Subgroup Robustness Grows On Trees: An Empirical Baseline Investigation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Many methods for fair, robust, or disparity-minimizing machine learning have been proposed across the machine learning community, but thorough empirical evaluation of their subgroup robustness is lacking. In this work, we address this gap in the context of tabular data, where sensitive subgroups are clearly-defined, real-world fairness problems abound, and prior works often fail to compare to state-of-the-art tree-based tabular data models. We conduct an empirical study of several previously-proposed methods for fair and robust learning alongside state-of-the-art tree-based methods for tabular data and other baseline methods. Over experiments with more than $300,000$ model configurations on eight datasets, we show that tree-based methods have surprisingly strong subgroup robustness properties even when compared to robustness- and fairness-enhancing methods. We show that three families of metrics – those derived from accuracy, robust losses, and fairness metrics – show strong empirical agreement within their family, but display little or no correlation across these families for non-tree models. As a result, the best tree-based models tend to show good performance over a range of metrics, while robust or group-fair models can show brittleness, with significant performance differences across different metrics for a fixed model. Our work suggests that tree-based ensemble models make an effective baseline for tabular data, and are a sensible default when subgroup robustness is desired.

1 Introduction

Over the past decade, the field of machine learning (ML) has seen a dramatic expansion along two related lines. On one hand, concerns about the fairness of ML models, and more broadly their performance on data outside the training distribution, have grown [43]. These concerns have been raised in both theoretical and empirical works demonstrating the vulnerability of models to learn biases from data or suffer performance drops under distribution shifts [37, 59, 28]. On the other hand, an abundance of methods have been proposed to address these limitations. These include *fairness* methods used to equalize metrics across groups, as well as *distributional robustness* methods which optimize for some worst-case distribution within a bounded distance of the training distribution.

Our work begins from the observation that, while "fairness" and "robustness" are distinct concepts, methods in both areas often share the (sometimes implicit) criteria that models should have low *disparity* (variation in performance over all subgroups should be minimized; cf. [68, 15, 27, 36]) and good *worst-group performance* (the lowest accuracy over all subgroups should be maximized; cf. [67, 55]) – that is, both fairness and robustness typically entail improving *subgroup robustness*.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

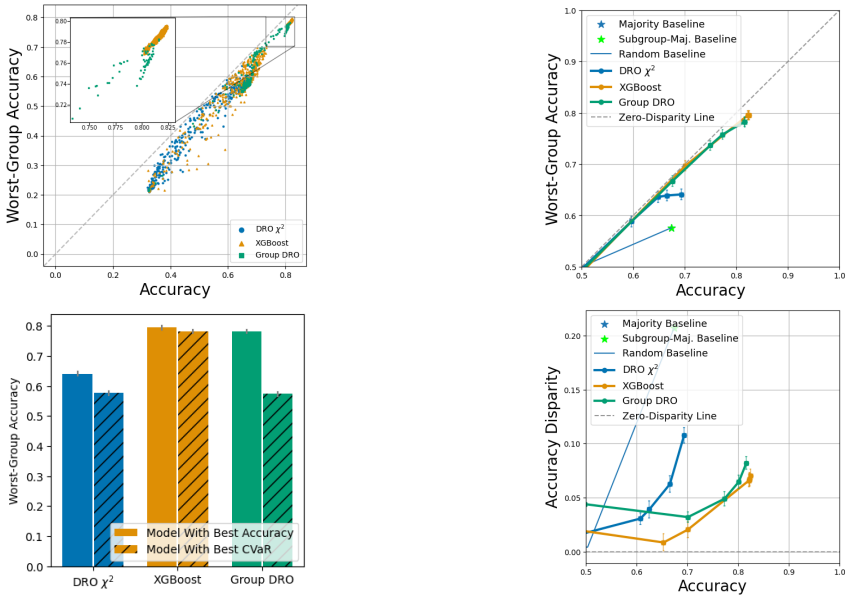

Figure 1: Results from the ACS Income dataset [17]. (a) Top left: Tree-based methods such as XGBoost show similar subgroup robustness, with sometimes better overall performance, as robustness-enhancing or disparity-mitigation methods. (b) Top right: Model performance frontiers corresponding to (a) show similar accuracy-robustness frontiers for XGBoost and DRO/Group DRO. (c) Lower left: Tree-based methods' worst-group accuracy is robust to model selection metrics. (d) Tree-based methods show similar accuracy disparity, with sometimes better overall performance, as robustness-enhancing or disparity-mitigation methods.

