Fairness by Explicability and Adversarial SHAP Learning*

James M. Hickey, Pietro G. Di Stefano, and Vlasios Vasileiou

Experian UK&I and EMEA DataLabs, London, UK
james.hickey@experian.com

Abstract. The ability to understand and trust the fairness of model predictions, particularly when considering the outcomes of unprivileged groups, is critical to the deployment and adoption of machine learning systems. SHAP values provide a unified framework for interpreting model predictions and feature attribution but do not address the problem of fairness directly. In this work, we propose a new definition of fairness that emphasises the role of an external auditor and model explicability. To satisfy this definition, we develop a framework for mitigating model bias using regularizations constructed from the SHAP values of an adversarial surrogate model. We focus on the binary classification task with a single unprivileged group and link our fairness explicability constraints to classical statistical fairness metrics. We demonstrate our approaches using gradient and adaptive boosting on: a synthetic dataset, the UCI Adult (Census) Dataset and a real-world credit scoring dataset. The models produced were fairer and performant.

1 Introduction

The last few decades have seen machine learning algorithms become even more performant and leverage larger varieties of data. These advances have led to wide-spread adoption of machine learning in nearly every industry. The potential damage and wider societal harm that could be caused by large-scale automated decisioning systems is palpable amongst regulators, industry practitioners and consumers [8,32,23]. Two specific concerns that have emerged center on the interpretability and fairness of the decisions resulting from these algorithms. These are not unjustified with cases of unfair decisioning systems manifesting in multiple domains from criminal recidivism [8] to credit worthiness assessment. In the European Union, these concerns have manifest in the General Data Protection Regulation [12,17] that enshrines each individual's right to fair and transparent processing. This combined societal and legislative scrutiny has resulted in model interpretability and algorithmic fairness coming to the fore in research [11,29].

At the broadest level, the concept of algorithmic fairness tackles whether members of specific unprivileged groups are more likely to receive unfavourable decisions from the predictions of a machine learning system. Recent advances

* Supported by Experian Ltd.

have enabled modellers to incorporate fairness at every point of the model building process [14,29,38,9]. One embodiment incorporates fairness constraints into the training procedure [18,5,10,15,46,31,27,1,27], typically these constraints rely on statistical measures of fairness and are subject to drawbacks [20] and trade-offs. These measures rely on *a priori* worldviews and do not incorporate the role of external model auditing or decision explicability in their fairness criteria. This is poorly aligned with how these issues are dealt with in industry, where external actors often question the model fairness through building surrogate explanatory models, even if mentally, using the information available to them.

To address these issues, we propose a new definition of fairness we dub "Fairness by Explicability". Under this definition, if an external actor's surrogate model cannot produce a *narrative* (i.e., a set of explanations) against the fairness of a particular model, then that particular model can be considered *explicably fair*. This definition explicitly frames the perception of an algorithm's fairness as one determined by a combination of an auditor's worldview, data availability, model interpretability framework and measurement/modelling approach. It can be considered complementary to the existing ways of evaluating a model's fairness, since while those may capture risk arising from non-adherence to regulatory requirements, our new "fairness by explicability" viewpoint captures the additional and independent risk that may arise from analyses performed by one's own clients [28].

To enforce our "Fairness by Explicability" definition, we leverage model interpretability methodologies [33,25,48] to incorporate fairness constraints through adversarial learning. More explicitly, we utilize the SHAP [25,24] values of a surrogate adversary model in two ways. The first works by constructing a differentiable fairness regularization term. The second is a modification to the classic AdaBoost algorithm [13] to include adversarial attribution values in the weight updates.

We link our fairness approach to statistical fairness [39] via the construction of an appropriate surrogate model. Our approaches are illustrated using a synthetic dataset, the UCI Adult Census Dataset [4], and a commercial credit scoring dataset[1]. These datasets present a diverse evaluation set, with the real-world dataset providing assurance that these approaches are viable in industrial applications. The structure of the papers is as follows: in Section 2 we introduce our notation; in Section 3 we provide a brief account of SHAP values and Section 4 discusses statistical fairness measures. Section 5 introduces the "Fairness by Explicability" worldview and in Section 6 we present our SHAP-regularized algorithms before discussing the results of the experiments in Section 7. We then state our conclusions and highlight areas of further research in Section 8.

2 Notation

To measure the fairness of any algorithm output one needs to define the task objective, the un-/privileged groups to measure fairness against and the favourable

[1] Private and internal to Experian

outcomes. For the remainder of this paper, we focus on binary classification tasks with a single privileged group indicator Z. We denote the other covariates present with $\mathbb{X}$ and the combination of Z with those covariates by $\tilde{\mathbb{X}}$. Furthermore, and without loss of generality, we define the value of 1 for the target Y and the corresponding model outcomes $\hat{Y}$ as the favourable label. Model outcomes are constructed by applying a threshold to the scores $\bar{Y}$. For each instance i, we denote the corresponding values with the appropriate lowercase symbol and subscript, i.e. y_i, z_i, x_i, etc. In this case, x_i and $\tilde{x}_i$ denote vectors and the value of the jth covariate is given by x_{ij} and $\tilde{x}_{ij}$.

