An Empirical Comparison of Bias Reduction Methods on Real-World Problems in High-Stakes Policy Settings

Hemank Lamba
Carnegie Mellon University
Pittsburgh, PA
hlamba@andrew.cmu.edu

Kit T. Rodolfa*
Carnegie Mellon University
Pittsburgh, PA
krodolfa@cmu.edu

Rayid Ghani*
Carnegie Mellon University
Pittsburgh, PA
rayid@cmu.edu

ABSTRACT

Applications of machine learning (ML) to high-stakes policy settings — such as education, criminal justice, healthcare, and social service delivery — have grown rapidly in recent years, sparking important conversations about how to ensure fair outcomes from these systems. The machine learning research community has responded to this challenge with a wide array of proposed fairness-enhancing strategies for ML models, but despite the large number of methods that have been developed, little empirical work exists evaluating these methods in real-world settings. Here, we seek to fill this research gap by investigating the performance of several methods that operate at different points in the ML pipeline across four real-world public policy and social good problems. Across these problems, we find a wide degree of variability and inconsistency in the ability of many of these methods to improve model fairness, but post-processing by choosing group-specific score thresholds consistently removes disparities, with important implications for both the ML research community and practitioners deploying machine learning to inform consequential policy decisions.

1. INTRODUCTION

There has been a recent increase in the use of machine learning models to support decisions in high stakes domains with societal impact, including informing bail decisions [20; 70; 6], hiring [61], healthcare delivery [57; 62] and social service interventions [9, 22, 60]. These decisions affect critical aspects of people's lives and if not done responsibly, can hurt already vulnerable and historically-disadvantaged communities. This combination of increased use, increased potential for improving social outcomes, and increased risk of harm has prompted questions from researchers, policymakers, citizens, and the media about the role these models can play in exacerbating (or reducing) existing inequities [36, 58, 54; 17], giving rise to a growing area, FairML, focused on dealing with issues of bias and fairness in building and using machine learning systems. FairML research has grown to span issues around defining bias in machine learning models, enumerating a variety of metrics that can be used to measure model bias [71], detecting instances of it through audit tools [67; 12], and methods for reducing (or mitigating the impact of) bias in ML models [17]. In this work, we focus on bias reduction methods, which can be broadly categorized into three groups based on the stage of analysis at which they are applied:

1. *Pre-processing methods* typically involve changing the data in some manner *before* building models.
2. *In-processing methods* typically involve using ML models/methods that are explicitly designed to deal with bias *in the process of building models*, such as using regularization approaches.
3. *Post-processing methods* typically involve adjusting scores, thresholds, or model selection *after the model predictions* have been generated.

Despite active research and the development of several new methods in the area of FairML in recent years, there has been a lack of extensive empirical evaluation across them to assess their effectiveness on real-world problems and data. The majority of this research has typically focused on achieving abstract, general-purpose definitions of fairness, and evaluated on benchmark data sets (such as the adult data set [23]) or limited data sets (such as COMPAS [50]). While that is a reasonable starting point, benchmark data sets often do not reflect the richness, nuances, and constraints of real-world problems, making it unclear for both researchers and practitioners how to assess the applicability and effectiveness of these methods or make decisions around which ones to use under given circumstances in real-world situations.

In this paper, we attempt to fill this empirical gap by presenting a **comprehensive empirical evaluation** of different bias reduction strategies over **four real-world problems** that come from public policy and social good settings. We want to emphasize that real-world problems are not just data sets but rather a combination of business/policy problems, the corresponding machine learning formulation and evaluation metrics, an extensive set of features generated in the feature engineering process, a large and varied set of ML models and hyperparameters, and a validation methodology and metric(s) that mirrors the deployment scenario. To that end, we describe the analytical formulation for each of these problems, the feature engineering process, parameters of temporal model selection using a wide variety of ML models and hyperparameters, and then evaluate the effectiveness of a variety of bias reduction strategies in reducing specific disparities while preserving as much of the original evaluation metric of interest as possible. We believe that this paper not only fills a critical gap today for researchers

*These authors contributed equally.

and practitioners of FairML but also provides a framework for researchers proposing new methods to follow when reporting the effectiveness of their work.

2. RELATED WORK

The focus of this paper is not on the entire process of building ML systems that lead to fair and equitable outcomes but more narrowly on methods that are used to reduce the bias in the predictions of ML models. With that focus, as mentioned earlier, bias reduction methods can be categorized broadly into three categories, based on the phase of the analysis pipeline to which they are applied: (a) Pre-processing, (b) In-processing and, (c) Post-processing.

2.1 Pre-processing

Pre-processing approaches assume that the bias in the ML models is caused by certain variables in the data or by the distribution of the data being used to train and validate the ML models. Most of the pre-processing approaches thus try to modify the data by either removing the sensitive variable (gender or race for example) or by changing the data distribution (with respect to the sensitive variable) by sampling.