Our work evaluates subgroup robustness in the context of *tabular* data, for several reasons. First, tabular data is widely used in practice, being the most common format in areas with important fairness impacts such as medicine, finance, and recommender systems [31, 57]. Second, tabular data often directly encodes sensitive attributes with respect to which subgroup robustness is desired (race, gender, income level), and these attributes are often declared by the individuals represented in the dataset. Finally, tabular data is challenging to model. It is often heterogeneous, containing mixed data types; it lacks the spatial, linguistic, or temporal structures common to other data (image, text, audio) where machine learning has historically excelled; and it can be relatively small, containing personal information which cannot be scraped at Internet scale.

In this work, we conduct a series of experiments to jointly compare methods from the fairness and robustness literature. Despite the fact that many works often use similar metrics and datasets, there is a lack of large-scale empirical studies comparing relevant methods to each other, and to strong baselines. The latter, our analysis shows, is especially important, as strong baselines can outperform even well-tuned state-of-the-art disparity-mitigation methods with a fraction of the computational cost. Our analysis includes a set of tree-based models which achieve strong performance on tabular data but are absent from most works on robustness, which largely only compare to deep learning-based approaches despite their tendency to perform poorly on tabular data [31]. This lack of effective baselines is particularly concerning in light of recent works across several other areas of machine learning which have shown that simple baselines can perform surprisingly well against state-of-the-art techniques [40]. In our experiments, we endeavor to tune both the "fair" or "robust" models and a representative suite of baseline methods for tabular data to obtain estimates of performance "frontiers" for each technique. To this end, we explore a wide space of architectures, regularization schemes, and training hyperparameters, across five tabular datasets and 16 model classes. Our work is an empirical contribution concerned not with proposing new methods, but with conducting a novel and rigorous evaluation of existing methods for learning subgroup-robust models from tabular data. Our contributions are as follows:

Trees Are Subgroup-Robust Learners: We show that modern tree-based methods are surprisingly robust to subgroup shift, achieving competitive or better performance against state-of-the-art methods for fairness and robustness (Fig. 1a,1b) over a large hyperparameter and model search space. To support the above analysis, we apply several techniques for the analysis of hyperparameter sweeps such as model performance frontier curves (Fig. 1b,d), hyperparameter sensitivity analysis (Figs. 7,8), and model feasibility regions (Table 3, Figures 18,19).

Trees Show Consistent Performance Across Accuracy and Robustness Metrics: We empirically investigate the relationship between three types of metrics (accuracy, robustness, fairness). We show that metrics within these groups often agree, but the relationships across metric groups (i.e. robust CVaR risk and accuracy) are inconsistent. One consequence of this finding is that model selection techniques affect tree-based models and "robust" or fairness-enhancing methods differently: tree-based methods show consistent subgroup robustness across metrics (Fig. 1c), while robust and fairness-enhancing methods are "metric-brittle" and tend to display poor subgroup robustness according to other metrics (e.g. worst-group accuracy).

2 Related Work

2.1 Fairness-Enhancing Methods for Supervised Learning

A wide variety of works have addressed fairness in the context of machine learning, where "fairness" is often measured by the equalization of some metric over groups [5, 16]. Most methods can be characterized as performing either pre- in-, or post-processing, which attempt to ensure fairness considerations are met in different stages of the modeling lifecycle. *Preprocessing* [66, 13] attempts to modify the input data to meet fairness criteria, while preserving the structure of the inputs and their relationship to the prediction target. *Inprocessing* uses a modified training procedure to explicitly optimize for a fairness-aware objective during model training. This includes using constrained or reduction-based optimization [1, 65], or including explicit regularizers designed to encourage fairness [7, 9, 36, 49]. *Postprocessing* [26, 46] operates on the predictions of a model, modifying them to achieve fairness criteria.

The impact of fairness under subpopulation shift is analyzed in [41], which shows theoretically that enforcing fairness during training can harm or improve the model, under certain conditions, and [58] conducts a causal analysis of fairness under covariate shift. Perhaps the work most closely related to the current study is [20], which evaluates a set of pre-, in-, and postprocessing algorithms, demonstrating that many of these techniques show instability (variations in performance over different train-test splits) and that several fairness metrics are empirically correlated. [20] does not, however, evaluate robust learning techniques, modern tree-based techniques (besides CART), or neural methods, and empirical evaluation of recent methods is needed.

2.2 Distributionally Robust Learning

While efforts to learn models with good performance beyond the training distribution date back decades, recent works have adapted and extended this to deep learning. For example, several variants of distributionally robust optimization (DRO) have been proposed for the training of neural networks with robustness guarantees [29, 39, 55, 54, 67]. While these approaches are not always explicitly oriented toward fairness, they are frequently evaluated in terms of their performance benefits for minimizing performance gaps between sensitive subgroups in real-world data [67, 55, 27].