3 SHapley Additive Explanations (SHAP)

SHapley Additive Explanations, or SHAP values [25,24], provide a unified framework for interpreting model predictions. This approach was built off the insight that many other modern explanatory frameworks such as LIME [33] and DeepLIFT [36] could be recast as variants of a generic additive feature attribution paradigm. In this paradigm, a simplified explanatory model σ is built to explain the original prediction f using simplified binary input vectors $\tilde{x}'_i \in \{0,1\}^M$, where M is the number of features and i is the instance label. These simplified inputs are related to the original feature vectors $\tilde{x}_i$ through the mapping $\tilde{x}_i = h_{\tilde{x}_i}(\tilde{x}'_i)$ and the local explanatory model is given by:

$$\sigma(\tilde{x}'_i) = \phi_0^{i,f} + \sum_{j=1}^{M} \phi_j^{i,f} \tilde{x}'_{ij}. \tag{1}$$

The local feature effect of feature j for model f is $\phi_j^{i,f}$ and global explanations are calculated via the statistics of these values across a dataset. The different explanatory frameworks, e. g. LIME, emerge from specific choices of the mapping function $h_{\tilde{x}_i}$, the kernel weighting of instances in the objective ($\pi_{\tilde{x}}$) and any additional regularization terms $\Omega(\sigma)$ used to fit σ. These choices influence the properties of the surrogate model. In Ref. [25], they showed that only one σ satisfies these 3 desirable properties:

1. Local Accuracy: $f(\tilde{x}_i) = \sigma(\tilde{x}'_i) = \sum_{j=0}^{M} \phi_j^{i,f}$, when $\tilde{x}_i = h_{\tilde{x}_i}(\tilde{x}'_i)$.
2. Missingness: $\tilde{x}'_{ij} = 0 \implies \phi_j^{i,f} = 0$.
3. Attribution Consistency: for any two models f, f', the ordering of the differences of the model output when a feature is present vs missing is reflected in their respective attributions of that feature.

Its attributions ϕ_j are the same Shapley Values first identified in cooperative game theory [35,22,42,37]:

$$\phi_j^{\bullet,f} = \sum_{z' \subseteq \tilde{x}'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_{\tilde{x}}(z') - f_{\tilde{x}}(z' \setminus j)]. \tag{2}$$

Here, $|z'|$ is the number of non-zero entries in z', $z' \setminus j$ denotes setting the jth element of z' to 0 and the summation is over all z' where the non-zero entries are a subset of the non-zero entries of $\tilde{x}'$. These SHAP values can be estimated for a generic model using KernelSHAP [25] while for specific model families there are efficient computational methods and analytic approximations [24,37].

4 Metrics and Statistical Fairness

To estimate fairness metrics one requires a dataset of N instances with Y and Z as well as the outcomes. Given this data, the appropriate fairness metric is often defined by the worldview(s) [41] of those auditing the outcomes. These worldviews tend to fall into three broad categories: "We're all equal" [2], "What you see is what you get" [11,34] and causal [19,47,21,40,10,7]. The first two categories are statistical in nature and we now discuss their application to the binary task domain.

Statistical fairness metrics relate to the conditional probabilities involving Y, $\hat{Y}$ and Z. The "We're all equal" worldview has numerous group fairness metrics associated with it. These metrics measure any differences in outcome given group membership and seek to balance said outcomes. Contrastingly, "What you see is what you get" asserts that the observed data captures the underlying "truth" and typically prefers to offer individuals similar outcomes conditional on Y. In this work, we consider two of the most common statistical fairness metrics from these categories: "statistical parity" difference (SPD) and "equality of opportunity" difference (EOD). More formally, these are defined as:

$$\text{SPD} = |P(\hat{Y} = 1|Z = 1) - P(\hat{Y} = 1|Z = 0)|, \tag{3}$$

$$\text{EOD} = |P(\hat{Y} = 1|Y = 1, Z = 1) - P(\hat{Y} = 1|Y = 1, Z = 0)|. \tag{4}$$

Note that a target SPD value can also be calculated by replacing $\hat{Y}$ with Y respectively in Eq. 3. Both of these measures are estimated from a specified dataset, their value of zero denotes a maximally fair model, and both have trade-offs [20] and limitations. For example, SPD can be minimized through randomly modifying outcomes while ignoring all other covariates $\mathbb{X}$ and so can be viewed as a lazy penalization. Contrastingly, minimizing EOD may not reduce any gap in the rate of favourable outcomes between the groups.

5 Fairness by Explicability

The traditional statistical fairness metrics presented in Section 4 are not explicitly linked to the domain of model interpretability, let alone interpretability frameworks such as SHAP [25] or LIME [33]. These measures emerge from the worldviews of individuals auditing the model outcomes for fairness. Typically, when trying to understand observations, a human agent (an external actor/auditor) will construct a surrogate model to obtain explanations for their

observations. The role of Z in these explanations determines whether the outcomes constructed are perceived as fair or not. Building on this idea, we propose a new worldview to capture the mechanism by which model decisions are evaluated by external actors.

