Article:
Pelegrina, G.D., Siraj, S. orcid.org/0000-0002-7962-9930, Duarte, L.T. et al. (1 more author) (2024) Explaining contributions of features towards unfairness in classifiers: A novel threshold-dependent Shapley value-based approach. Engineering Applications of Artificial Intelligence, 138 (Part B). 109427. ISSN 0952-1976

https://doi.org/10.1016/j.engappai.2024.109427

Explaining contributions of features towards unfairness in classifiers: A novel threshold-dependent Shapley value-based approach

Guilherme Dean Pelegrina, Sajid Siraj, Leonardo Tomazeli Duarte, Michel Grabisch

[a] *Mackenzie Presbyterian University (UPM), 930 Consolação St, São Paulo, 01302-907, São Paulo, Brazil*
[b] *Centre for Decision Research, Leeds University Business School, Leeds, LS6 1AN, UK*
[c] *School of Applied Sciences - University of Campinas (UNICAMP), 1300 Pedro Zaccaria St, Limeira, 13484-350, São Paulo, Brazil*
[d] *Centre d'Économie de la Sorbonne - Université Paris I Panthéon-Sorbonne, 106-112 bd de l'Hôpital, Paris, 75013, France*

Abstract

A number of approaches has been proposed to investigate and mitigate unfairness in machine learning algorithms. However, as the definition and understanding of fairness may vary in different situations, the study of ethical disparities remains an open area of research. Besides the importance of analyzing ethical disparities, explainability in machine learning is also a relevant issue in Trustworthy Artificial Intelligence. Usually, both fairness and explainability analysis are based on a fixed decision threshold, which differentiates the positive cases from the negative ones according to the predicted probabilities. In this paper, we investigate how changes in this threshold can impact the fairness of predictions between protected and other groups and how features contribute towards such a measure. We propose a novel Shapley value-based approach as a tool to investigate how changes in the threshold values change the contribution of each feature towards unfairness. This gives

Preprint submitted to Engineering Applications of Artificial Intelligence October 3, 2024

us an ability to evaluate how fairness measures vary for different threshold values and which features have the higher (or lower) impact on creating ethical disparities. We demonstrate this using three different case studies that are carefully chosen to highlight different unfairness scenarios and features contributions. We also applied our proposal as a feature selection strategy, which contributed to decrease unfair results substantially.

Keywords:

Interpretable Machine Learning, Shapley value, Fairness, Feature contribution

1. Introduction

It is often argued that decision making by people contain several psychological biases [1, 2] and that may be avoided if we rely on machine-based intelligence. Based on the level of human involvement, the authors in [3] explained and structured operations into different types. On one hand, they define the fully automated tasks as "out of the loop" meaning that humans are not involved, while on the other hand, the "on the loop" tasks were defined as those where decisions are solely depending on humans. Although it appears that excluding humans from decision process should remove the issues and biases related to humans, it is not that simple as machines are often trained with the data that contain biases in itself [4]. The issue of unfairness in Machine Learning (ML) algorithms has been widely debated in recent years and a number of approaches has been proposed to mitigate unfairness [5, 6]. One may cite, for instance, works that exploit the concepts of equal opportunity [7] and accuracy parity [8]. However, as the concept

and definition of fairness may vary in different context, this remains an open area of research. Regardless of its definition, fairness is closely related to the explainability of algorithms. This is because to investigate fairness, there should be a way to explain the results and performance of machine learning algorithms.

One can define the term explainability as the ability of an algorithm to provide information that can help evaluate causalities, similarities, and/or uncertainties which in turn helps decision makers towards understanding the model outcomes [9, 10]. Miller [11] described this field as the intersection of ML with the subject areas of human-computer interaction and social sciences. Some of the explainability tools focus on improving the interpretability of data, for example, by reducing the dimension of feature space [12]. Others focus on the machine learning models, for example, by explaining the outcomes generated by ML algorithms [13]. Among such approaches, one may cite the well-known method proposed by [14], called SHAP, which has been addressed in several recent works [15, 16, 17].

The term explainable AI, or XAI in short, can be used at two very different levels: the global and the local ones [13, 10]. The global level interpretations are important to assess and audit the explainability of algorithms overall, for example, which feature contributes more towards explaining the outcomes [18]. The local level interpretations, however, are more suited to explain results (or predictions) provided for local samples [14], for example, why a loan has been refused to a certain applicant. While both levels of explainability can be useful in different situations, a single instance of local explainability might not be sufficient for evaluating fairness. The notion of

group fairness demands investigation of multiple cases in order to compare the algorithm's performance for different groups (for example, male group versus female group). Although some of the explainability strategies are only applicable to specific algorithms [19], our focus remains on model-agnostic algorithms that can be applied to different machine learning techniques equally. In this context, the use of Shapley value-based approaches has gained more attention due to its algorithm-agnostic characteristic [14], versatility [20, 21] and other useful properties [22, 23].

The issue of fairness in machine learning has also gained attention in recent years [24, 25]. For example, this has been proposed as a constrained optimization problem where the objective is to minimize the mis-classification probability whilst imposing an upper bound on an unfairness measure [26]. Fair solutions have also been exploited by means of multi-objective approaches [27, 28]. In this case, both algorithm performance and fairness concerns are optimized simultaneously. Although mitigation of unfairness in ML algorithms has been investigated, there is a little done in the field of explaining the contribution of features towards unfair results generated by these algorithms. Usually explainability for ML classifiers is done considering a single (pre-assigned/fixed) value of threshold that is used to differentiate the positive cases from the negative ones. Changing this threshold might not only impact the performance of classifiers but can also affect fairness for a protected group against the other groups. [15] proposed the use of Shapley values to investigate fairness in machine learning models and to explain the trade-off between accuracy and fairness. However, they did not investigate thresholds used for classification and their impact on fairness, whether using equalized

odds [7] or any other method of quantifying fairness.

A recent work [29] involved generating various different decision thresholds to estimate the contributions of features towards the quality of predictions, measured in terms of Receiver-operating Characteristics (ROC) curve and the Area under the ROC curve (AUC). Although the explanation of these curves can help analysts in feature engineering, it may also help investigate the issues of fairness, for example, by explaining the contribution of features towards disparities in quality of predictions between the protected and the privileged groups. Therefore, we see a gap in the literature in providing a mechanism to evaluate features contributions toward unfair results along with thresholds[1]. Instead of an analysis for a single (and predefined) threshold, it is of interest to verify how fairness and features contributions varies for different threshold values. Aiming at overcoming this gap, in this paper, we investigate the range of decision thresholds to identify different levels of fairness along with the performance of classifiers. Our proposal is based on the Shapley value, which indicates features contributions towards both performance (for protected and privileged groups) and fairness measures. From the experimental results, we attest that our proposal can be useful to (i) evaluate which features impact disparate results, (ii) investigate the presence of features acting as proxies and (iii) observe how features contribute differently towards different sensitive groups. Moreover, we also apply our proposal as

[1]Indeed, most of the explainable AI techniques, such as LIME [30] or SHAP [14], deal with local interpretability, i.e., with the purpose of explaining the outcome of a sample of interest. Although SHAP can be extended to global interpretability, it provides the contributions toward the predicted classes, instead of group fairness measures along with thresholds.

a tool to assist feature selection. Indeed, by removing features with relevant contributions toward unfairness, we are able to mitigate ethical disparities.

The next section describes the background with related literature, followed by the proposed technique (Section 3). The numerical experiments are presented in Section 4, and then, Section 5 concludes the discussion.

2. Background

This section discusses the theoretical background used in our proposal. Firstly, we present the adopted notations and, thereafter, we discuss some performance and fairness metrics frequently used to evaluate machine learning models. Finally, we define the Shapley value and how it has been used as a feature attribution method for ML explainability.