A particular form of shift (and robustness) evaluated by these works is *subgroup shift* (also called subpopulation shift) [43], where the balance of subgroups in the data shifts between training and testing. Evaluating performance on individual subgroups is an extreme version of subgroup shift. In the context of fairness, these subgroups are often defined by one or more discrete sensitive attributes, such as race, gender, or income level, and intersections between these sensitive groups can often identify groups most susceptible to performance disparities [12].

2.3 Models for Tabular Data

Deep learning-based approaches largely have not achieved the transformative performance gains on tabular data that they have with other data modalities such as images, text, and audio [31, 11, 57]. The existing state-of-the-art for learning with tabular data is widely acknowledged to be tree-based methods, and in particular gradient-boosted trees [25, 47, 57, 11]. This includes GBM [21, 22], LightGBM [33], XGBoost [14], and CatBoost [18, 48], which often show only small differences in performance between them. Several deep learning-based approaches have recently attempted to fill the gap of deep learning-based solutions for tabular data [31, 2, 30, 47, 32]. However, empirical analyses have shown that the reported performance of these methods does not generalize well to other datasets, and gradient boosting methods still achieve better performance with less tuning and a considerably smaller computational budget. For example, [57] and [11] both show, in separate analyses, that gradient boosting consistently outperforms these state-of-the-art tabular deep learning methods across several datasets, and tree-based methods achieve competitive performance with the deep learning approaches proposed in [23, 30].

While the overall predictive performance of these tabular methods has been evaluated, their fairness properties, subgroup performance, and robustness to subgroup shift have not been carefully investigated or compared with subgroup-robust learning methods. Additionally, despite the widely-known strong performance of these models on tabular data, we are aware of no prior work which compares tree-based methods to "robust" neural network-based learners, despite the fact that the latter are commonly evaluated on tabular data (e.g. [29, 36, 67, 55]). Notably, [1] evaluates GBM in conjunction with their proposed inprocessing method, but does not evaluate subpopulation robustness; [17] evaluates GBM with fairness methods but does not compare to robust methods and does not tune the GBM with fairness methods.

3 Setup

Our work is primarily concerned with empirically evaluating the sensitivity of various supervised learning methods to subpopulation shift. Below, we introduce our formal model, and then describe the main axes of variation in our experiments. This includes a large set of models, many of which have not been directly compared in previous works; a set of metrics used across the disparate literature on robustness, subpopulation shift, and fairness in machine learning; the hyperparameters used for each; and a suite of eight real-world fairness datasets.

3.1 Preliminaries

Our work evaluates the task of learning a model $f_\theta \in \mathcal{F}$, a function from model class $\mathcal{F}$ parameterized by θ. The parameters are learned from a dataset $\mathcal{D} \coloneqq (x_i, y_i)_{i=1}^n \sim \mathcal{P}$ where $\mathcal{P}$ is the data-generating distribution. This matches the case, most common in practice, where a single model is learned by estimating $min_\theta \mathbb{E}(\mathcal{L}(y, f_\theta(x)))$ and deployed for all users. We evaluate the binary classification context, where $y \in \{0, 1\}$ and $f_\theta(x) = \hat{y} \in [0, 1]$ is the predicted probability that $y_i = 1$. Each $x_i \in \mathbb{R}^d$ can be partitioned into $x_i = (x_{i,1}, \ldots, x_{i,d-1}, a)$ where a is a sensitive attribute. For notational convenience, we represent a as a single feature here, but in our data, a is typically defined as the concatenation of multiple binary sensitive attributes (e.g. gender = female, race = white). Let $\mathcal{D}_a \coloneqq \{(x_i, y_i) \sim \mathcal{P} | a_i = a\}$ denote the subgroup of the dataset with a given sensitive identity.

Following many of the previous works in both fairness [16] and robustness [67, 36], our analysis focuses on the performance of f_θ on each $\mathcal{D}_a \forall a \in A$. We are particularly interested in the worst-group performance and the loss disparity, respectively defined as

$$\mathcal{L}_{\text{WorstGroup}} \coloneqq min_a \mathbb{E}_{\mathcal{D}_a}[\mathcal{L}(f_\theta)] \quad \text{and} \quad \mathcal{L}_{\text{DISP}} \coloneqq max_{a,a'} |\mathcal{L}_a - \mathcal{L}_{a'}| \tag{1}$$

. These metrics can be thought of as assessing the sensitivity of f_θ to subgroup shift, evaluating the worst-group shift from $\mathcal{D}$ to $\mathcal{D}_a$ (note that subgroup shift is itself a form of covariate shift [43]).

3.2 Models

Our goal is to conduct a thorough empirical comparison of the subgroup robustness of a suite of methods for tabular data. We provide a more detailed description of all models used in Section B.