Omission of sensitive variables has been widely explored in the past [29, 69]. This approach is based on the assumption that if machine learning model is not given the protected variable as a feature, the model that is trained will not be dependent on the protected variable, making the model unbiased. Unfortunately, this assumption is often overly-optimistic (and violated) in real-world problems where several other features, including ones relevant to the prediction problem, may be strongly correlated with the protected attribute. Recent work has described how omission of sensitive variables for training models often may not affect bias reduction (or even increase biases) despite decreasing model accuracy [16, 42, 24]. Despite these well-documented limitations, we included this strategy in the present exploration of fairness-enhancing methods because this notion of "fairness through unawareness" nevertheless persists and has commonly been posited by policymakers, decision-makers (in governments, non-profits, and corporations), and students we have worked with. Notably, other researchers have proposed more nuanced approaches to modifying the input data to remove correlations with the protected attribute in addition to the attribute itself. Although we do not explore this direction for pre-processing here, we refer the reader to [26] for an example of this approach in the context of disparate impact as a measure of fairness.

Resampling involves modifying the distribution of the training data by either over- or under-sampling examples to reduce disparities when the modified data is used in model training. Calders et al. [15] explored three different sampling techniques to fix existing bias in the data distribution to ensure that a model (in their case, Naive Bayes) trained on the modified data is more fair. Similarly, Iosifidis et al. [39] used clustering across sensitive attribute and labels to come up with representative training data to train models, and Kamiran et al. [42] explored multiple techniques involving sampling and re-weighing of training instances as pre-processing steps before applying machine learning models. Other popular preprocessing techniques involve relabelling and perturbation [44], details of which we omit from the paper.

2.2 In-processing

In-processing bias reduction methods generally include regularization or constrained optimization approaches to account for fairness metrics while solving their underlying classifier's optimization problem. Regularization adds penalty terms to the objective function of the classifier such that it is penalized for unfair solutions, whereas constrained optimization generally introduces fairness as a hard constraint in order to directly reject solutions that fail to satisfy fairness criteria. Kamishima et al. [43] proposed a regularization technique that uses mutual information of the sensitive attribute and prediction class, penalizing any increase in conditional probability on a specific subgroup. Zafar et al. [75] extended on this work by introducing fairness constraints into the objective function of the underlying classifier. One challenge faced by these approaches, however, is that these constraints often yield a non-convex objective function, making the optimization problem inherently difficult. To address this issue, Zafar proposed an efficient method for solving the resulting non-convex formulation. Similar techniques for different fairness metrics and even general classes of metrics have also been proposed in the literature [18]. Jiang et al. proposed an approach that minimizes Wasserstein-1 distances between classifier output and sensitive information [41]. Heidari et al. proposed a Rawlsian concept of fariness that can be introduced as a constraint into any convex loss-minimization algorithm [33]. Similar methods have also been extended to neural-network based models [52] as well as decision trees [3].

2.3 Post-processing

Post-processing methods are generally agnostic to the machine learning models used, and modify the outputs to improve fairness in predictions or classifications. This involves training meta-models with fairness constraints [18, 25] or directly thresholding or modifying model scores to improve fairness [32, 65]. Hardt et al. proposed methods for **direct post-hoc adjustments to scores** (or binary predicted classes) from trained classifiers to achieve either equalized odds or equality of opportunity by choosing group-specific thresholds that meet these fairness goals [32]. Recently, we have extended on this work, applying similar methods across a number of policy contexts and finding little or no trade-off in model accuracy in doing so [65, 64].

Another fairness-enhancing strategy that can be applied on top of a range of underlying machine learning methods involves **decoupling the training or selection of classifiers**, as proposed by Dwork and colleagues [25]. This approach starts from the hypothesis that a model trained to do well on the entire population might not fully capture differences in predictiveness of features or other important patterns across groups and posits that training separate models for each protected group might better pick up on these nuances. Because fully decoupling the models might significantly reduce the available training data (particularly for small groups), they also suggest exploring different levels of transfer learning between groups, giving a relative weight to training examples from the protected group or rest of the population (so, at the other extreme, one might train models across the full population, but select best-performing models for each group rather than a single overall model).

Other authors, including Celis et al [18] as well as Menon and Williamson [55], have proposed methods that perform a **constrained optimization to train a meta-model** to

improve the fairness of a prediction score generated by a model. These methods seem particularly useful where membership in the protected groups is not known apriori but can be estimated (for instance, [18] describes estimating a joint probability distribution over outcomes and sensitive attributes). However, when group membership is known, these methods will generally result in stretching or shifting within-group score distributions without reordering in a manner equivalent to choosing separate thresholds for each group (for more detail, see our discussion in the supplemental materials from [64]).

Finally, and perhaps most simply, fairness can be incorporated into the process of **model selection**. After training a large set of different model types and hyperparameter values, the validation set performance of these different trained models can be assessed both in terms of traditional accuracy metrics (such as AUC-ROC, precision@k, or other confusion matrix based metrics) as well as fairness metrics appropriate to the context. Choosing a model to deploy then becomes an optimization problem over two dimensions, with a Pareto frontier reflecting a menu of potential trade-offs between these two goals of accuracy and fairness [72]. In practice, the trade-offs presented by this frontier might be a function of inherent properties of the data and problem as well as the extent to which the grid search that was performed covers the possible space of model types and hyperparameters. Although relatively straightforward in nature and implementation, relying entirely on model selection is somewhat arbitrary as it relies entirely on finding a model specification that performs well on both fairness and accuracy metrics without taking active steps to ensure or improve fairness.