2.1. Setup and Notations

Assume a binary classification setting where $\mathbf{X} \in \mathbb{R}^{n \times m}$ represents the m-dimensional dataset with n samples and $\mathbf{y} \in \{0,1\}^n$ is the associated vector of labels. Generally, class 1 (the "positive" class) indicates a benefit. Consider $\mathbf{d} \in [0,1]^n$ as the vector of predicted probabilities which indicate the likelihood of belonging to class 1. Given an instance $x^{(i)}$ and a predefined threshold t, $x^{(i)}$ is classified as class 1 if $d^{(i)} \geq t$. The vector of all predicted labels is represented by $\hat{\mathbf{y}}$, i.e., for a given instance $x^{(i)}$, if $d^{(i)} \geq t$, $\hat{y}^{(i)} = 1$.

Suppose a classification task whose samples can be split into sensitive groups defined by $G \in \{a, b\}$ (e.g., blacks and whites or men and women). Without loss of generality, consider that a and b are the protected and the privileged groups, respectively. By splitting $\mathbf{X}$ into these groups, one may define $\mathbf{X} = [\mathbf{X}_a; \mathbf{X}_b]$, where $\mathbf{X}_a \in \mathbb{R}^{n_a \times m}$ and $\mathbf{X}_b \in \mathbb{R}^{n_b \times m}$ are m-dimensional

datasets with n_a and n_b samples, respectively. The vector of labels, predicted probabilities and predicted labels can also be split similarly.

2.2. Performance metrics

In order to evaluate the performance of a classifier, one generally considers metrics used to construct the confusion matrix. They are defined as follows:

- True positive (TP): Number of instances correctly classified as class 1.
- True negative (TN): Number of instances correctly classified as class 0.
- False positive (FP): Number of instances wrongly classified as class 1.
- False positive (FN): Number of instances wrongly classified as class 0.

Based on these metrics, the classifier's performance is evaluated by means of the following rates:

- Positive predictive value ($PPV = \frac{TP}{TP+FP}$): ratio between the number of class 1 correctly classified as class 1 and the total number of instances classified as class 1. This ratio is also called Precision.
- Negative predictive value ($NPV = \frac{TN}{TN+FN}$): ratio between the number of class 0 correctly classified as class 0 and the total number of instances classified as class 0.
- True positive rate ($TPR = \frac{TP}{TP+FN}$): percentage of class 1 correctly classified as class 1. It is also called Sensitivity or Recall.

- False positive rate ($FPR = \frac{FP}{FP+TN}$): percentage of class 0 wrongly classified as class 1.

All the aforementioned metrics can also be defined by conditioning on the sensitive groups. For example, TP_a means the number of instances in group a correctly classified as class 1 and FPR_b is the percentage of class 0 wrongly classified as class 1 for the individuals belonging to group b.

2.3. Measures of fairness

Besides the performance metrics presented in the previous subsection, machine learning models have been evaluated in terms of disparate results against protected groups. We present in the sequel a brief description of some fairness metrics used in our analysis (see [31, 24] for further details and other metrics).

2.3.1. Statistical parity (or Demographic parity)

A classifier satisfies statistical parity (SP) [32] if both groups have the same probability of being classified as the positive class, i.e., $\frac{TP_G+FP_G}{n_G}$ must be the same regardless group G. Therefore, the SP strategy is associated with the minimization of the following cost function:

$$f_{SP} = \left| \frac{TP_a + FP_a}{n_a} - \frac{TP_b + FP_b}{n_b} \right|. \tag{1}$$

The idea supported by this definition is that equivalent positive outcomes (such as receiving a credit) should be similar for individuals, regardless if its group.

2.3.2. Predictive equality

A classifier satisfies predictive equality (PE) if both groups have equal FPR. In other words, $P(\hat{Y} = 1 \| Y = 0, G = 1) = P(\hat{Y} = 1 \| Y = 0, G = 0)$. One here attempts to minimize

$$f_{PE} = \left| \frac{FP_1}{FP_1 + TN_1} - \frac{FP_0}{FP_0 + TN_0} \right|. \tag{2}$$

In this metric, the aim is to assign the same amount of positive outcomes to individuals of different groups that actually belong to the negative class.

2.3.3. Equalized odds

A classifier satisfies equalized odds (EO) if both groups have equal TPR and FPR. In other words, $P(\hat{Y} = 1 \| Y = 1, G = 1) = P(\hat{Y} = 1 \| Y = 1, G = 0)$ and $P(\hat{Y} = 1 \| Y = 0, G = 1) = P(\hat{Y} = 1 \| Y = 0, G = 0)$. The definition of equalized odds was proposed by [7] to remedy previously noted flaws with demographic parity [32]. The goal is to minimize

$$f_{EO} = \left| \frac{TP_1}{TP_1 + FN_1} - \frac{TP_0}{TP_0 + FN_0} \right| + \left| \frac{FP_1}{FP_1 + TN_1} - \frac{FP_0}{FP_0 + TN_0} \right|. \tag{3}$$

2.4. Shapley values in machine learning interpretability

The Shapley value [33] is a classical solution concept in game theory. Consider a scenario in which a set $M = \{1, 2, \ldots, m\}$ of m players join a coalition in order to achieve a common goal. For example, energy storage systems owners could join coalitions in order to save individual costs [34]. In such a scenario, the Shapley value ϕ_j associated with each player j will indicate how much he/she should receive when sharing the whole benefit

achieved by the coalition of all players. Mathematically, it is defined as follows:

$$\phi_j = \sum_{A \subseteq M \setminus \{j\}} \frac{(m - |A| - 1)!\,|A|!}{m!} \left[\upsilon(A \cup \{j\}) - \upsilon(A)\right], \tag{4}$$

where $|A|$ indicates the cardinality of subset A and $\upsilon(A)$ is the game payoff (or benefit) when only players in A join the coalition. For the coalition of all players, $\upsilon(M)$ indicates the total benefit.

Among the several properties satisfied by the Shapley value (see [35] for further details), one is of interest in machine learning interpretability: efficiency. This property states that the payoff of the grand coalition $\upsilon(M)$ can be decomposed into the sum of the individuals Shapley values. Mathematically, efficiency means $\sum_{j=1}^{m} \phi_j = \upsilon(M) - \upsilon(\emptyset) = \upsilon(M)$ (in game theory, one generally assumes $\upsilon(\emptyset) = 0$). In machine learning, this property allows us to explain the contributions of features from a predefined baseline ($\upsilon(\emptyset)$) to an achieved goal ($\upsilon(M)$). Therefore, how to defined both baseline and goal is an important task when adopting the Shapley value for ML interpretability (see [29] for further details).

One associates both baseline and goal in machine learning to the game payoff $\upsilon(\cdot)$ in game theory. Clearly, the definition of both elements depends on what one would like to explain. For instance, [36] used the Shapley value to evaluate the contribution of features towards the coefficient of determination in linear regression models. In this case, the payoff $\upsilon(A)$ is the coefficient of determination when features in A are available in the linear regression model. When no feature is available, $\upsilon(\emptyset) = 0$, and when all feature is available,

the coefficient of determination $\upsilon(M)$ is maximal. In the well-know SHAP method [14], the authors adopted the Shapley value as a model-agnostic feature attribution approach to explain local predictions. In this case, given a sample of interest x^*, one may interpret how much the associated characteristics (features values $x_1^*, x_2^*, \ldots, x_m^*$) contributes towards the obtained prediction or classification. Then, $\upsilon(\emptyset)$ represents the expected prediction when all features values in x^* are unknown and $\upsilon(M)$ is the actual prediction assigned to x^*.