Definition 1. *Consider a model trained by an auditor to predict $\bar{Y}$ using Z and, optionally, a combination from $\{Y, \not{X}\}$. If this model does not detect any difference in the Z attribution between the $Z = 0, 1$ groups, then the predictor model is explicably fair with respect to the auditor.*

We dub this worldview "Fairness by Explicability". The precise measure of fairness one attains is determined by: the population examined by the auditor, the interpretability framework used, how attributions are calculated and aggregated, and the auditor model developed. This definition can be specialized into a *strong* "Fairness by Explicability" form by further requiring that total attribution for Z is also reduced to zero.

Auditors are usually interested in the average attribution of the two groups given a population of data. This informs the metrics used to quantify how "explicably fair" a model is. These are:

$$\text{FE} = |\frac{\sum_{i,s.t.Z=1} \phi_Z^{i,l}}{N_1} - \frac{\sum_{i,s.t.Z=0} \phi_Z^{i,l}}{N_2}|, \tag{5}$$

$$\text{SFE} = \frac{\sum_i |\phi_Z^{i,l}|}{N}, \tag{6}$$

where $\phi_Z^{i,l}$ is the SHAP value of Z for instance i for auditor model l, N is the total number of instances in the dataset and $N_{1(0)}$ is number of examples when $Z = 1(0)$. FE measures the difference in mean attribution between the two groups. When it is minimized the model is considered fair according to our "Fairness by Explicability" definition. The second metric (SFE) measures the total attribution of Z across the population, when minimized the auditor model concludes that the model satisfies the strong version of "Fairness by Explicability" and, by definition, the first metric is also zero.

From this discussion, "Fairness by Explicability" may appear intuitive but difficult to implement and, in general, being "explicably fair" does not provide any guarantees of statistical fairness. However, an initial informal connection to the prior fairness worldviews can be made through consideration of specific form of the auditor models. Intuitively, removing the dependency on Z as measured by an external l will tend to reduce $\bar{Y}$'s dependency on Z. This will generally lead to improved SPD and EOD, although the decision policy plays a large role in how these two connect.

6 Achieving Fairness by Explicability

We now present two different approaches for imposing "Fairness by Explicability" directly into the training process of gradient-based and adaptive boosting

(specifically AdaBoost) algorithms. These approaches rely on inserting a surrogate model g directly into the iterative training procedures. The form of g is then chosen to account for the examination of an anticipated external auditor whose model is l. Both approaches require Z during the training phase only, hence any sensitive attributes defining Z do not need to be supplied at prediction time. In addition to this presentation, we also discuss how the approaches can be linked to the SPD and EOD.

6.1 SHAPSqueeze

The first approach to imposing "Fairness by Explicability" uses a series of differentiable regularizations to penalize unfair attributions. We consider a differentiable loss function of the form:

$$\mathcal{L}_{\text{fair}} = (1 - \lambda) * \mathcal{L}_o + \lambda * \mathcal{R}, \tag{7}$$

which we can optimize through gradient-based methods, e.g. stochastic gradient descent. At each iteration, a surrogate model g is fit to the $\bar{Y}$ values. From g, the SHAP values of Z, and optionally Y, are used to calculate the appropriate regularization term ($\mathcal{R}$). In this work, $\mathcal{L}_o$ is the binary cross-entropy. Considering the case where l and g are identical, when the associated $\mathcal{R}$ is minimized then the attributions to Z will be zero and *strong* "Fairness by Explicability" is satisfied by the model scores $\bar{Y}$.

The specific form of g we examine is a linear regression model, see the first row of Table 1. We approximate the SHAP values of interest by:

$$\phi_Z^{i,g} = \beta(z_i - \mathbb{E}[Z]). \tag{8}$$

Equation (8) directly relates the SHAP values of Z to its model coefficient, β, and the specific realisation of Z for instance i. The regularization $\mathcal{R}$ is then simply the sum of the squares of these SHAP values scaled by a constant C, see Table 1. This constant is used to make the size of the gradients coming from $\mathcal{R}$ and $\mathcal{L}_o$ comparable, while λ is used to adjust the balance between these two quantities. Moreover, we note that the explicability fairness metrics in Eq. 5 are proportional to β in this instance. Therefore, these specific g and $\mathcal{R}$ will seek to eliminate the linear dependence of the model predictions $\bar{Y}$ on Z. Consequently, we expect reductions in the SPD as the model becomes explicably fairer.

To conclude, we note that the use of linear regression makes both the model fitting and SHAP value derivative calculations computationally efficient to perform. However, the approach described is applicable to any g whose SHAP values are differentiable with respect to $\bar{Y}$ and so parametric/kernel regression models could also be employed. In combination with adding more features, this can allow for the consideration of more complex auditors with different worldviews.