3. COMPARISON SETUP

This section describes our setup to conduct the empirical evaluation across bias reduction methods. We describe the specific methods we chose to compare, the policy contexts for the problems we use to conduct that empirical evaluation, and the specific experimental setup for each real-world problem (the data used, features generated, models built, evaluation metric, protected group, and bias metric).

3.1 Methods to Compare

While a large number of bias reduction methods exist in each category we describe in Section 2 (Pre-processing, In-processing, and Post-processing), in this paper, we focus on a few representative methods from each category to compare with each other. The methods chosen for this study are described below.

3.1.1 Pre-Processing Methods

Removing the Protected Attribute: For each problem domain, we define a set of protected attributes and remove those from the data before performing any ML modeling.

Sampling: We apply sampling to our training sets with respect to the protected group in three ways: a) changing the marginal distribution of the protected and non-protected subgroups, b) changing the label distribution within the protected and non-protected subgroups, and c) changing both simultaneously. Here, we implemented the six sampling strategies described in Table 1 reflecting a set of reasonable a priori hypothesis about how these distributions in the training data might influence model fairness.

To formalize our sampling approaches, we define *Protected* as the protected value/group (such as Race=Black) and *NonProtected* as the set of values that are considered Non-Protected (such as Race=White). The (binary) label variable is represented as Y with values 0 and 1. $P^0(\cdot)$ represents a probability distribution in the original dataset and $P'(\cdot)$ represents a probability distribution after resampling. With those definitions, each of our three sampling settings are:

(A) Balances the data by changing the ratio of Protected to Non-Protected while preserving the original label distribution within each group.

The goal is to achieve:

$$\frac{P'(NonProtected)}{P'(Protected)} = \alpha$$

while preserving the original label distribution within *Protected* and *NonProtected* such that

$$P'(Y = 1 \mid NonProtected) = P^0(Y = 1 \mid NonProtected)$$
$$\text{and } P'(Y = 1 \mid Protected) = P^0(Y = 1 \mid Protected)$$

In Table 1, Strategy 1 uses this approach with $\alpha = 1$.

(B) Balances the label distribution across each subgroup: Protected and NonProtected. The goal is to achieve:

$$P'(Y = 1 \mid NonProtected) = \beta_{NP}$$
$$P'(Y = 1 \mid Protected) = \beta_P$$
$$\text{such that } \frac{\beta_{NP}}{\beta_P} = \gamma$$

while preserving the original marginal distributions for Protected and NonProtected such that:

$$P'(NonProtected) = P^0(NonProtected)$$
$$\text{and } P'(Protected) = P^0(Protected)$$

In Table 1, Strategy 2 uses this approach (with $\beta_P = \beta_{NP} = 0.5$ and $\gamma = 1$), as does Strategy 3 (with $\beta_P = \beta_{NP} = P^0(Y = 1 \mid NonProtected)$ and $\gamma = 1$) and Strategy 4 (with $\beta_{NP} = P^0(Y = 1 \mid NonProtected)$ and $\beta_P = 0.5$).

(C) Adjusts the marginal distribution of Protected and Non-Protected as well as the label distributions by setting α, β_P, β_{NP}, and γ as described above.

In Table 1, Strategy 5 uses this approach (with $\alpha = 1$ and $\beta_P = \beta_{NP} = 0.5$) as does Strategy 6 (with $\alpha = 1$, and $\beta_P = \beta_{NP} = P^0(Y = 1 \mid NonProtected)$).

Note that in each strategy, in order to balance two distributions, we can either *undersample* from the majority distribution or *oversample* from the minority distribution. In case of oversampling, we randomly sample (with duplicates allowed) to generate more examples[1] increasing the total number of examples as little as possible while achieving the desired distributions. When undersampling, we remove as few examples as possible in order to achieve the desired distributions. Also note that we only sample in each training set while keeping the distribution of the validation sets the same as in the original data.

[1]For oversampling, we do not make use of methods such as SMOTE [19] as each feature might have a specific set of constraints and this method does not take into account the overall joint distribution.

Table 1: Sampling strategies used in this study.

	Ratio: Protected to Non-Protected	Label Dist. Protected	Label Dist. Non-Protected
1	1:1	Original	Original
2	Original	50-50	50-50
3	Original	Same as Non-Protected	Original
4	Original	50-50	Original
5	1:1	50-50	50-50
6	1:1	Same as Non-Protected	Original

3.1.2 *In-Processing Methods*

In this paper, we focus on in-processing through constrained optimization to reduce model disparities. This approach includes fairness metrics in the objective function and seeks to produce predictions that maximize accuracy while taking fairness into account.