Most of the work in explainable machine learning uses the Shapley value to interpret features contributions toward performance metrics such as accuracy. However, only few works attempt to explain such contributions towards unfair results. An example is the work conducted by Begley et al. [15], where the authors aggregate local Shapley values in order to globally explain both performance and fairness measures. Moreover, most of the explainable approaches only focus on results achieved by assuming a single threshold (and, therefore, performance or fairness metrics in a single scenario). In this paper, we borrowed the idea proposed by Pelegrina and Siraj [29] to explain the overall performance of classifiers based on ROC curve and on the area under the ROC curve. As for such analysis where the authors investigated the classifier performance along with thresholds, our proposal in this work is to evaluate the impact of each feature in biased results for different predefined thresholds. We detail our proposal in the next section.

3. Explaining unfair results through Shapley values

In this section, we discuss the use of the Shapley values in our proposal to assign contributions of features towards both performance and fairness measures.

3.1. Features contributions towards performance measures

Generally, fairness implies equal performance measures for the considered groups of individuals. For instance, to ensure the fairness of a model in terms of predictive equality, it is imperative that both groups of individuals exhibit equal false positive rates. Nevertheless, disparate model performances among diverse groups of individuals are a common occurrence in numerous applications. In this context, a crucial aspect involves interpreting how each feature contributes to the observed unfair result. Such an interpretation is vital for the redesign of the machine learning model, enabling the mitigation of inherent inequalities.

In Section 1, we highlighted some existing approaches that are used to evaluate features contributions towards performance measures. Typically, these approaches calculate the impact of features on model performance using a single predefined threshold. However, in this paper, as delineated in [29], we extend our analysis to incorporate contributions across varying thresholds. Assume $p_{t,G}(A)$ as the model performance for group G and predefined threshold t when a set of features, expressed by the set A, is considered in the training step. We define the payoff as follows:

$$\upsilon_{t,G}^{PERF}(A) = p_{t,G}(A) - p_{t,G}(\emptyset), \tag{5}$$

where $p_{t,G}(\emptyset)$ represents the performance of a random classifier (i.e., when no features is available in training step). Note that, in accordance with the definition of a game, $\upsilon_{t,G}^{PERF}(\emptyset) = p_{t,G}^{PERF}(\emptyset) - p_{t,G}^{PERF}(\emptyset) = 0$. Based on $\upsilon_{t,G}^{PERF}(A)$ for all $A \in \mathcal{P}(M)$, where $\mathcal{P}(M)$ is the power set of M, it is possible to calculate the Shapley value of feature j as

$$\phi_j^{PERF,t,G} = \sum_{A \subseteq M \setminus \{j\}} \frac{(m - |A| - 1)!\,|A|!}{m!} \left[\upsilon_{t,G}^{PERF}(A \cup \{j\}) - \upsilon_{t,G}^{PERF}(A)\right]. \tag{6}$$

Due to the efficiency property, $\sum_{j=1}^{m} \phi_j^{PERF,t,G} = \upsilon_{t,G}^{PERF}(M) - \upsilon_{t,G}^{PERF}(\emptyset) = \upsilon_{t,G}^{PERF}(M)$, i.e., the model performance for group G when all features are available ($\upsilon_t(M)$) can be decomposed by the sum of the marginal contributions of each feature. Note that, depending on the adopted performance measure, $\phi_j^{PERF,t,G}$ may be either positive or negative. For instance, if we are evaluating overall accuracy and by assuming that the inclusion of a feature into the model training would not decrease its performance, we would expect $0 \leq \phi_j^{PERF,t,G} \leq \upsilon_t(M)$. In this scenario, we have clear bounds, as the game is non-decreasing and normalized by the accuracy based on all features (see [37] for more details about such a game, usually called capacity). However, if we are looking at the area under the Precision-Recall curve (see [29] for further details), $\phi_j^{PERF,t,G}$ could be negative and, therefore, it is difficult to define bound conditions for such a value.

In a biased scenario, the performances $\upsilon_{t,G}^{PERF}(M)$ as well as the features contributions $\phi_j^{PERF,t,G}$ for $G = \{a, b\}$ may be different and, therefore, we may interpret which features are creating disparate results. Figure 1 illustrates the process of interpreting the disparate results by comparing the

performance measures of different groups of individuals. In summary, the process involves calculating performance measures and features contributions for all coalitions of features $A \in \mathcal{P}(M)$, thresholds $t_0, t_1, \ldots, t_s$ and groups a and b. Let us recall that groups a and b as the protected and the privileged groups, respectively. As in this process one calculates $p_{t,a}^{PERF}(M)$ and $p_{t,b}^{PERF}(M)$, one can compare the performance measure for both groups and visualize disparities along with thresholds. One illustrates this result on the center right plot of Figure 1, which assumes the true positive rate as the performance measure[2]. Moreover, in the bottom right plot in Figure 1, one also illustrates a comparison between features contributions towards the TPR for groups a and b along with thresholds. Note that, for a fixed threshold t, the sum of the Shapley values is equal to the difference between the actual TPR (by using all features) and the random classifier performance which, in this case, is given by $1 - t$. It is worth mentioning that the shaded area in both figures indicate the standard deviation from the averaged value by taking into account the k-fold cross-validation strategy.

So far, we discussed how to use the Shapley values in order to compare performance measures for different groups of individuals and interpret disparities in features contributions. However, unfair results in machine learning are frequently evaluated by means of fairness measures. Therefore, one may also interpret features contributions directly on fairness measures. We elaborate this in the sequel.

[2]It is worth highlighting that the overall TPRs are close to the TPRs from the privileged group due to an imbalance in the dataset used to create this illustrating example. Indeed, there are more instances from the privileged group in comparison with the protected group.

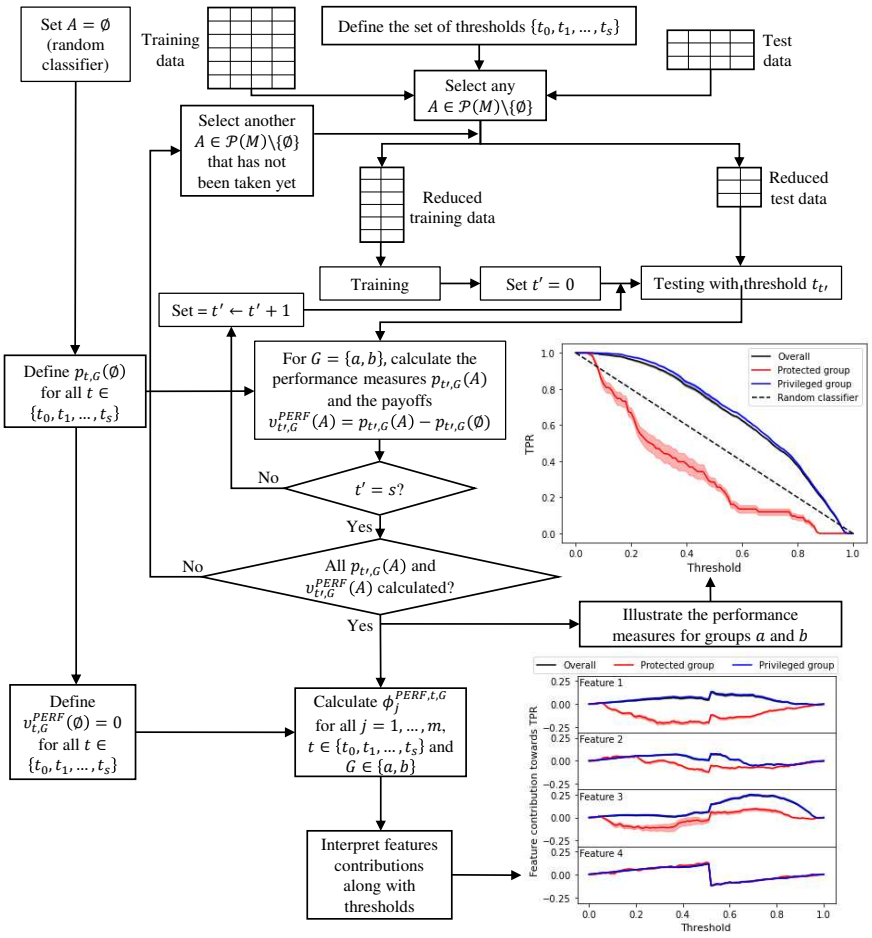


Figure 1: The proposed scheme to evaluate features contributions towards disparities in performance measures.