Fairness-Enhancing Models: We evaluate the LFR preprocessing method of [66]; the inprocessing method of[1], and the postprocessing method of [26]. As in [1, 17], we use GBM as the base learner for each model; however, unlike [17], we also tune the base learner parameters.

Robust Models: We evaluate both the CVaR and χ^2 constraint forms of DRO via [39]; the CVaR and χ^2 constraint forms of DORO [67]; Group DRO [55]; and Maximum Weighted Loss Discrepancy [36]. Each robust optimization method is used with a fully-connected multilayer perceptron (MLP), as in most previous works when using tabular data [67].

Tree-Based Models: We evaluate GBM, LightGBM [33], XGBoost [14], and Random Forest models.

Baseline Models: As baselines, we include L_2-regularized logistic regression, Support Vector Machines, and fully-connected neural networks (MLP) with standard (non-DRO) ERM optimization.

Hyperparameter Sweeps: For each model, we conduct a grid search over a large set of hyperparameters. We give the complete set of hyperparameters tuned for each model in Section F. Our hyperparameter search for each model is extensive by design in order to ensure a reliable comparison of the best-performing models from each class. We expand the initial grid for continuous hyperparameters when a model on the Pareto frontier is at the edge of the grid. When one method includes another as a "base" learner (e.g. LFR preprocessing with GBM, or DRO with MLP), we include the full tuning grid for the base model (i.e. we explore the cross-product of all MLP hyperparameters with all DRO hyperparameters). We note that previous works tend to either compare against a fixed baseline architecture and/or training hyperparameters (e.g. [67, 31]), perform manual tuning ([39]), or do not tune hyperparameters of the base model when using fairness-enhancing techniques ([17]).

3.3 Metrics

One goal of our work is to assess the empirical relationship between the model evaluation *metrics* used in the robustness, fairness, and classification literatures. This relationship is not explored in existing work, and these models optimize for a wide variety of different objectives during training. Additionally, there have been several metrics used across prior works to compare the performance of models, evaluate their fairness with respect to subgroups, and to measure their robustness to various shifts or outliers. We draw several metrics, from prior works and report them for our experiments.

For each metric $\mathcal{L}$ (e.g. loss, accuracy, Equalized Odds), we can not only measure the overall empirical performance of a model, but we can measure the *worst-group* loss and *disparity* over subgroups of the data, as defined in Equation (1). Worst-group and disparity measures, for various formulations of the loss functions above, are a widely-used measure of fairness and robustness [10, 67, 55, 54, 27, 37]. In particular, we use the $\mathcal{L}_{\text{DISP}}$ and $\mathcal{L}_{\text{WorstGroup}}$ with binary accuracy as the loss function. Accuracy is widely used in practice; directly interpretable; and invariant to rescaling of the model's predictions.

We also use the **CVaR risk** metric from the robustness literature. CVaR risk is measured over a set of inputs $\mathcal{D} = (x_i, y_i)_{i=1}^N \sim \mathcal{P}$ and measures the worst-case weighted loss, according to some loss function ℓ, at level α over the inputs in $\mathcal{D}$:

$$\mathcal{L}_{\text{CVaR}}(\mathcal{D}, \mathcal{P}) := \sup_{q \in \Delta^N} \sum_{i=1}^{N} q_i \ell(f_\theta; x) \quad \text{s.t. } ||q||_\infty \leq \frac{1}{\alpha N} \tag{2}$$

(2) is the risk function optimized by the DRO CVaR model [39], but it is also used more widely as a measure of the tail risk of a classifier (cf. [67]).

Additional fairness and robustness metrics are reported in Section G and defined in Section C.

Dataset	n	Features	Target	Sensitive Attributes
ACS Income	499350	20	Income $\geq$ 56k	Race, Sex
ACS Public Coverage	1138289	19	Public Ins. Coverage	Race, Sex
Adult	48845	14	Income $\geq$ 56k	Race, Sex
BRFSS	175745	28	Diabetes	Race, Sex
Comm. & Crime	1994	113	High Crime	Income Level, Race
COMPAS	7215	10	Recidivism	Race, Sex
German Credit	1000	22	Credit Risk	Age, Sex
LARC	169032	26	At-Risk (Grade)	URM Status, Sex

Table 1: Overview of datasets used.