Zafar and colleagues [75, 74] proposed a constrained optimization method centered on a fairness notion they described as "disparate mistreatment." A model can be said to have disparate mistreatment when misclassification rate for the protected and non-protected group are different, and their work described optimization problems using either False Positive Rate (FPR) or False Negative Rate (FNR) as a measurement of misclassification. Formally, this optimization problem (for FNR) is defined by:

$$\min L(\theta)$$
$$\text{s.t. } P(\hat{y} \neq y \mid z = 0, y = 1) - P(\hat{y} \neq y \mid z = 1, y = 1) \leq \epsilon$$
$$P(\hat{y} \neq y \mid z = 0, y = 1) - P(\hat{y} \neq y \mid z = 1, y = 1) \geq -\epsilon$$

where, L is the loss function (over model parameters θ), $\hat{y}$ prediction, y original label, z is the protected attribute, and ϵ denotes the tolerance boundaries for a fair output.

For our problem settings, we focus on True Positive Rate (TPR) disparities (also referred to "equality of opportunity" by Hardt [32]) as the appropriate metric of fairness (see the discussion on problem settings below, as well as in [66, 64]). However, because $TPR = 1 - FNR$, we make use of Zafar's method to equalize FNR. In doing so, we used a very small value of $\epsilon = 0.0001$ to find solutions which remove disparities entirely.

Recently, open source toolkits such as FairLearn [13] have also been introduced which try to reduce biases, according to a given metric in classification problems. However, we do not include FairLearn in this study setting because it only generates binary predicted class labels rather than a continuous score. This makes it poorly suited to our problem settings where we focus on choosing the k highest-risk entities for intervention based on an organization's resource constraints (as discussed in more detail below). In other work, we have explored heuristics such as sampling to select top k predictions from the output of FairLearn but found that it performed poorly since it wasn't designed for that purpose [68].

3.1.3 *Post-Processing Methods*

We define the post-processing class of methods as any method that is applied once the model has been built, typically in adjusting the scores that the models produced or using different thresholds to create classification decisions. We describe several such methods above and discuss here the methods we explored in the present work.

Post-Hoc Adjustments: Here we expand on some of our recent work [65, 64] using a method to equalize TPRs across groups while keeping the total number of individuals selected constant, reflecting the "top k" setting of the policy problems we consider (see the discussion on problem settings below for more details). In short, because TPR increases monotonically with depth in a predicted score, we can find a single solution (up to randomized tie breaking) with equalized TPR across groups by adjusting the score thresholds for each group while keeping the total number of individuals selected constant. In practice, these threshold adjustments are made on the model scores in one validation split (say, at time $t = 0$) to decompose the overall number of individuals to select by group,[2] then these group-specific target numbers are applied to a subsequent validation set to evaluate how well this fairness-enhancing strategy generalizes into the future. Note that, as mentioned above, some of the meta-model approaches such as those described in [18, 55] can be shown to be mathematically equivalent to choosing different score thresholds when protected group membership is known (rather than modeled) and a unique equitable solution exists, as is the case here. As such, we don't explore those methods separately from these post-hoc adjustments through group-specific thresholding.

Composite Models: Following the proposal of Dwork and colleagues [25], we investigated two options for building composite models from models trained or selected for their performance on subgroups. On the one extreme, we simply used the grid of models trained on the full population but performed model selection separately for each subgroup (reflecting the complete transfer learning approach described by Dwork). On the other extreme, we trained separate models just with examples from each subgroup (the fully decoupled approach in Dwork) and added these to the model grid for subgroup-specific model selection. One challenge with implementing these composite models, however, is that the scores from the separate models chosen for different subgroups have not been calibrated and cannot be assumed to be comparable. As such, one needs to determine how to appropriately choose a total "top k" set of individuals across these different models. Because we were making use of these composite models in the interest of improving fairness, a natural means of choosing these thresholds was to apply the same method choosing TPR-equalizing thresholds described above. It is somewhat challenging to determine whether fairness improvements seen from these composite strategies are more a result of the group-specific thresholds or decoupling the model building or selection itself. However, one hope here would be that the decoupling should improve the accuracy of model predictions on the subgroups, so success for these methods ideally would show not just similar disparity mitigation to post-hoc adjustments but also improved overall accuracy metrics at the same level of fairness.

[2]For instance, if a program can intervene on 100 individuals, this process might break that down into 75 Black individuals and 25 white individuals. Because score distributions are likely to change over time, group-specific "top k" values are used rather than score thresholds to ensure the total number of targeted individuals remains fixed.

Model Selection: As noted above, an additional simple approach that falls under our umbrella of post-processing strategies is to account for fairness metrics in the process of model selection. However, this approach is not only very sensitive to the machine learning method/hyperparameter grid explored but also relies on some degree of luck that specifications with favorable trade-offs will be found. Here, we explored two options by which fairness could be included in the model selection process:

- Setting a "Disparity Constraint" reflecting a largest acceptable disparity. Here, we only consider models with disparity no higher than a certain value, then choose the model with the highest precision among these. Note that it may be possible that no models have a low enough disparity to meet the criteria, in which case we choose the model closest to this cut-off (making it a soft constraint and guaranteeing a model will always be chosen).

- Setting an "Accuracy Constraint" reflecting a largest acceptable loss in accuracy to improve fairness. Here, we only consider models with precision@k within a given number of percentage points below the best model, then choose the model with lowest disparity among these. Note that because this constraint is relative to the performance of the most-accurate model, there will always be at least one meeting the criteria, so this is a hard constraint.