3.2. Features contributions towards fairness measures

Besides evaluating features contributions towards the model performance, one may also interpret their impacts on fairness measures. Assume $f_{t,p}(A)$ as the fairness measure associated with performance measures $p^{PERF}_{t,a}(A)$ and $p^{PERF}_{t,b}(A)$, for a predefined threshold t when only features in A are available

in training step. The payoff is defined by

$$\upsilon_t^{FAIR}(A) = f_{t,p}(A) - f_{t,p}(\emptyset), \tag{7}$$

where $f_{t,p}(\emptyset)$ represents the fairness measure of a random classifier[3]. One may also note that $\upsilon_t^{FAIR}(\emptyset) = f_{t,p}^{FAIR}(\emptyset) - f_{t,p}^{FAIR}(\emptyset) = 0$, i.e., in accordance with the definition of a game. Based on $\upsilon_t^{FAIR}(A)$ for all $A \in \mathcal{P}(M)$, we may calculate the Shapley value of feature j as

$$\phi_j^{FAIR,t} = \sum_{A \subseteq M \setminus \{j\}} \frac{(m - |A| - 1)!\,|A|!}{m!} \left[\upsilon_t^{FAIR}(A \cup \{j\}) - \upsilon_t^{FAIR}(A)\right]. \tag{8}$$

The interpretation in the considered fairness measure is given by the sum of the marginal contribution of each feature. Indeed, given the efficient property of Shapley values, $\sum_{j=1}^{m} \phi_j^{FAIR,t} = \upsilon_t^{FAIR}(M) - \upsilon_t^{FAIR}(\emptyset) = \upsilon^{FAIR}(M) = f_{t,p}(M) - f_{t,p}(\emptyset)$. In the fairest scenario, we normally expects $f_{t,p}(M) = f_{t,p}(\emptyset)$ and therefore, $\upsilon^{FAIR}(M) = 0$. However, if $\upsilon^{FAIR}(M) > 0$, there is a disparity between groups a and b. The marginal contributions of features given by $\phi_j^{FAIR,t}$ will then highlight which features are creating disparate results and can be seen as a source of bias. As in the case of the performance measure, $\phi_j^{FAIR,t}$ can be either positive (feature j entail disparate results) or negative (feature j contributes to improve fairness).

We illustrate in Figure 2 the steps to interpret the contributions of fea-

[3]Although several fairness measure should be zero when considering a random classifier, we decided to keep $f_{t,p}(\emptyset)$ in our proposal in order to generalize the idea for any fairness measure.

tures on fairness measures. As in the Subsection 3.1[4], the process here also involves calculations on all coalitions of features $A \in \mathcal{P}(M)$ and thresholds $t_0, t_1, \ldots, t_s$. As a first interpretation, one may visualize the fairness measure along with the thresholds. This result is presented on the center right plot of Figure 2, which assumes the statistical parity as the fairness measure. The fairness measure along with the thresholds can be decomposed on individual contributions of features. The bottom right plots in Figure 2 illustrates the marginal contribution of each feature towards the fairness measure. Note that the sum of these marginal contributions is equal to the curve presented in the center right plot of Figure 2. Moreover, if there is an interest in analyzing the Shapley values for a single threshold t, one may use a waterfall plot as illustrated in the bottom left plot of Figure 2. As in Figure 1, the shaded area in both center and bottom right plots and the whiskers in the waterfall plot indicate the standard deviation from the averaged value by taking into account the k-fold cross-validation strategy.

4. Experiments

This section outlines the experiments conducted to validate the efficacy of our proposed method in interpreting feature contributions toward disparate outcomes in machine learning. While our experiments utilized the Random Forest classifier [38] with 5-fold cross-validation, it is essential to note that our approach is model-agnostic, as discussed in Section 3. Consequently, other

[4]Clearly, in a computational point of view and in order to avoid double calculating the performance measures, the schemes presented in Figures 1 and 2 could be merged into a single one. However, aiming at providing an easier visualization of both proposals, prefer to split both of them into two processes.

classifier can be employed. The explanations derived from our approach reveal the contributions of features towards the unfair outcomes produced by the trained classifier.

In order to evaluate our proposal, we examined three real datasets: COMPAS [39], Law School Admission Council (LSAC) [40] and Adult income [41]. For each dataset, we computed disparities related to a sensitive feature and interpreted the individual contributions of each feature towards such disparities. For this purpose, we only consider the sensitive feature to split the dataset into two groups when calculating the performance/fairness measures (i.e., we do not use such a feature in training step). The remaining features will, then, explain the achieved disparity for each threshold. We also conducted a preprocessing step in all datasets in order to ensure that the two classes are balanced. In this case, we used a re-sampling strategy that randomly eliminates samples from the over-represented class until balancing the dataset.

Subsequently, we delve into the results obtained for each dataset. All codes and datasets are openly accessible at the following URL: https://github.com/shaprob/FairShap.

4.1. COMPAS dataset

As a first experiment, we considered the COMPAS dataset [39], released by ProPublica[5] in 2016. In this dataset, one assigns recidivism risk scores to defendants based on a set of numerical and categorical features describing them. We considered seven input features, namely *sex* (male or female),

[5] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

age_cat (age intervals - less than 25, greater than 45 and between 25 and 45 years old), *juv_fel_count* (number of juvenile felonies), *juv_misd_count* (number of juvenile misdemeanors), *juv_other_count* (number of other prior juvenile convictions), *priors_count* (number of prior crimes) and *c_charge_degree* (crime degree - misdemeanor or felony). As the COMPAS dataset also provides the *race* of each defendant, we used this information as the sensible feature. For the purpose of evaluating unfair results between blacks and whites, we only considered African-American (protected group) or Caucasian (privileged group) defendants. After under-sampling the data by removing defendants from other races, we achieved 5048 samples.

It is known from the literature [39] that in this dataset one has disparities in false positive rates when comparing blacks and whites defendants. In other terms, the rate that blacks are wrongly classified as recidivists is higher than this rate for Caucasians. In order to investigate this unfair result, we provide in Figure 3 some interpretations with respect to the false positive rates and Predictive Equality, along with thresholds. One may clearly see from Figure 3a that, regardless the adopted threshold, blacks achieved higher FPRs in comparison with whites. An interesting result was achieved in terms of the feature contributions. As can be seen in Figure 3c, *age_cat* led to the highest contributions toward the unfair results. This finding can be explained by the data distribution with respect to the age categories, race and classes. For instance, among the defendants under 25 years old, 66% are considered recidivists. Moreover, within this age category, 70% are blacks. On the other hand, for those greater than 45 years old, only 24% are considered recidivists and 42% are blacks. Although race was removed from the training

step, the data distribution of age category and recidivism risk scores carries information from race and, therefore, the age category can be seen as a source of bias towards race.