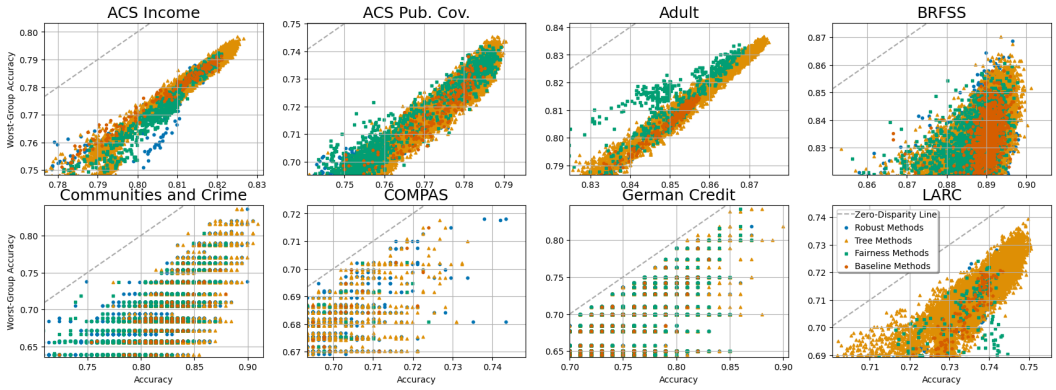

Figure 2: Overall Accuracy vs. Worst-Group Accuracy of robust, fairness-enhancing, tree-based, and baseline models over eight tabular datasets. Dashed lines indicate $y = x$, when worst-group accuracy equal to overall accuracy (zero accuracy disparity). See Supplementary Figure 9 for more detailed results by algorithm. Note: "discretization" artifacts are due to small test/dataset sizes.

3.4 Datasets

We evaluate the 16 models over a set of eight datasets covering a variety of prediction tasks and domains. We use a set of two binary sensitive attributes from each dataset, for a total of four nonoverlapping subgroups in each dataset. A summary of the datasets used in this work is given in Table 1. We provide additional details on each dataset, along with critical framing and context regarding these prediction tasks and their representations of individuals, in Section A.

4 Results: Tree Models are Subgroup-Robust Learners

Our main finding is that tree-based models are subgroup-robust learners; indeed, tree-based models match or exceed the performance of distributionally robust and fairness-enhancing models in terms of best overall accuracy, worst-group accuracy, and accuracy disparity on all datasets evaluated.

A summary of our results is shown in Figures 2, 3, and 4. Following previous works, our main evaluation metrics are accuracy-based: overall accuracy, worst-group accuracy, and accuracy disparity. Over the wide sweep of datasets and model configurations evaluated, modern tree-based models – GBM, LightGBM, Random Forest and XGBoost – all achieve subgroup performance characteristics on par with, or better than, the set of robust or fairness-enhancing models evaluated. We note that this is particularly surprising because none of the tree-based models explicitly optimize for robustness, fairness, or subgroup performance in any way; in contrast, the distributionally robust learners and fairness-enhancing techniques explicitly optimize for such criteria and in some cases (DORO, DRO, Group DRO) provide explicit guarantees of various forms of robustness. A similar analysis showing accuracy *disparity* is shown in Supplemental Figure 14, and a complete set of model performance observations from each algorithm and dataset, are in Supplementary Section G.

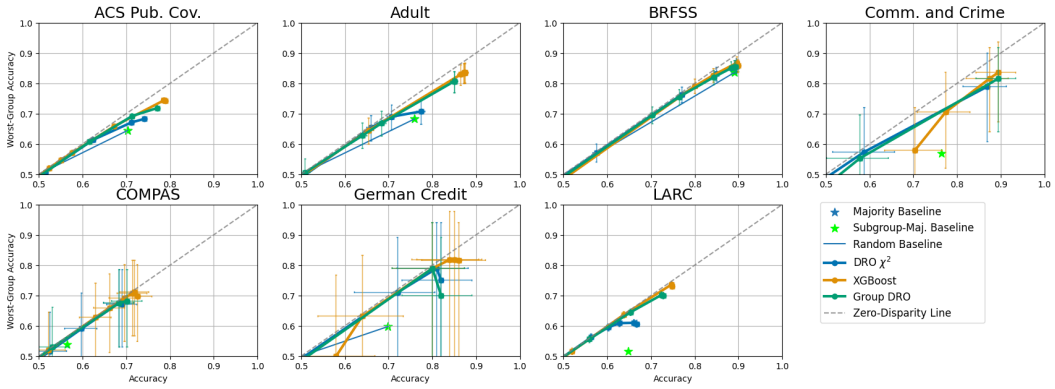

Figure 3: Model performance frontiers, formed by tracing the convex envelope of model performance. Tree-based models achieve comparable and sometimes improved frontiers with the highest-performing robustness methods. (See also Fig. 1b.)