For each type, we explored eight levels of the constraint, from placing little or no weight on fairness to strongly selecting for fair models. For Disparity Constraints, these included allowing disparities up to 5.0, 2.0, 1.5, 1.3, 1.2, 1.1, 1.05, or 1.0 (that is, exact equity). For the Accuracy Constraints, these included allowing a decrease in precision of up to 0.0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.50, and 0.60 percentage points.

3.2 Problems, Data, and Experimental Setup

Our empirical evaluation of these methods was done on three real world problems that we have worked on in collaboration with various government agencies. These span mental health and criminal justice (with Johnson County, Kansas), housing safety inspections (with San Jose, CA), and education outcomes (with the Education Ministry of El Savador). Since the data for these problems is confidential and not available publicly, we also replicate this empirical evaluation on a crowdfunding problem from DonorsChoose[3] where the data is publicly available. This will allow other researchers and practitioners to replicate our work before applying it to their own problems. In general, these problem settings involve six elements:

1. **Features**: Each project we use in this study went through an extensive feature engineering process. As is typically done in real-world ML systems, the features generated included raw and transformed information about the entities of interest (such as demographics) as well as temporal and spatial aggregations (while respecting temporal boundaries in train and validation sets to avoid leakage).

[3]http://wwww.donorschoose.org

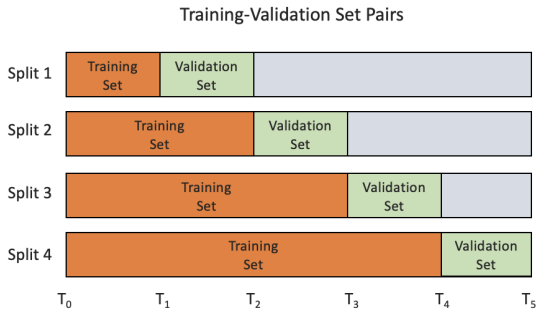


Figure 1: The temporal validation approach used in these settings to capture the non-stationary nature of the data and guard against leakage. Time is used to split the available data into a series of training and validation sets, testing for generalization performance on "future" data relative to model training.

2. **Label**: In each of the problem domains, the decision on the definition of the label is part of the formulation process and is done in collaboration with the partnering organization. In all of these problems, the label was determined by the occurrence of an event at some point in the future from the time of prediction, for example, an individual being booked into jail in the next 12 months or a crowdfunding project failing to get fully funded in the next 4 months.

3. **Train and Validation Splits**: Since most real-world prediction problems are temporal in nature and violate stationary distribution assumptions, we use temporal validation to split our datasets into train and validation sets [38]. These train and validation sets are usually temporally sequential in nature, where each candidate model is trained on data from "past" data and validated on "future" data (see Figure 1 for a diagram).

4. **Models**: We train a wide variety of model and hyperparameter combinations, including logistic regression, tree-based models, and ensembles such as random forests and boosted trees. The reasoning behind a wide grid was both to understand the effectiveness of different models along both the "accuracy" and bias dimensions as well as to provide the model selection process with as much diversity as possible. The model types and hyperparameters used for each problem are listed in Table 2.

5. **Choice of Bias Metric**: In all of these problems, a key decision to make is the choice of the appropriate bias metric(s). We use the Fairness Tree (Figure 2), a framework developed and used in [65] to inform that choice. Since in all the problems we describe below, we are supporting assistive interventions (i.e. reducing disparities in false negatives is more important than those in false positives), and have limited resources to intervene compared to the number of people that need support, the Fairness Tree framework leads us to choose Recall (True Positive Rate) Disparity as the primary bias metric.

