \neq h(G(x))\}$$

The positive and negative partitions of $F(h, G)$ are denoted by $F^+(h, G)$ and $F^-(h, G)$.

$$F^+(h, G) = \{x \in S \mid h(x) > h(G(x))\}$$

$$F^-(h, G) = \{x \in S \mid h(x) < h(G(x))\}$$

If S and S' are equal, then G would be an identity function and then as per the definition of $F(h, G)$, we would expect empty flipsets. Implies model cannot discriminate based on protected attributes.

Notice, if the size of positive and negative flipset are non-zero, it implies demographic parity (For Proof, refer [1])

When h is biased, we get the following useful information

- Relative sizes of $F^+(h, G)$ and $F^-(h, G)$ give an idea group fairness.
- Their absolute sizes hint to possible discrimination at sub-group/individual level.
- if the distributions of the flipsets are different from S, we gain information about which subgroup may be discriminated against.

Experiment and Result

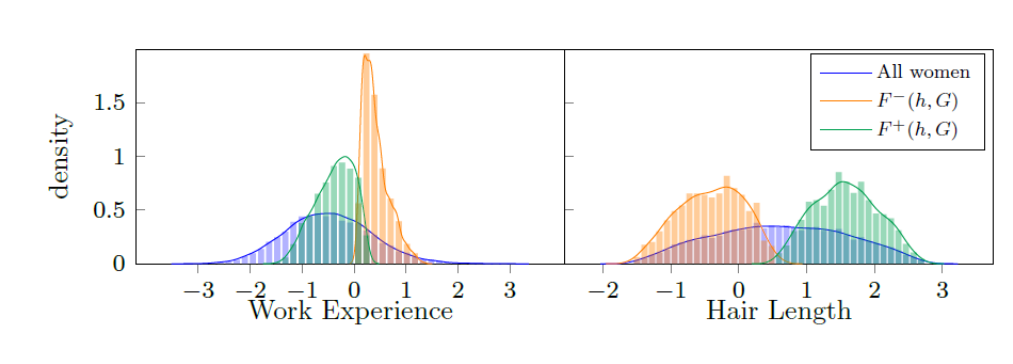

Figure 2: Caption

References i


E. Black, S. Yeom, and M. Fredrikson.
Fliptest: fairness testing via optimal transport.
In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 111–121, New York, NY, USA, 2020. Association for Computing Machinery.


M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian.
Certifying and removing disparate impact, 2015.


M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva.
Counterfactual fairness, 2018.


Z. C. Lipton, A. Chouldechova, and J. McAuley.
Does mitigating ml's impact disparity require treatment disparity?, 2019.

References ii


F. Tramèr, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin.
Fairtest: Discovering unwarranted associations in data-driven applications, 2016.

THANK YOU!