6.2 SHAPEnforce

The classic AdaBoost algorithm [13] trains a model that is a weighted linear combination of weak classifiers. The training process is iterative, with each weak

Table 1. The surrogate models and regularizations considered in this work.

algorithm	surrogate - g	regularization
SHAPSqueeze	$\bar{Y} = \beta Z + \alpha$	$\mathcal{R} = C \sum_i (\phi_Z^{i,g})^2$
SHAPEnforce		$\mathcal{P} = \begin{cases} -\phi_Z^{i,g}, \text{if } y_i = 1 \\ 0, \text{otherwise} \end{cases}$

learner (k_m) being fitted to a reweighted version of the training data. After R iterations, the outputted model is given by $C_R = \sum_{m=1}^{R} \alpha_m k_m$. We consider learners that output a score and whose classification output, $\{0, 1\}$, is obtained by thresholding. Traditionally, AdaBoost generates the instance weights for the m^{th} training round, ω_i^m, by scaling the previous iteration's weights $\omega_i^{(m-1)}$. Instances k_m that are incorrectly classified have their weights enhanced by e^{α_m}, while correctly classified instances are downweighted by $e^{-\alpha_m}$. As training proceeds, the algorithm increasingly focuses on erroneous examples to improve its predictive performance. To incorporate "Fairness by Explicability" into AdaBoost, we adjust its reweighting process to consider the SHAP values $\{\phi_j^{i,g}\}, i = 1, \ldots, N$, of the features $\{j\}$ of a surrogate g. This SHAP weighting is introduced through a penalty function ($\mathcal{P}(\{\phi_j^{i,g}\})$) and fairness regularization weight (λ) which trades off the original weight update with the new penalty.

In effect, this forces weak learners to not only focus on erroneous examples but also those with specific SHAP values as determined by g and $\mathcal{P}$. This pushes the algorithm to improve its predictions on instances with specific SHAP values and is dubbed "SHAPEnforce". Furthermore, in contrast to SHAPSqueeze, it is fully non-parametric and only requires that the SHAP values of g can be computed.

Algorithm 1 presents the pseudo-code for "SHAPEnforce". The learning approach can be qualitatively interpreted as a two-player game. Expanding on this view, at each stage the predictive learner makes a move by constructing a weak learner and attempts to reweight the training data as-if the surrogate had not acted up to that point. Similarly, once the learner is constructed the surrogate acts to reweight the dataset in its own best interest. The regularization weight λ then controls the resulting outcome between these two competing actions.

In this work, we consider a linear surrogate model trained on data where $Y = 1$ whose form and associated $\mathcal{P}$ is shown in Table 1. We again approximate the SHAP values using Eq. 8. The $\mathcal{P}$ considered is local in nature and, conditioned on $Y = 1$, will downweight any examples with positive SHAP values while upweighting those with negative values. This forces the predictor model to focus on instances where the Z attributions have a negative impact on the favourable outcome and where the weak learner has made mistakes when the target is favourable, i.e. $Y = 1$. By focusing on the examples with negative Z attribution, their Z attribution will be increased at the next round, hence the explicability fairness, as determined by an equivalent l, will tend to increase. This choice of $\mathcal{P}$ further reflects the intuition that unprivileged groups are likely

Algorithm 1 SHAPEnforce

Input: training examples $\{(x_i, y_i, z_i)\}_{i=1}^{N}$, specification of favourable outcome, a surrogate model g, a SHAP penalty function $\mathcal{P}$, and the number of boosting rounds R.

INITIALIZE weights $\omega_i^1 = 1/N, \forall i$.
for m=1,...,R **do**
  Fit a weak learner $k_m(x)$ using the training data with weights ω_i^m.
  Compute the probability of favourable outcome, $\bar{y}_i^m$, and the predicted label $\hat{y}_i^m$ from k_m.
  Fit g - taking features and the target from $\{(x_i, y_i, z_i, \bar{y}_i^m)\}$.
  Compute the $\phi_j^{i,g}$, and the corresponding weight adjustment $\mathcal{P}(\{\phi_j^{i,g}\})$.
  Compute $e_m \leftarrow \mathbb{E}_{\omega^m}[\mathbb{1}_{(y \neq k_m(x))}]$.
  Compute $\alpha_m = \log((1 - e_m)/e_m)$.
  Update the instance weights: $\omega_i^{(m+1)} \leftarrow \omega_i^m[(1-\lambda) * e^{\alpha_m(\mathbb{1}_{(y \neq k_m(x))} - \mathbb{1}_{(y = k_m(x))})} + \lambda e^{\mathcal{P}(\{\phi_j^{i,g}\})}]$.
  Set $\omega_i^{m+1} \leftarrow \frac{\omega_i^{m+1}}{\sum_i \omega_i^{m+1}}$.
end for
Output: Classifier $C_R(x) = \sum_{m=1}^{R} \alpha_m k_m(x)$.

to have unfavourable predictions from weak learners and hence negative Z attribution. Furthermore, with the focus on examples where $Y = 1$ we expect this modification to reduce the EOD.