The unfair scenario can also be interpreted directly from the fairness measure. As showed in Figures 3e and 3d, age category is the feature highly associated with disparate outcomes. It is interesting to remark here that there are a lot of dispersion around the averaged Shapley values for *priors_count* and *c_charge_degree* (see the shaded are in Figures 3b and 3d). Therefore, there are uncertainties in the Shapley values estimation for these two feature. However, there are less uncertainties for age category, with positive contributions regardless the adopted threshold. Moreover, note that the disparity decreases as the threshold value increases. For instance, for $t = 0.8$, one practically achieved fairness for predictive equality. However, for this threshold we possibly pay the price of decreasing the TPR. Another choice could be $t \approx 0.35$, which significantly reduces the disparity on FPRs (around 50%) in comparison with the default choice of $t = 0.5$.

4.2. LSAC dataset

In this second application of our proposal, we deal with interpretability in unfair results associated with the LSAC (Law School Admission Council) dataset [40] (see [42] for further details). The goal is to predict whether a student will pass the bar exam on the first try. For this purpose, the students are described by the following features: *decile1b* (decile based on the grades in the first year), *decile3* (decile based on the grades in the thrid year), *lsat* (score), *ugpa* (undergraduate GPA), *zfygpa* (first year law school GPA), *zgpa* (cumulative law school GPA), *fulltime* (full-time or part-time work) and

fam_inc (family income bracket). Besides these features, *male* (whether the student is male or female) and *race* (white or non-white) is also provided. In our analysis, both of them are assumed as sensible features. However, when evaluating disparities with respect to race (resp., gender), we considered *male* (resp., *race*) as an input feature in model training. This dataset is highly unbalanced and, after under-sampling, we obtained 3672 samples (originally, there were more than 45k samples).

Unfair results in the LSAC dataset are frequently associated with race. We provide in Figures 4 and 5 the interpretations on the probability of success in the bar exam and on the Statistical Parity, respectively. One may see that, when evaluating fairness in race, there is a huge disparity between the two groups (see Figure 4a). The probability of positive outcome (i.e., being classified as success in the exam) for non-whites is much lower than for whites. This disparate result can be explained by the features contributions presented in Figure 4c, where most of them (specially *lsat*) contribute more to classify whites as the positive class than non-whites. On the other hand, as can be seen from Figure 4b, the probability of favorable outcomes is practically the same regardless the gender and the adopted threshold. Moreover, in this scenario, all features contribute equally to classify as successfully passing the bar exam both males and females (see Figure 4d). Therefore, we could note that features have different impacts when evaluating disparities for different sensitive features. Moreover, the performance and fairness measures have different shapes, which indicate that, in a threshold analysis, the choice of an appropriate threshold to enhance fairness should be conducted individually. However, in a scenario where two or more sensitive features are considered

simultaneously (what we generally refer as to intersectional fairness), a trade-off analysis should be conducted.

The aforementioned findings can also be seen in Figure 5. The statistical parity presented in Figures 5a and 5b attests the existence of disparities between races and the absence of unfair outcomes related to gender, respectively. With respect to the contributions of features (see Figures 5c and 5d), while there are contributions towards unfair results for race, they are almost zero for gender. Finally Figure 5e presents the contributions of features along with thresholds, where *lsat* appears as the most relevant information that leads to the disparate results.

4.3. Adult income dataset

In this last scenario, we considered the Adult income dataset. The goal in this dataset is to predict whether a person makes over 50K per year. As *gender* is one of the available information, disparities can be noted when comparing males and females. Therefore, we assumed gender as the sensible feature in our analysis, with female and male being the protected and the privileged groups, respectively. The remaining features used to train the model are[6] *age* (intervals - less than 25, greater than 60 and between 25 and 60 years old), *workclass* (Private or Non-private), *educational-num* (numerical value associated with the education degree - the greater the better), *marital-status* (married, never-married or other), *relationship* (Wife, Own-child, Husband, Not-in-family, Other-relative or Unmarried), *race* (White, Asian-Pac-Islander,

[6]We preprocessed this dataset in order to group categories of some features. Moreover, we also removed categorical features with high number of categories. Most of these changes have already been discussed in [42].

Amer-Indian-Eskimo, Black or Other), *capital-gain*, *capital-loss*, *hours-per-week* (hours of work per week) and *native-country* (US or non-US). The total number of samples are 22416.

As in this dataset one observes less females being classified as making over 50k per year (in comparison with males), we investigated such a disparity by means of the Equalized Odds. Recall that in Equalized Odds one considers true positive and false positive rates. Figures 6a and 6b present a comparison between sensitive groups with respect to FPRs and TPRs, respectively. Both figures attest that there are less females classified as making over 50k per year than males, either if the classification is correct (the TPRs) or incorrect (the FPRs). The interpretations in terms of equalized odds are presented in Figure 6c. The unfair outcomes can be explained by the features contributions provided in Figures 6e and 6d. It is interesting to note that both *marital-status* and *relationship* are the features that contribute the most toward the unfair result. Indeed, as some categories in *relationship* describe marital status, such features are somehow redundant. This explains the similar results in terms of contribution towards the equalized odds. Moreover, *relationship* is a feature that can be seen as a proxy for gender. For instance, the indication that a person is a Wife or a Husband is practically the same that saying the this person is a woman or a man, respectively. Therefore, even if gender is removed from training step, the use of proxy features such as *relationship* is a source of bias that can lead to disparate results.

4.3.1. Conducting feature selection

Although the main goal of this paper is to propose an approach to evaluate features contributions toward unfair results along with threshold values,

our method could be helpful in feature engineering. As we detect features with high impact on disparate results, this information can be useful to conduct feature selection aiming at improving fairness. Indeed, by removing features that contribute to increase unfairness, one expects to mitigate ethical disparities.

In this subsection, we attest our proposal as a feature selection strategy on the Adult income dataset. As achieved in the previous section, both *marital-status* and *relationship* have high impact towards unfair results. Therefore, aiming at improving fairness, we could remove both features from the analysis[7]. By adopting the remaining features into the classifier, the obtained FPRs, TPRs and Equalized Odds are presented in Figure 7. As can be seen in Figures 7a and 7b, the performance of both protected and privileged groups are much closer in comparison with the model with features *marital-status* and *relationship* (see, for instance, Figures 6a and 6b. Therefore, we can attest the reduction of disparate results when features with high impact on unfairness are removed from the dataset. This finding can also be verified in Figure 7c, where the equalized odds decreased along with thresholds.

5. Conclusions

In this paper, we propose to investigate performance and disparity across various decision threshold(s) and quantifying the contribution of different features towards these two objectives. We have demonstrated the usefulness of our proposed approach with the help of three different case studies in-

[7]For the scope of this paper, we only consider improving fairness. Therefore, we do not further evaluate the impact on the performance measure by removing both features.

volving real datasets. For example, in the recidivism dataset, we showed the disparity decreased with the increasing value of threshold. In the adult income dataset, our proposed approach identified that the use of proxy features such as relationship is a source of bias that can lead to disparate results, even if the gender attributes is excluded whilst training the model. In the LSAC dataset, we noted that features may have different contributions towards different sensitive features. Indeed, in this dataset, some features impacted the statistical parity associated with race, however, they do not contribute towards disparity associated with gender.

In summary, this provided us a tool to identify the trade-offs between the quality of prediction and the disparities between protected and other groups. Moreover, one may identify which features contribute the most for both measures. This highlights an imminent use of our proposal. As illustrated in the Adult income dataset, the user may adopt our proposal to help conducting feature engineering by selecting features that do not entail disparate outcomes. As a result, one may improve fairness. However, it is important to see this feature selection impact into the model performance. Therefore, we see as a future perspective a trade-off analysis looking performance and fairness when conducting feature selection based on features contributions.

As another future perspective, we believe that it will be useful to validate the practical usefulness of our proposal through experimental studies, for example, to assess its impact in real world problems by doing survey-based studies and collecting feedback from stakeholders. Another important area of work will be to develop an interactive software tool for stakeholders that can help visualize the trade-off between performance and disparity in machine

learning classifier; and therefore, enabling them to choose a decision threshold based on their preferences. Finally, we would like to extend our approach to deal with multi-class classification problems. In this case, once we adapt both performance and fairness measures to have a single measure for each coalition of features, we are able to calculate the payoff are, therefore, the Shapley value and feature contributions.