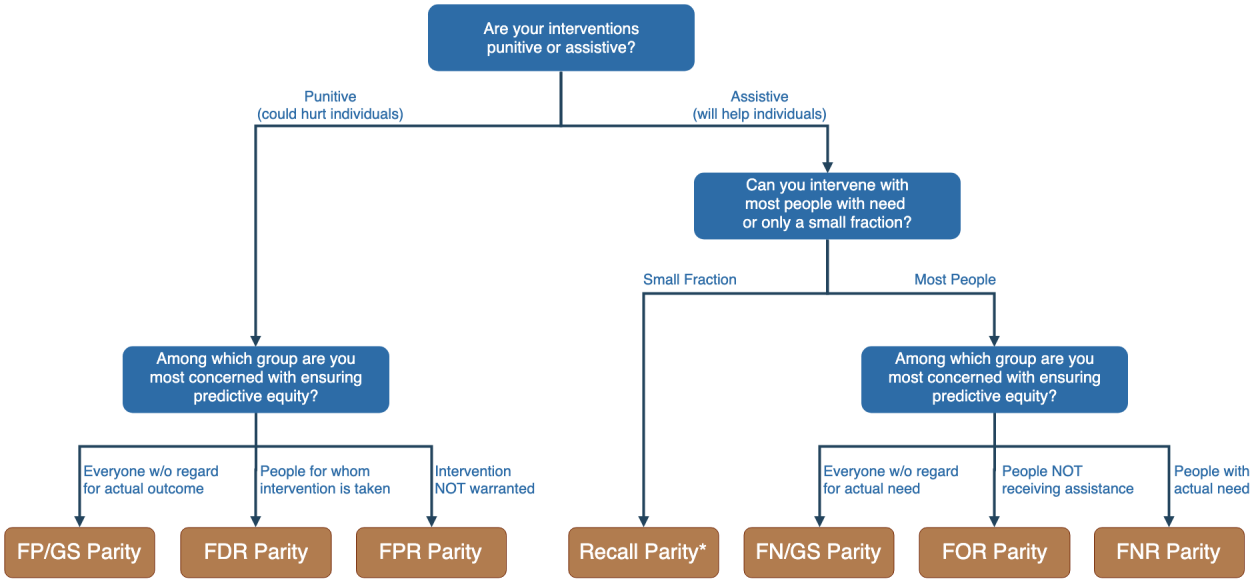


Figure 2: Fairness Tree framework to help identify appropriate fairness metrics based on the intended use. The metrics in the leaf nodes are: False Negative Rate (FNR), False Omission Rate (FOR), False Negatives Adjusted to Group Size (FN/GS), Recall/True Positive Rate (TPR), False Positive Rate (FPR), False Discovery Rate (FDR), and False Positives Adjusted to Group Size (FP/GS).

6. **Evaluation Methodology**: For each temporal validation set we calculate the evaluation metric as well as the bias metric (with respect to the protected group) for all models. These results are aggregated by calculating the mean and standard errors.

Each of the four policy problems used for the present empirical evaluation are described in detail below, including details about the underlying data set, the performance and fairness metrics of interest, and the protected group for bias and fairness analysis.

3.2.1 *Mental Health Outreach - Johnson County KS*

Untreated mental health conditions often result in a negative spiral, which can culminate in repeated periods of incarceration with long term consequences both for the affected individual and the community as a whole [30]. Surveys of inmate populations have suggested a high prevalence of multiple and complex needs, with 64% of people in local jails suffering from mental health issues and 55% meeting criteria for substance abuse or dependence [40]. The criminal justice system is poorly suited to address these needs, yet houses three times as many individuals with serious mental illness as hospitals [27].

Since 2016, Johnson County, KS, has partnered with our group to help them break this cycle of incarceration by identifying individuals who might benefit from outreach with mental health resources and are at risk for future incarceration. While the Johnson County Mental Health Center (JCMHC) currently provides services to the jail population, needs are generally identified reactively, for instance through screening instruments individuals fill out when entering jail. The new program being developed will supplement these existing approaches by adding a new automatic referral system for people who are at risk of being booked into jail, with the hope that they can be outreached to reduce their risk of returning to jail.

Through our partnership, we obtained administrative data from their mental health center, jail system, police arrests, and ambulance runs. ML modeling was focused on Johnson County residents with any history of mental health need who had been released from jail within the past three years. Early results from this work were described in [9]. A field evaluation of the predictive model is ongoing at the time of this writing, but validation on historical data demonstrated a 12% improvement over a baseline based on the number of bookings in the prior year and 4.8-fold increase over the population prevalence.

3.2.2 *Housing Safety Inspections - San Jose, CA*

The Multiple Housing team in San Jose's Code Enforcement Office is tasked with protecting the occupants of properties with three or more units, such as apartment buildings, fraternities, sororities, and hotels. They do so by conducting routine inspections of these properties, looking for everything from blight and pest infestations to faulty construction and fire hazards (see [35] and [45] for a discussion of the importance of housing inspections to public health). Although the city of San Jose inspects all of the properties on its Multiple Housing roster over time, and expects to find minor violations at many of them, it is important that they can identify and mitigate dangerous situations early to prevent accidents. With more than 4,500 multiple housing properties in San Jose, CA – many of which comprise multiple buildings and hundreds of units – it is not possible for the city to inspect every unit every year. San Jose recently instituted a tiered approach to prioritizing inspections, inspecting riskier properties more frequently and thoroughly. Although the tier system helped focus inspections on riskier

Table 2: Data and Experimental Setup for our four problems.

	Mental Health and Criminal Justice	**Housing Safety Inspections**	**Student Outcomes**	**Education Crowdfunding**
Prediction Task	Jail booking within the next 12 months	Housing unit having a violation within the next year	Student not returning to school next year	Project not getting fully funded within 4 months
Timespan	2013-01-01 to 2019-04-01	2011-01-01 to 2017-06-01	2009-01-01 to 2018-01-01	2010-01-01 to 2014-01-01
# of entities	61,192	4,593	801,242	210,310
Feature Groups	Demographics Mental Health History Past Diagnosis Mental Health Programs Police Interactions Past Jail Incarceration Jail Booking Details	Building Permits Past Citations Past Violations House Prices Census Data	Age Relative to Grade Repeated Grades Rural/Urban Academic History Dropout History Gender Illness Family Information	Funding Request Details Donation Details Past Funding Rates Project Description
# of Features	3,465	1,657	220	319
Base Rate	0.12	0.43	0.25	0.24
Evaluation Metric	Precision@500	Precision@500	Precision@10000	Precision@1000