7 Computational Experiments

To evaluate our algorithms we consider three binary classification datasets: a synthetic dataset, the UCI Adult dataset [4], and a commercial Credit Risk dataset. The train/test splits are shown in Table 2. The datasets were preprocessed so categorical variables were one-hot encoded and numeric variables were converted to their standard score.

We exemplify the SHAPSqueeze objectives using XGBoost [6]. In each experiment, we evaluate the algorithms predictive performance, as measured by accuracy/precision and ROC AUC, as well as measuring the SPD and EOD. To determine these quantities, we use a fixed threshold policy. For SHAPSqueeze, in the case of the synthetic and UCI Adult dataset, this threshold is 0.5 while a more risk-averse threshold of 0.85 is set for the commercial Credit Risk dataset. This higher threshold better reflects real-world business practices in this domain. SHAPEnforce, being a modification to AdaBoost, is less calibrated than the SHAPSqueeze implementation and so a threshold of 0.5 is used in all cases. Additionally, we build linear regression auditor models on the test set to measure the explicability fairness. The equations defining l are the same as the g employed, and so the explicability fairness is given by the coefficient β of the fitted l, see Table 1. Note for SHAPEnforce, l is built on the data subset where $Y = 1$.

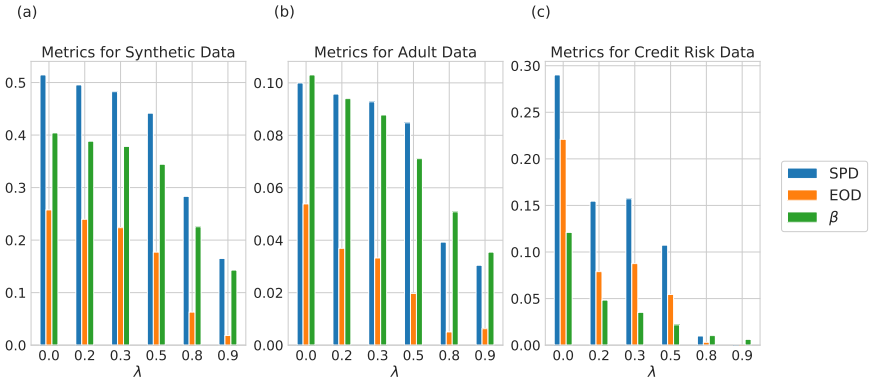

Fig. 1. Fairness metrics for SHAPSqueeze plotted with varying regularization strength for the (a) synthetic, (b) Adult and (c) Credit Risk test datasets. We set $C = 1$ for the synthetic data, $C = 10$ for Adult and $C = 100$ for the Credit Risk evaluations.

Table 2. Datasets used for the algorithm evaluation.

dataset	train size	test size
Synthetic	75000	25000
Adult	32561	16281
Credit Risk	48112	23697

7.1 Datasets

Synthetic Data The synthetic dataset was generated to exhibit a very large SPD. To construct this, the distribution of $\mathbb{X}$ is conditional on Z and Y is determined by Z and $\mathbb{X}$. Specifically, Z was sampled from a Bernoulli distribution and $\mathbb{X}$ contains three sets of covariates: "safe" covariates $\mathbb{X}_s \sim \mathcal{N}(0, 1)$, "proxy" covariates ($\mathbb{X}_p$) and "indirect effect" covariates ($\mathbb{X}_i$). The latter two are sampled from $\mathcal{N}(Z, 1)$. From this, we set the log-odds of the binary target (S_Y) are then given by $0.25\mathbf{w} \cdot (\mathbb{X}_i + \mathbb{X}_s) + 1.25Z$, $\mathbf{w}$ is a vector of ones. The target Y is then sampled from Bern $\left(\frac{1}{1+e^{-S_Y}}\right)$. Using this approach we sampled a dataset with 10 safe, 4 indirect effect and 2 proxy variables. Furthermore, the sampled dataset was such that approximately 90% of the favourable outcomes were obtained by the privileged group.

Adult Census The goal is to predict whether a person will have an income below or above $50k. In this dataset, we consider the variable sex as our protected attribute and removed race, marital status, native country and relationship from our models. The other covariates measure financial information, occupation and education.

Private Credit Risk Dataset In this dataset, we are trying to infer a customer's default probability given curated information on their current account transactions. We are interested in removing bias related to age. We binarize the age variable dividing our examples in two groups, an "older" (unprivileged) group of people over 50 and a "younger" group of people under 50 years old.