Acknowledgments

Work supported by São Paulo Research Foundation (FAPESP) under the grants #2020/09838-0 (BI0S - Brazilian Institute of Data Science), #2020/10572-5 and #2021/11086-0. L. T. Duarte would like to thank the National Council for Scientific and Technological Development (CNPq, Brazil) for the financial support.

References

[1] J. M. Sheffield, R. Smith, P. Suthaharan, P. Leptourgos, P. R. Corlett, Relationships between cognitive biases, decision-making, and delusions, Scientific Reports 13 (2023) 9485.

[2] B. De Martino, D. Kumaran, B. Seymour, R. J. Dolan, Frames, biases, and rational decision-making in the human brain, science 313 (2006) 684–687.

[3] S. H. Ivanov, Automated decision-making, foresight 25 (2023) 4–19.

[4] S. Barocas, A. D. Selbst, Big data's disparate impact, California Law

Review 104 (2016) 671. URL: http://lawcat.berkeley.edu/record/1127463.

[5] A. Petrović, M. Nikolić, M. Jovanović, M. Bijanić, B. Delibašić, Fair classification via Monte Carlo policy gradient method, Engineering Applications of Artificial Intelligence 104 (2021). doi:10.1016/j.engappai.2021.104398.

[6] P. A. Grabowicz, N. Perello, A. Mishra, Marrying Fairness and Explainability in Supervised Learning, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, Seoul, Republic of Korea, 2022, pp. 1905–1916. doi:10.1145/3531146.3533236.

[7] M. Hardt, E. Price, N. Srebro, Equality of opportunity in supervised learning, in: Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 2016, pp. 3315–3323.

[8] T. Quan, F. Zhu, Q. Liu, F. Li, Learning fair representations for accuracy parity, Engineering Applications of Artificial Intelligence 119 (2023). doi:10.1016/j.engappai.2023.105819.

[9] G. Xu, T. D. Duong, Q. Li, S. Liu, X. Wang, Causality Learning: A New Perspective for Interpretable Machine Learning, ArXiv ID: 2006.16789 (2020). URL: http://arxiv.org/abs/2006.16789.

[10] N. Burkart, M. F. Huber, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245–317. doi:10.1613/JAIR.1.12228.

[11] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.

[12] D. Charte, F. Charte, M. J. del Jesus, F. Herrera, An analysis on the use of autoencoders for representation learning: Fundamentals, learning task case studies, explainability and challenges, Neurocomputing 404 (2020) 93–107. doi:10.1016/j.neucom.2020.04.057.

[13] C. Molnar, Interpretable machine learning, 2021. URL: https://christophm.github.io/interpretable-ml-book/.

[14] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[15] T. Begley, T. Schwedes, C. Frye, I. Feige, Explainability for fair machine learning, arXiv preprint arXiv:2010.07389 (2020).

[16] S. F. Nimmy, O. K. Hussain, R. K. Chakrabortty, F. K. Hussain, M. Saberi, Interpreting the antecedents of a predicted output by capturing the interdependencies among the system features and their evolution over time, Engineering Applications of Artificial Intelligence 117 (2023) 105596. doi:10.1016/j.engappai.2022.105596.

[17] W. Cai, A. B. Kordabad, S. Gros, Energy management in residential microgrid using model predictive control-based reinforcement learning and Shapley value, Engineering Applications of Artificial Intelligence 119 (2023) 105793. doi:10.1016/j.engappai.2022.105793.

[18] B. Rozemberczki, L. Watson, P. Bayer, H. T. Yang, O. Kiss, S. Nilsson, R. Sarkar, The Shapley Value in Machine Learning, in: IJCAI International Joint Conference on Artificial Intelligence, Vienna, Austria, 2022, pp. 5572–5579. doi:10.24963/ijcai.2022/778.

[19] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence 2 (2020) 56–67. doi:10.1038/s42256-019-0138-9.

[20] E. Albini, J. Long, D. Dervovic, D. Magazzeni, Counterfactual Shapley Additive Explanations, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, Seoul, Republic of Korea, 2022, pp. 1054–1070. doi:10.1145/3531146.3533168.

[21] D. Watson, Rational Shapley Values, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 2022, pp. 1083–1094. doi:10.1145/3531146.3533170.

[22] K. Aas, M. Jullum, A. Løland, Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, Artificial Intelligence 298 (2021) 103502. doi:10.1016/j.artint.2021.103502.

[23] G. D. Pelegrina, L. T. Duarte, M. Grabisch, A k-additive Choquet integral-based approach to approximate the SHAP values for local inter-

pretability in machine learning, Artificial Intelligence 325 (2023) 104014. doi:10.1016/j.artint.2023.104014.

[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, arXiv preprint arXiv:1908.09635 (2019). URL: http://arxiv.org/abs/1908.09635.

[25] M. Kearns, A. Roth, The ethical algorithm: The science of socially aware algorithm design, Oxford University Press, New York, USA, 2019. doi:10.1017/CBO9781107415324.004.

[26] Y. Wang, V. A. Nguyen, G. A. Hanasusanto, Wasserstein robust classification with fairness constraints, 2021. arXiv:2103.06828.

[27] N. Martinez, M. Bertran, G. Sapiro, Minimax pareto fairness: A multi objective perspective, in: Proceedings of the 37th International Conference on Machine Learning (ICML 2020), 2020, pp. 6711–6720.

[28] S. Liu, L. N. Vicente, Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach, Computational Management Science 19 (2022) 513–537. doi:10.1007/s10287-022-00425-z.

[29] G. D. Pelegrina, S. Siraj, Shapley value-based approaches to explain the quality of predictions by classifiers, IEEE Transactions on Artificial Intelligence (2024) 1–15. doi:10.1109/TAI.2024.3365082.

[30] M. T. Ribeiro, S. Singh, C. Guestrin, "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, 2016, pp. 1135–1144.

[31] S. Verma, J. Rubin, Fairness definitions explained, in: 2018 ACM/IEEE International Workshop on Software Fairness, IEEE, 2018, pp. 1–7.

[32] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. Zemel, Fairness through awareness, in: 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012), Cambridge, MA, USA, 2012, pp. 214–226. doi:10.1145/2090236.2090255. arXiv:1104.3913.

[33] L. S. Shapley, A value for n-person games, in: W. Kuhn, A. W. Tucker (Eds.), Annals of mathematics studies: Vol. 28. Contributions to the theory of games, Vol. II, Princeton University Press, Princeton, 1953, pp. 307–317.

[34] L. Han, T. Morstyn, M. McCulloch, Incentivizing Prosumer Coalitions With Energy Management Using Cooperative Game Theory, IEEE Transactions on Power Systems 34 (2019) 303–313.

[35] H. P. Young, Monotonic solutions of cooperative games, International Journal of Game Theory 14 (1985) 65–72.

[36] S. Lipovetsky, M. Conklin, Analysis of regression in game theory approach, Applied Stochastic Models in Business and Industry 17 (2001) 319–330.

[37] M. Grabisch, P. Miranda, Exact bounds of the Möbius inverse of monotone set functions, Discrete Applied Mathematics 186 (2015) 7–12. doi:10.1016/j.dam.2015.01.016.

[38] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[39] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias - propublica, Propublica (2016). URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[40] L. F. Wightman, LSAC national longitudinal bar passage study, 1998.

[41] B. Becker, R. Kohavi, Adult, UCI Machine Learning Repository, 1996. doi:doi.org/10.24432/C5XW20.

[42] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsi, A survey on datasets for fairness-aware machine learning, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12 (2022) 1–59.