Model Types and Hyperparameters (as specified by scikitlearn parameters used in the experiments)	`Decision Tree`
	Max Depth:
	(1,2,3)
	Min Samples Split:
	(10, 50, 100)
	
	`Random Forest`
	Num Estimators:
	(100, 1000, 5000)
	Min Samples Split:
	(10, 25, 100)
	Max Depth:
	(5, 10, 50)
	
	`Logistic Regression`
	Penalty:
	(l1, l2)
	C:
	(0.001, 0.01, 0.1, 1, 10)
		`Decision Tree`
		Criteria:
		(gini, entropy)
		Max Depth:
		(1,2,3,5,10,20,50)
		Min Samples Split:
		(10, 20, 50, 1000)
		
		`Random Forest`
		Max Features:
		(sqrt, log2)
		Criteria:
		(gini, entropy)
		Num Estimators:
		(100, 1000, 5000)
		Min Samples Split:
		(10, 20, 50, 100)
		Max Depth:
		(2, 5, 10, 20, 50, 100)
		
		`Extra Trees`
		Max Features:
		(sqrt, log2)
		Criterion:
		(gini, entropy)
		Num Estimators:
		(100, 1000, 5000)
		Min Samples Split:
		(10, 20, 50, 100)
		Max Depth:
		(2, 5, 10, 50, 100)
		
		`Logistic Regression`
		Penalty:
		(l1, l2)
		C:
		(0.001, 0.01, 0.1, 1, 10)
			`Decision Tree`
			Max Depth:
			(1, 5, 10, 20, 50, 100)
			Min Samples Split:
			(2, 5, 10, 100, 1000)
			
			`Extra Trees`
			Num Estimators: (100)
			Max Depth: (5, 50)
			
			`Logistic Regression`
			Penalty:
			(l1, l2)
			C:
			(0.0001, 0.001, 0.1, 1, 10)
			
			`Random Forest`
			Num Estimators: (100, 500)
			Min Samples Split: (2, 10)
			Class Weight:
			(Balanced Subsample, Balanced)
			Max Depth: (5, 50)
				`Random Forest`
				Num Estimators:
				(100, 500, 1000)
				Min Samples Split:
				(10, 50)
				Max Depth:
				(10, 50, 100)
				
				`AdaBoost`
				Num Estimators: (500, 1000)
				`Decision Tree`
				Max Depth:
				(1, 5, 10, 20, 50, 100)
				Min Samples Split:
				(2, 5, 10, 100, 1000)
				
				`Logisitic Regression`
				C:
				(0.0001, 0.001, 0.01, 0.1, 1, 10)
				penalty:
				(l1, l2)

Train and Validation Sets	Temporal Block: 4 months	Temporal Block: 2 months	Temporal Block: 1 year	Temporal Block: 3 months
Protected Group	Race	Median Income	Age Relative to Grade	Poverty Level

properties, the new system has its limitations. The city evaluates tier assignments for properties infrequently (every 3 to 6 years), and these adjustments require a great deal of expertise and manual work while leaving out a rich amount of information.

In order to provide a more nuanced view of properties' violation risk over time and allow for more efficient scheduling of inspections, the Code Enforcement Office partnered with us to develop a model to predict the risk that a serious violation would be found if a given property was prioritized for inspection (similar tools have been developed for allocating fire inspections in New York [7] and health inspections in Boston [28]). Evaluation of the model on historical data indicated that it could provide a 30% increase in precision relative to the current tier system and the model's predictive accuracy was confirmed during a 4-month field trial in 2017.

3.2.3 *Improving Educational Outcomes - El Salvador*

Each year from 2010 through 2016, 15-29% of students enrolled in school in El Salvador did not return to school in the following year. This high dropout rate is cause for serious concern, with significant consequences for economic productivity, workforce skill, inclusiveness of growth, social cohesion, and increasing youth risks [11, 8]. El Salvador's Ministry of Education has programs available to support students with the goal of reducing these high dropout rates, but the budget for these programs is not large enough to reach every student and school in El Salvador.

Predictive modeling has been deployed to help schools identify students at risk of dropping out in several contexts [49, 4, 14] and El Salvador partnered with us in 2018 to make use of these methods to focus their limited resources on the students at highest risk of not returning each year. Student-level data was provided by the Ministry of Education, including demographics, urbanicity, school-level resources (e.g., classrooms, computers, etc), gang and drug violence, family characteristics, attendance records, and grade repetition. For the present study, we focused on the state of San Salvador and identifying the 10,000 highest-risk students, considering annual cohorts of approximately 300,000 students and drawing on 5 years' of prior examples as training data.

3.2.4 *Education Crowdfunding - DonorsChoose*

Since the projects above used confidential and sensitive data and were done under data use agreements, we are not able to make that data publicly available. For our work to be easily reproducible, we include a fourth problem in this study where the data is available publicly, focused around crowdfunding for education by the organization DonorsChoose.