7.2 Results

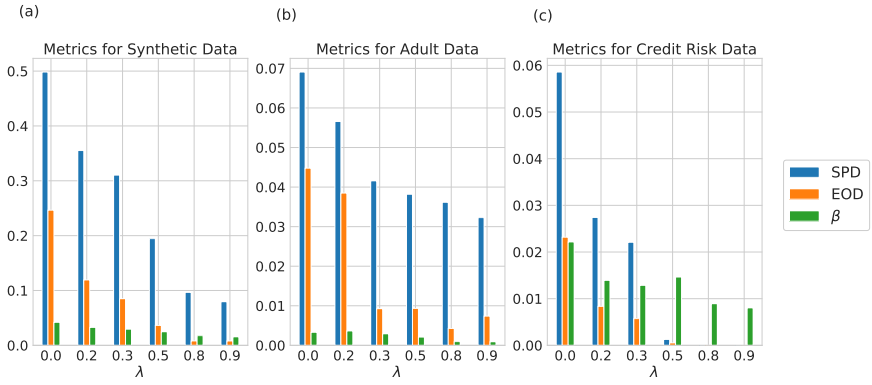

Fig. 2. Fairness metrics for SHAPEnforce, using the penalty $\mathcal{P}$ in Table 1, plotted with varying regularization strength. Results for the synthetic, Adult and Credit Risk test datasets are shown in (a), (b) and (c) respectively.

Results for SHAPSqueeze on the test datasets are shown in Fig. 1. We observe that across all 3 datasets increasing λ induces fairness as observed by reductions in SPD, EOD and β. We set $C = 1$ for the synthetic dataset, $C = 10$ for Adult and for the Credit Risk dataset we set $C = 100$. These values were chosen to ensure the mean gradients from $\mathcal{L}_o$ and $\mathcal{R}$ in the intermediate stages of training, i.e. ~ 100 iterations, when $\lambda = 0.5$ were on the same order of magnitude and effective . For the synthetic data, we observe β drops from 0.404 at $\lambda = 0$ to 0.142 at $\lambda = 0.9$. It is accompanied by a tolerable drop in the AUC and accuracy of 0.04 in both cases. Similarly, the SPD is reduced by roughly 0.35 while the EOD is almost eliminated, taking a value of 0.018 at $\lambda = 0.9$. Increasing λ further, β approaches zero and is faithful to our *strong* "Fairness by Explicability" definition.

We observe the same patterns for the fairness metrics when SHAPSqueeze is applied to the Adult and Credit Risk datasets. In the former case, we observe a reduction of roughly 0.04 in accuracy and AUC with increasing λ, at $\lambda = 0.9$ these take values of 0.80 and 0.83 respectively. In the latter case, the precision is reduced by ≈ 0.12 and the AUC drops by ≈ 0.07 as we change λ from 0 to 0.9. Contrastingly, for Adult, we observe an increase in precision (from 0.76 to

0.98) as the fairness regularization increases the scores beyond the classification threshold. A similar effect is seen in the Credit Risk dataset where we observed an increase in the accuracy from 0.744 to 0.837 as λ was increased to 0.9. This increased accuracy is attributed to the conservative threshold of 0.85 employed. This threshold also results in the SPD and EOD being eliminated at $\lambda = 0.9$ as the regularization pushes all of the scores above 0.85. At this point β is roughly 0.006 demonstrating that even when the SPD and EOD are zero a model may not be 100% explicably fair. This highlights the differences in fairness definition and, in particular, the use of $\bar{Y}$ and not $\hat{Y}$ when measuring explicable fairness. To avoid this scenario one would either reduce C or select a different λ value. At $\lambda = 0.7$, the model has SPD, EOD and β values of 0.035, 0.013 and 0.014 respectively. It is also performant with tolerable drops in the AUC (0.06) and precision (0.1) observed.

The results for SHAPEnforce are presented in Fig. 2. In all cases, we observe the EOD, SPD, AUC and accuracy decrease with increasing λ. For the synthetic data, the accuracy drops by approximately 0.07 from 0.829 to 0.76 as we increase λ. This is accompanied by a drop of ≈ 0.03 in the AUC from 0.868 to 0.839 as we change λ from 0 to 0.9. Compared to the statistical fairness metrics, we observe smaller improvements in the explicable fairness. Furthermore, the decreasing trend of β is less pronounced and consistent compared to SHAPSqueeze. This was expected for two reasons. Firstly, the unregularized AdaBoost model is explicably fairer than XGBoost and so there is less explicable unfairness to remove. Secondly, we expected the heuristic nature of the modification provides no guarantees on explicable fairness and so the magnitude of the reduction is not guaranteed. For the synthetic dataset we observe a decrease in β of $\approx 63\%$ as we increase λ from 0 to 0.9. Moving to the Adult results, we observe β decreases by approximately 72% on changing λ from 0 to 0.9. The SPD and EOD are reduced to 0.03 and 0.007 respectively with tolerable drops in accuracy (0.02) and AUC (0.01) observed. For the Credit Risk data, we again observe explicable fairness improvements, on the order of 64% as we increase λ. This is accompanied with the SPD and EOD being eliminated for $\lambda > 0.7$. Similar to SHAPSqueeze, this elimination is due to the regularization pushing all scores below the threshold for $\lambda > 0.7$. In practice one would use a model from another λ, such as $\lambda = 0.2$, where the SPD and EOD are reduced by roughly 53% and 64% respectively while the precision and AUC take values of 0.85 and 0.84 respectively. This represents a drop of ≈ 0.01 for the former and < 0.01 for the latter. However, at this point, β is only reduced by 37% compared to $\lambda = 0$.