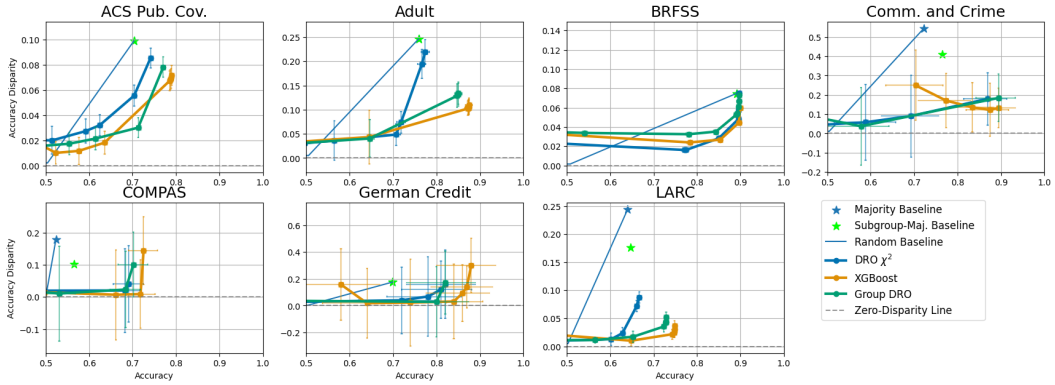

Figure 4: Model disparity frontiers, formed by tracing the convex envelope of model disparity. Tree-based models achieve comparable and sometimes improved frontiers compared to the highest-performing robustness methods, particularly in high-accuracy regions. (See also Fig. 1d.)

To summarize over the large hyperparameter sweeps, we propose two techniques. First, we use *model performance frontiers* to measure the best possible set of tradeoffs between ($\mathcal{L}$ and $\mathcal{L}_{\text{WorstGroup}}$) and ($\mathcal{L}$ and $\mathcal{L}_{\text{DISP}}$). Model performance frontiers represent the envelope of the convex hull of all observations of $\mathcal{L}(f_\theta) \in \mathcal{F}$. An example is shown in Figure 1b and 1d, with the remaining four dataset in Figure 3 and 4. The frontiers allow us to compare the best-possible tradeoff between accuracy and worst-group accuracy (Fig 3; higher is better) or between accuracy and accuracy disparity (Fig 4; lower is better).

Second, we propose *feasible model regions* as a measure of the full space of $(\mathcal{L}, \mathcal{L}_{\text{WorstGroup}})$ values observed over $f_\theta \in \mathcal{F}$. By comparing the area of the feasible model regions, we can compare the sensitivity of $(\mathcal{L}, \mathcal{L}_{\text{WorstGroup}})$ over the space of possible models $f_\theta \in \mathcal{F}$. We provide pseudocode for computing both the feasible model regions and model performance frontiers in Algorithm 1 of Section D. We give a more detailed analysis of feasible model regions in the Supplementary Section D; here we only note that the area encompassed by the FMR for DRO-based methods is consistently the largest across all datasets, as shown in Table 3.

These results are particularly important in light of the large differences in resources required to achieve similar levels of performance between robust models and the tree-based models: for example, the full hyperparameter grid sweep of size $12k$ XGBoost on the largest dataset in our study, ACS Income, completed in 1 CPU-day; DRO χ^2 sweep of 3250 training runs completed in 58 *GPU*-days.

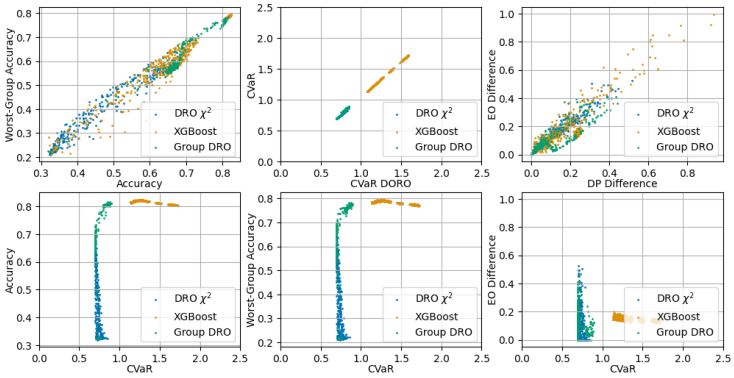

Figure 5: Empirically, "complementary" metrics (§5, top row) are closely correlated for all models. Non-complementary metrics (bottom row) are not correlated consistently across models. Accuracy can vary widely for "robust" models with a fixed robust (CVaR) risk, while for tree-based models, the best (lowest) CVaR risk is typically associated with the best (highest) accuracy.

5 Results: Relationship Between Fairness, Robustness, Performance Metrics

Our results show that, over all models and datasets, metrics which we refer to as "complementary" – those measuring the same event on different subsets, or with different conditioning – are strongly correlated with each other: accuracy and worst-group accuracy; CVaR and CVaR DORO; and Demographic Parity Difference and Equalized Odds Difference all show strong correlation with each model class $\mathcal{F}$; we show these correlations on Adult in Fig. 5. The median ρ values over all datasets and algorithms are 0.87, 0.99, 0.79 for the three pairs of metrics, respectively (we show the complete set of correlations for each model and metric in Figure 16). These results show that model selection based on one metric from each pair (Overall Accuracy) is likely to lead to a model with strong performance for the other metric in the pair (e.g. Worst-Group Accuracy).