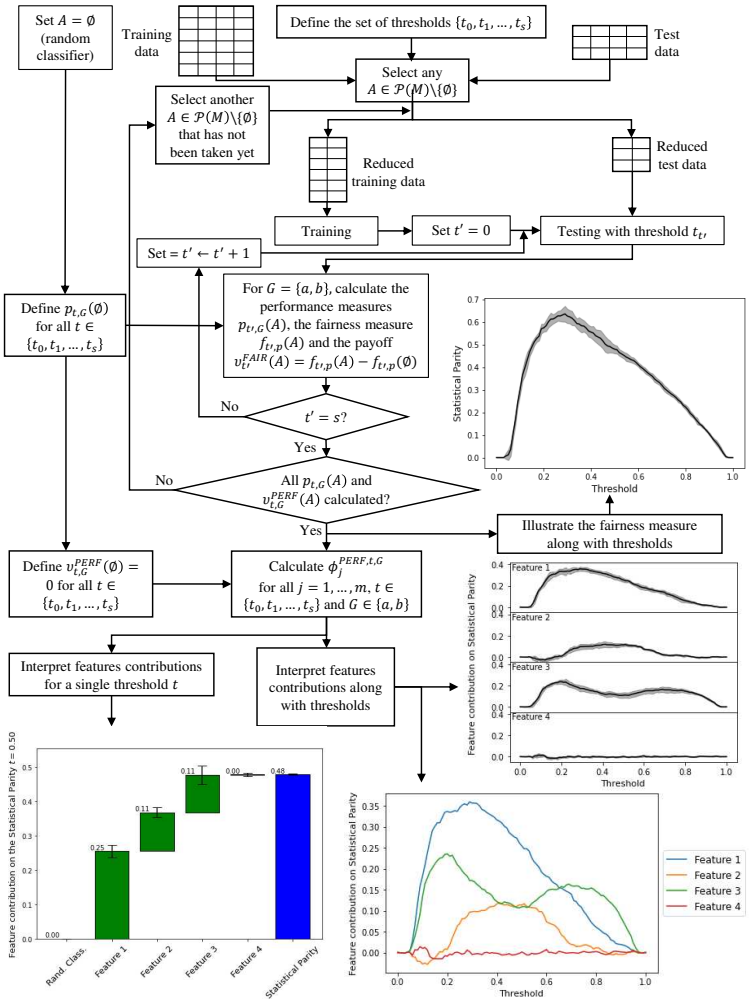
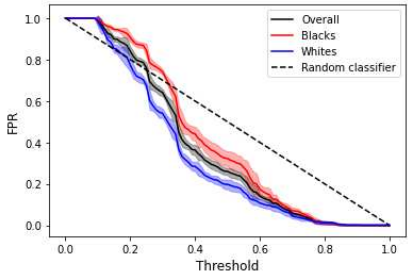
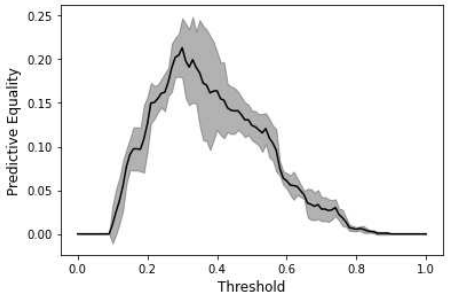


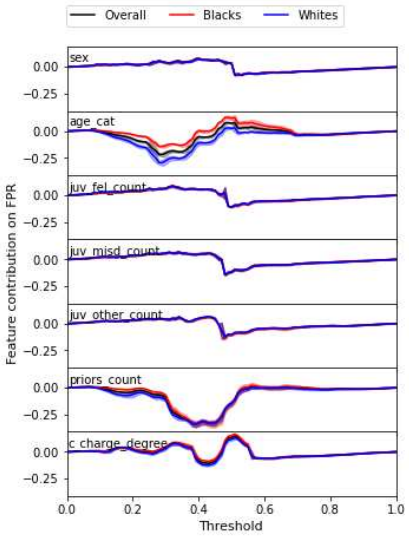
Figure 2: The proposed scheme to evaluate features contributions towards fairness measures.



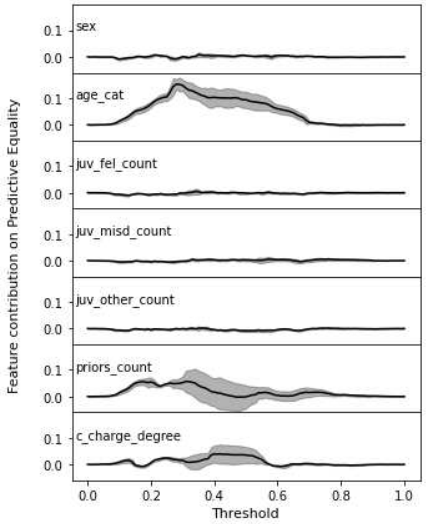
(a) FPRs along with thresholds.



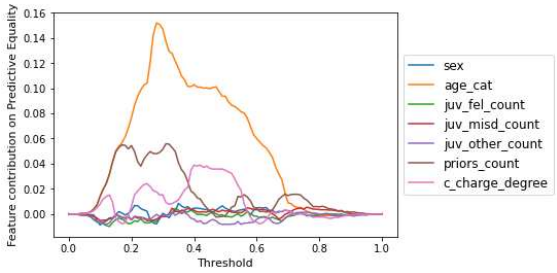
(b) Predictive Equality along with thresholds.



(c) Contributions of features towards FPRs (split).

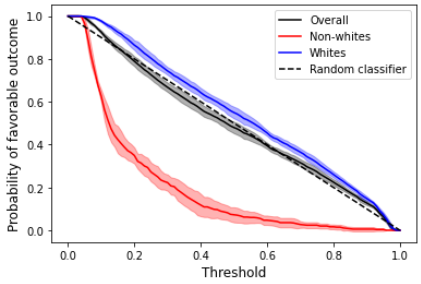


(d) Contributions of features towards Predictive Equality (split).

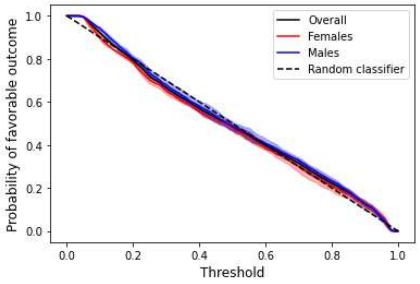


(e) Contributions of features towards Predictive Equality.

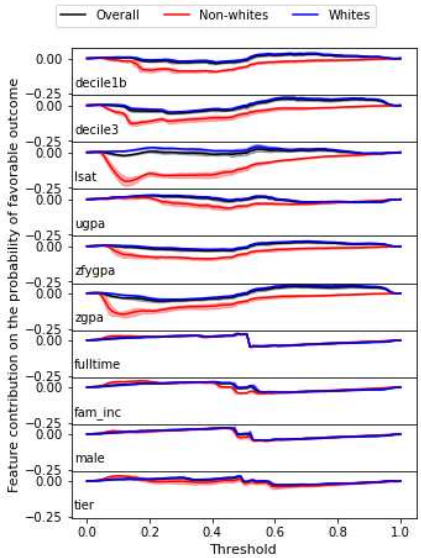
Figure 3: Interpreting disparate results towards FPRs and Predictive Equality - COMPAS dataset.