8 Conclusions

In this work, we developed a novel fairness definition, "Fairness by Explicability", that gives the explanations of an auditor's surrogate model primacy when determining model fairness. We demonstrated how to incorporate this definition into model training using adversarial learning with surrogate models and SHAP values. This approach was implemented through appropriate regularization terms

and a bespoke adaptation of AdaBoost. We exemplified these approaches on 3 datasets, using XGBoost in combination with our regularizations, and connected our choices of surrogate model to "statistical parity" and "equality of opportunity" difference. In all cases, the models trained were explicably and statistically fairer, yet still performant. This methodology can be readily extended to other interpretability frameworks, such as LIME [33], with the only constraint being that $\mathcal{R}$ must be appropriately differentiable. Future work will explore more complex surrogate models and different explicability scores in the proposed framework.

9 Related Work

In recent years, there has been significant work done in both model interpretability, adversarial learning and fairness constrained machine learning model training.

Interpretability: Ref. [25] provided a unified framework for interpreting model predictions. This framework unified several existing frameworks, e.g. LIME [33] and DeepLift [36], and it can be argued to be the "gold standard" for model interpretability. It provides both local and global measures of feature attribution and through the KernelSHAP algorithm, is model agnostic. Further work has introduced computationally efficient approximations to the SHAP values of [25] for tree-based models [24]. Other works in interpretability have focussed on causality for model interpretability. These approaches provide insight into *why* the decision was made, rather than an explanation of the model predictive accuracy and are frequently qualitative in nature. Ref. [30] is a recent exception, where the counterfactual examples generated obey realistic constraints to ensure practical use and are examined quantitatively through bespoke metrics.

Adversarial Training: Ref. [3] used adversarial training to remove EOD while a framework for learning adversarially fair representations was developed in Ref. [26]. Similar, in Ref. [46] an adversarial network [16] was used to debias a predictor network, their specific approach compared favourably to the approach of [3].

Training Fair Models: typically, fair learning methodologies have tended to focus on incorporating statistical fairness constraints directly into the model objective function. Ref. [27] combined neural networks with statistical fairness regularizations but their form restricts their applicability to neural networks. Similarly, Ref. [15] trained a fair logistic regression using convex regularizations. These regularizations rely on empirical weights that represent the historical bias and were designed with proportionally fair classification rather than classical fairness measures in mind. Other works have viewed fair model training as one of constrained optimization [45,43,44] or have created meta-algorithms for fair classification [5].

In these works, the approaches to fair learning have tended to focus on fairness metrics associated with more traditional worldviews and less focus on model explicability. Similarly, the role of model explicability in fairness, to the authors' knowledge, has not been used directly in fair model training but instead research

has focussed on the consistency and transparency of explanations. Our work is novel as it places the role of model explicability at the core of a new fairness definition and develops an adversarial learning methodology that is applicable to adaptive boosting and any model trained via gradient-based optimization. In the former case, our proposed algorithm is fully non-parametric where the adversary can come from any model family provided the corresponding explicability scores, in this case SHAP values, can be computed.

10 Acknowledgements

We thank C. Dhanjal, F. Bellosi, G. Jones and L. Stoddart for their useful suggestions and discussions. We also thank Experian Ltd and J. Campos Zabala for supporting this work.

References

1. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 60–69. PMLR (10–15 Jul 2018), http://proceedings.mlr.press/v80/agarwal18a.html
2. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016)
3. Beutel, A., Chen, J., Zhao, Z., Chi, E.H.: Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:1707.00075 (2017)
4. University of California, I.: Census income dataset (1996), https://archive.ics.uci.edu/ml/datasets/census+income
5. Celis, L.E., Huang, L., Keswani, V., Vishnoi, N.K.: Classification with fairness constraints: A meta-algorithm with provable guarantees. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 319–328. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287586, https://doi.org/10.1145/3287560.3287586
6. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. KDD 2016, Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939785, https://doi.org/10.1145/2939672.2939785
7. Chiappa, S., Gillam, T.: Path-specific counterfactual fairness. Proceedings of the AAAI Conference on Artificial Intelligence **33** (02 2018). https://doi.org/10.1609/aaai.v33i01.33017801
8. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017). https://doi.org/10.1089/big.2016.0047, https://doi.org/10.1089/big.2016.0047, pMID: 28632438