In contrast, our experiments show that there is a severe lack of correlation across almost all of the non-complementary pairs, shown in the bottom row of Figure 5. In particular, our results show that models which achieve a strong "robust" risk – for example, a low CVaR DORO risk, one model selection rule used in [67] – can achieve very low accuracies. Indeed, the "robust" learning methods are the most susceptible to this discrepancy over metrics in our experiments, as shown in Figure 6. While these results confirm the finding in previous works that fairness metrics are correlated [20], our broader comparison shows a brittleness to model selection metrics outside these sets of complementary metrics, and reveal the very strict sense in which robustness guarantees apply only to a limited range of metrics. This discrepancy is of practical significance given that robust models are often selected using robust metrics (i.e. DORO CVaR, cf. [67]), while models in practice are frequently evaluated using other metrics (i.e. accuracy, fairness metrics).

6 Results: Robust Models Can Be Metric-Brittle

In this section, we explore the impact of various strategies for choosing a single $f_\theta \in \mathcal{F}$ over a set of models in each hyperparameter grid. While this problem is implicitly addressed in almost all previous works, the decision is often handled differently – for example, models are tuned and selected by hand [39], chosen based on worst-group performance [67], or based on robust risk metrics [55] – but the implications of various model selection methods for fairness and robustness are not well-understood. Our large-scale empirical study provides valuable data to help address this question.

Figure 6 demonstrates the downstream impact of the lack of correlation between robust risk metrics and worst-group accuracy: distributionally-robust models can suffer significant drops in worst-group accuracy, when evaluated based on their robust risk. In contrast, tree-based models show "metric

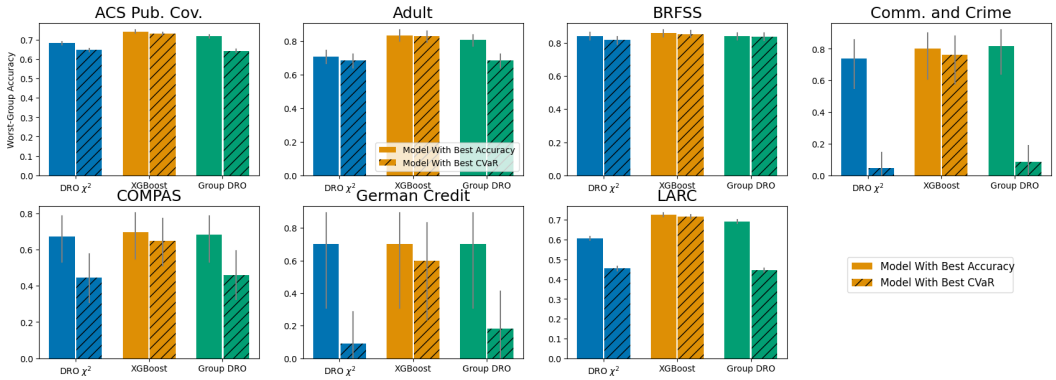

Figure 6: Worst-group accuracy for models with best overall accuracy (solid), and best CVaR (shaded). Tree-based models show better performance across non-complementary metrics, whereas robustness- and fairness-based models do not: worst-group accuracy drops significantly between best-accuracy and best-CVaR DRO χ^2 and Group DRO models (see also Fig. 1d).

robustness" – that is, tree-based models which perform best according to a robust risk measure still achieve worst-group accuracy near the highest-accuracy model of the same class (here, XGBoost).

7 Conclusion

The field of machine learning has made significant progress in the past several years in identifying and addressing various challenges which can contribute to real-world performance disparities. However, our results suggest that another form of progress not widely acknowledged as having implications for fairness or robustness – advances in tabular data modeling with tree-based algorithms – have similar, if not greater, impact in practice than the state of the art in robust neural network learning or fairness-enhancing methods. Our results suggest that tree-based algorithms are a surprisingly strong baseline for subgroup robustness in tabular data, and that future work should compare against, and improve upon, the subgroup robustness of tree-based models. Our work suggests that, in practice, tree-based ensembles are an effective default for tasks where subgroup robustness is desired.

In addition to being surprising, given the attention and effort dedicated to the recent advances in robust and fair learning, our results are important because they suggest that subgroup robustness on par with existing state of the art can be achieved with tree-based classifiers that are easier and computationally cheaper to tune and train than either robust or fairness-enhancing models, which tend to scale poorly with dataset size (in both n and d). Our work thus contributes a critical baseline, similar to how other recent works have contributed much-needed baselines in areas of machine learning affected by the rapid proliferation of new methods [52, 60, 40].