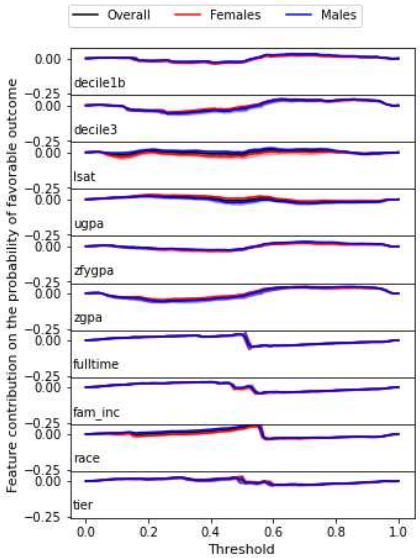
(a) Disparities between whites and non-whites.



(b) Disparities between males and females.

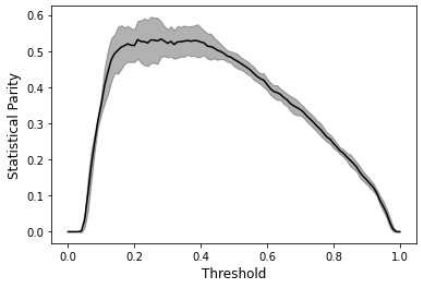


(c) Contributions of features (split) - Whites and non-whites.

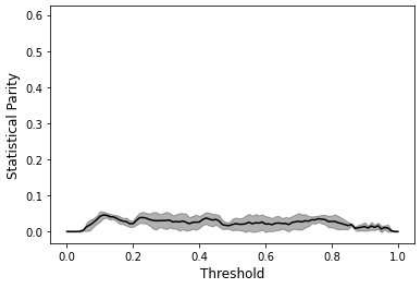


(d) Contributions of features (split) - Males and females.

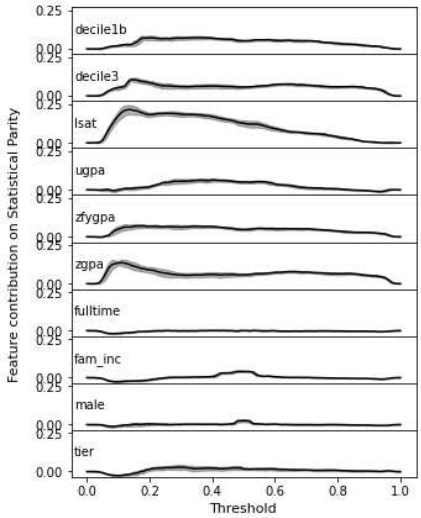
Figure 4: Interpreting disparate results towards the probability of success in the bar exam - LSAC dataset.



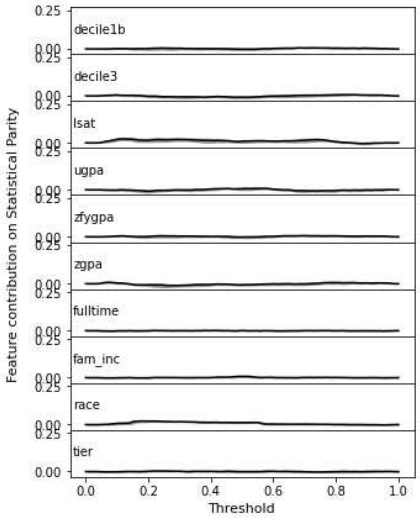
(a) Disparities between whites and non-whites.



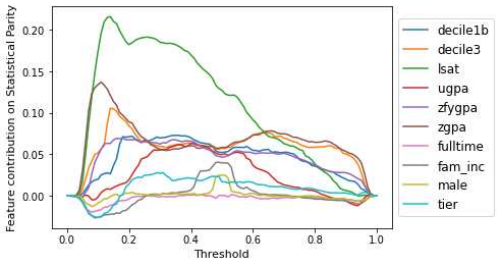
(b) Disparities between males and females.



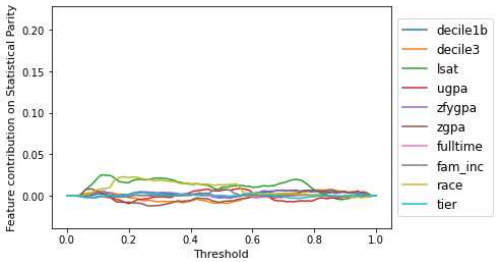
(c) Contributions of features (split) - Whites and non-whites.



(d) Contributions of features (split) - Males and females.

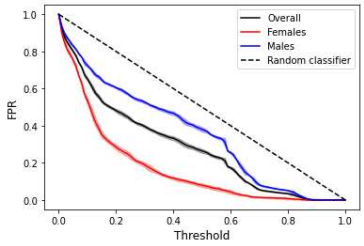


(e) Contributions of features - Whites and non-whites.

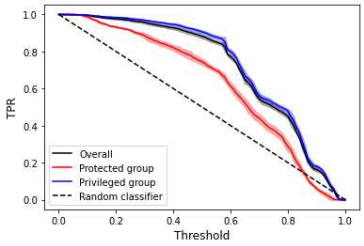


(f) Contributions of features - Males and females.

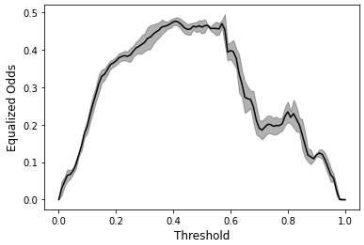
Figure 5: Interpreting disparate results towards the Statistical Parity - LSAC dataset.



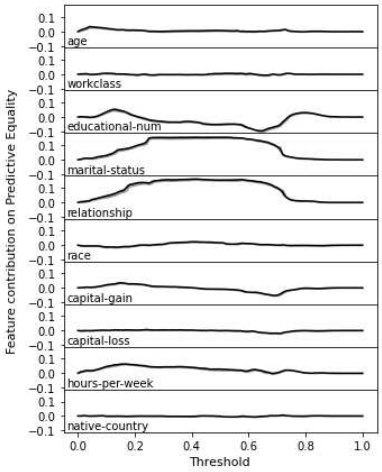
(a) FPRs along with thresholds.



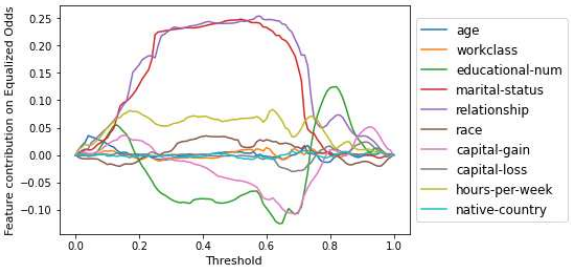
(b) TPRs along with thresholds.



(c) Equalized Odds along with thresholds.

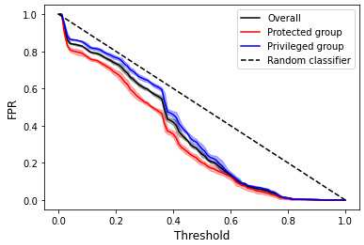


(d) Contributions of features towards Equalized Odds (split).

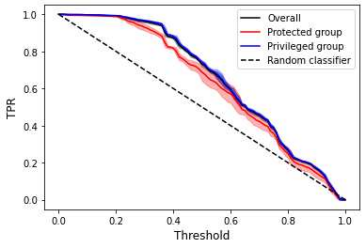


(e) Contributions of features towards Equalized Odds.

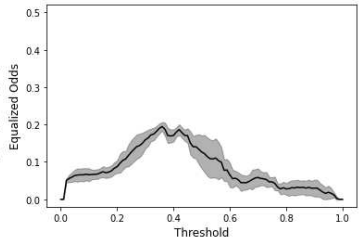
Figure 6: Interpreting disparate results towards FPRs, TRPs and Equalized Odds - Adult income dataset.



(a) FPRs along with thresholds.



(b) TPRs along with thresholds.



(c) Equalized Odds along with thresholds.

Figure 7: textcolorredVisualizing FPRs, TRPs and Equalized Odds after feature selection - Adult income dataset.