9. Corbett-Davies, S., Goel, S., Morgenstern, J., Cummings, R.: Defining and designing fair algorithms. In: Proceedings of the 2018 ACM Conference on Economics and Computation. p. 705. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3219166.3277556, https://doi.org/10.1145/3219166.3277556
10. Di Stefano, P., Hickey, J., Vasileiou, V.: Counterfactual fairness: removing direct effects through regularization. arXiv preprint arXiv:2002.10774 (2020)
11. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226. ITCS 2012, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2090236.2090255, https://doi.org/10.1145/2090236.2090255
12. European Parliament and Council of the European Union: Regulation (eu) 2016/679 regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (data protection directive). OJ **L 119**, 1–88 (2016), https://gdpr-info.eu/
13. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences **55**(1), 119–139 (1997). https://doi.org/https://doi.org/10.1006/jcss.1997.1504, http://www.sciencedirect.com/science/article/pii/S002200009791504X
14. Friedler, S., Choudhary, S., Scheidegger, C., Hamilton, E., Venkatasubramanian, S., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. pp. 329–338. FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, Inc (1 2019). https://doi.org/10.1145/3287560.3287589
15. Goel, N., Yaghini, M., Faltings, B.: Non-discriminatory machine learning through convex fairness criteria (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16476
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
17. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine **38**(3), 50–57 (Oct 2017). https://doi.org/10.1609/aimag.v38i3.2741, https://www.aaai.org/ojs/index.php/aimagazine/article/view/2741
18. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) Machine Learning and Knowledge Discovery in Databases. pp. 35–50. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
19. Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B.: Avoiding discrimination through causal reasoning (06 2017)
20. Kleinberg, J.: Inherent trade-offs in algorithmic fairness. In: Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems. p. 40. SIGMETRICS-18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3219617.3219634, https://doi.org/10.1145/3219617.3219634

21. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4066–4076. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf
22. Lipovetsky, S., Conklin, M.: Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry **17**(4), 319–330 (2001). https://doi.org/10.1002/asmb.446, https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.446
23. Lum, K., Isaac, W.: To predict and serve? Significance **13**(5), 14–19 (2016). https://doi.org/10.1111/j.1740-9713.2016.00960.x
24. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: Explainable ai for trees: From local explanations to global understanding. arXiv preprint arXiv:1905.04610 (2019)
25. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
26. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 3384–3393. PMLR, Stockholmsmässan, Stockholm Sweden (10–15 Jul 2018), http://proceedings.mlr.press/v80/madras18a.html
27. Manisha, P., Gujar, S.: A neural network framework for fair classifier. arXiv preprint arXiv:1811.00247 (2018)
28. Mansoor, S.: A viral tweet accused apple's new credit card of being 'sexist.' now new york state regulators are investigating. TIME Magazine (2019)
29. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)
30. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 607–617. FAT* 2020, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3351095.3372850, https://doi.org/10.1145/3351095.3372850
31. Nabi, R., Shpitser, I.: Fair inference on outcomes. Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence **2018**, 1931–1940 (Feb 2018), https://www.ncbi.nlm.nih.gov/pubmed/29796336, 29796336[pmid]
32. OŃeil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, USA (2016)
33. Ribeiro, M.T., Singh, S., Guestrin, C.: ẅhy 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. pp. 1135–1144. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939778, https://doi.org/10.1145/2939672.2939778
34. Roemer, J.E., Trannoy, A.: Equality of opportunity. In: Handbook of income distribution, vol. 2, pp. 217–300. Elsevier (2015)
35. Shapley, L.S.: A value for n-person games. pp. 307–317. Contributions to the Theory of Games 2.28 (1953)

36. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. pp. 3145–3153. JMLR.org (2017)
37. Strumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems pp. 647–655 (2014), https://doi.org/10.1007/s10115-013-0679-x
38. Suresh, H., Guttag, J.V.: A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002 (2019)
39. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. pp. 1–7. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3194770.3194776, https://doi.org/10.1145/3194770.3194776
40. Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law and Technology **31** (2018)
41. Yeom, S., Tschantz, M.C.: Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews. arXiv preprint arXiv:1808.08619v4 (2019)
42. Young, H.P.: Monotonic solutions of cooperative games. International Journal of Game Theory **14**, 65–72 (1985), https://doi.org/10.1007/BF01769885
43. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1171–1180. WWW 2017, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017). https://doi.org/10.1145/3038912.3052660, https://doi.org/10.1145/3038912.3052660
44. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P., Weller, A.: From parity to preference-based notions of fairness in classification. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 228–238. NIPS-17, Curran Associates Inc., Red Hook, NY, USA (2017)
45. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Singh, A., Zhu, J. (eds.) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 962–970. PMLR, Fort Lauderdale, FL, USA (20–22 Apr 2017), http://proceedings.mlr.press/v54/zafar17a.html
46. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. pp. 335–340. AIES 18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3278721.3278779, https://doi.org/10.1145/3278721.3278779
47. Zhang, L., Wu, Y., Wu, X.: A causal framework for discovering and removing direct and indirect discrimination. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17. pp. 3929–3935 (2017). https://doi.org/10.24963/ijcai.2017/549, https://doi.org/10.24963/ijcai.2017/549
48. Zhao, Q., Hastie, T.: Causal interpretations of black-box models. Journal of Business & Economic Statistics **0**(0), 1–10 (2019). https://doi.org/10.1080/07350015.2019.1624293, https://doi.org/10.1080/07350015.2019.1624293