Our findings are limited to the set of hyperparameters and models explored in our experiments – a *superset* of those considered across the prior works discussed above (e.g. [55, 67]). Our findings do not demonstrate that robust learning methods *cannot* achieve subgroup robustness on par with e.g. XGBoost, but merely that this is not achieved with the configurations widely used in the literature. Future work should investigate the subgroup robustness of non-MLP models trained using robust techniques.

While our experiments do not identify the *cause* of trees' subgroup robustness, it is likely that this is a consequence of their strong performance on tabular data. These findings can be also viewed as an analogue of the empirical relationship between in-distribution and out-of-distribution accuracy in computer vision documented in [42] but now demonstrated for the tabular domain. It is possible that these improvements are due to (i) an inductive bias in tree-based models beter suited to modeling differences in subpopulations in tabular data, or (ii) the *ensembling* used by the tree-based models in this work. We leave an identification of such factors to future work.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[2] Sercan O Arık and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *AAAI*, volume 35, pages 6679–6687, 2021.

[3] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*, 2021.

[4] Matias Barenstein. Propublica's compas data revisited. *arXiv preprint arXiv:1906.04711*, 2019.

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. 2017.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[7] Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. 2017.

[8] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[9] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[10] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

[11] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *arXiv preprint arXiv:2110.01889*, 2021.

[12] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[13] L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International Conference on Machine Learning*, pages 1349–1359. PMLR, 2020.

[14] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[15] Jianfeng Chi, Yuan Tian, Geoffrey J Gordon, and Han Zhao. Understanding and mitigating accuracy disparity in regression. In *International Conference on Machine Learning*, pages 1866–1876. PMLR, 2021.

[16] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[17] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[18] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[19] Petros Drineas, Michael W Mahoney, and Nello Cristianini. On the nyström method for approximating a gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(12), 2005.

[20] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338, 2019.

[21] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[22] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[23] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[24] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512, 2020.

[25] Vasyl Harasymiv. Lessons from 2 million machine learning models on kaggle, Dec 2015.

[26] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[27] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[28] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[29] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[30] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

[31] Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Well-tuned simple nets excel on tabular datasets. *Advances in Neural Information Processing Systems*, 34, 2021.

[32] Liran Katzir, Gal Elidan, and Ran El-Yaniv. Net-dnf: Effective deep modeling of tabular data. In *International Conference on Learning Representations*, 2020.

[33] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[34] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

[35] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 100–109, 2019.

[36] Fereshte Khani, Aditi Raghunathan, and Percy Liang. Maximum weighted loss discrepancy. *arXiv preprint arXiv:1906.03518*, 2019.

[37] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[38] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[39] Daniel Levy, Yair Carmon, John C Duchi, and Aaron Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860, 2020.

[40] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[41] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. Does enforcing fairness mitigate biases caused by subpopulation shift? *Advances in Neural Information Processing Systems*, 34:25773–25784, 2021.

[42] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

[43] Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[46] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[47] Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint arXiv:1909.06312*, 2019.

[48] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[49] Edward Raff, Jared Sylvester, and Steven Mills. Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 243–250, 2018.

[50] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

[51] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008.

[52] Steffen Rendle, Li Zhang, and Yehuda Koren. On the difficulty of evaluating baselines: A study on recommender systems. *arXiv preprint arXiv:1905.01395*, 2019.

[53] Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *arXiv preprint arXiv:1811.00731*, 2018.

[54] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

[55] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[56] Morgan Klaus Scheuerman, Aaron Jiang, Katta Spiel, and Jed R Brubaker. Revisiting gendered web forms: An evaluation of gender inputs with (non-) binary people. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–18, 2021.

[57] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[58] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.

[59] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[60] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pages 266–282. Springer, 2020.

[61] Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *arXiv preprint arXiv:2205.04610*, 2022.

[62] Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

[63] Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

[64] Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. *Advances in neural information processing systems*, 25, 2012.

[65] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

[66] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[67] Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pages 12345–12355. PMLR, 2021.

[68] Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, et al. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. *Advances in Neural Information Processing Systems*, 32, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sec. F.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Our work is not concerned with point estimates, so we do not display error bars in e.g. our scatterplots. However, we do fix random seeds for reproducibility and provide extensive empirical results over a wide grid of experimental parameters in order to provide a sense of the variability of our findings over different models, datasets, and learning hyperparameters.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section E.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use existing datasets, all of which are piblicly available, and the sources are cited and linked.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code to reproduce experiments.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We discuss issues with data representation in our introduction (which motivates the use of tabular data), as well as potential concerns raised regarding COMPAS and other datasets in Section A.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]