Many schools in the United States, particularly in poorer communities, face funding shortages [56]. Often, teachers themselves are left to fill this gap, purchasing supplies for their classrooms when they have the individual resources to do so [37]. The non-profit DonorsChoose was founded in 2000 to help alleviate these shortages by providing a platform where teachers post project requests focused on their classroom needs and community members can make contributions to support these projects. Since 2000, they have facilitated $970 million in donations to 40 million students in the United States [2]. However, approximately one third of all projects posted fail to reach their funding goal.

Here, we make use of a dataset DonorsChoose made publicly available for the 2014 KDD Cup (an annual data science competition) including information about projects, the schools posting them, and donations they received. Because the other case studies explored here focused on proprietary and often sensitive data shared with us under data use agreements that cannot be made publicly available, we included a case study surrounding this publicly-available dataset. While we have not partnered with DonorsChoose to deploy the machine learning system described, we otherwise treated this case study as we would any of our applied projects. Here, we consider a resource-constrained effort to assist projects at risk of going unfunded (for instance, providing a review and consultation) capable of helping 1,000 projects in a 2-month window, focusing on the most recent 2 years' of data available in the extract (earlier data had far fewer projects and instability in the baseline funding rates as the platform ramped up). This dataset is publicly available at kaggle.com [1].

4. RESULTS

Overall results across the different methods and problems we evaluated are shown in Figure 3. Each graph shows the relative performance of models with a given strategy in terms of the "performance metric" (namely precision@k on the x-axis) and fairness with respect to the protected group (namely True Positive Rate or Recall disparities on the y-axis), with error bars representing the 95% confidence interval over all temporal validation splits. The ideal model would have a value of 1.0 for both of these metrics – models appearing further to the right on the x-axis are more accurate while those appearing closer to the dashed y=1.0 line are more equitable (departures from this line in either direction reflect disparities favoring one or the other group). The blue circle in all of the graphs refers to the *Original* model — a term we use to specify the model built and selected only focused on maximizing the accuracy metric. All the other points are results from the bias reduction methods that we investigated. Note that in this figure we only include the best-performing sampling strategies (either in terms of fairness or accuracy) but discuss and show the wider range of sampling results in the graphs below. Likewise, we only show the composite models without decoupled training because the results from the two strategies were generally similar, but discuss and show these results in more detail below as well. Additionally, model selection approaches are omitted from Figure 3 because they generally required considerable decreases in precision@k to improve fairness, allowing us to focus the overall analysis on the nuance between the other methods (see Figure 7 and the related discussion for these results).

4.1 Overall Results

Across the four problems, a few general patterns seem to emerge from our experiments:

1. **Considerable disparities, ranging from 30-100%, were observed in the *baseline* models** for all four problems. That is, building models which optimize only for some measure of accuracy consistently resulted in appreciable biases if fairness was not actively pursued as an outcome. This is not a surprising outcome and a result that has been demonstrated in various

studies but an important point to keep in mind when building ML models.

2. There was **considerable variability across strategies and settings** in the effectiveness and ability of the fairness-enhancing methods considered here to remove these disparities, with **most methods showing only moderate success** or success only in a few settings. Comparisons across these methods is discussed in more detail below.

3. Only the two approaches which made use of **separate thresholds across groups (composite models and post-hoc adjustments) were consistently successful** in removing disparities and did so without any appreciable loss in model accuracy.

Below we discuss these results in more detail, examining the performance of each fairness-enhancing approach in turn.

4.2 Effect of Removing Sensitive Attribute

A common misconception in the context of algorithmic fairness is that simply omitting a sensitive attribute can help a model achieve fair predictions through "unawareness." Several authors [46, 59] have spoken to the fallacy of this concept, both as a result of correlations between protected attributes and other potentially relevant ones and because access to the sensitive attribute might help models pick up on patterns that improve accuracy for the protected group and result in lower disparities. However, we included this strategy here both for completeness as well as to understand and demonstrate how this approach might perform in practice. Unsurprisingly, then, the results in Figure 4 show **this strategy is inconsistent in the magnitude and direction of its impact across the four problems**. Although omitting the protected attribute did improve model fairness somewhat in the Education Crowdfunding and Student Outcomes contexts, in neither case did it fully remove the disparities from the initial model, and in the latter case these improvements came at the cost of a moderate decrease in precision@k. Moreover, in the Inmate Mental Health setting, removing the race attribute actually made the models somewhat less fair on average while in the Housing Safety context doing so had no effect on either fairness or accuracy. Taken together, these results are very consistent with the notion that "fairness through unawareness" by **removing the sensitive feature cannot be relied upon to improve the fairness of machine learning models**.

4.3 Effect of Sampling

The other pre-processing method we explored involved sampling of the training data. As discussed above, a number of parameters must be determined in choosing a sampling strategy: the relative distributions of the protected and non-protected subgroups, the label distributions within each group, and whether to over- or under-sample training examples to achieve the target distribution. Here, we explored six strategies (Table 1) that reflect combinations of three reasonable hypothesis:

- A 1:1 ratio between protected group and non-protected group training examples might tell the model to treat errors in each group as equally important, alleviating differential error rates.

- Equal label distributions within the two subgroups might tell the model to not treat protected group membership as important.

- A 50/50 label distribution in one or both subgroup might alleviate any issues arising from imbalance in the training set.

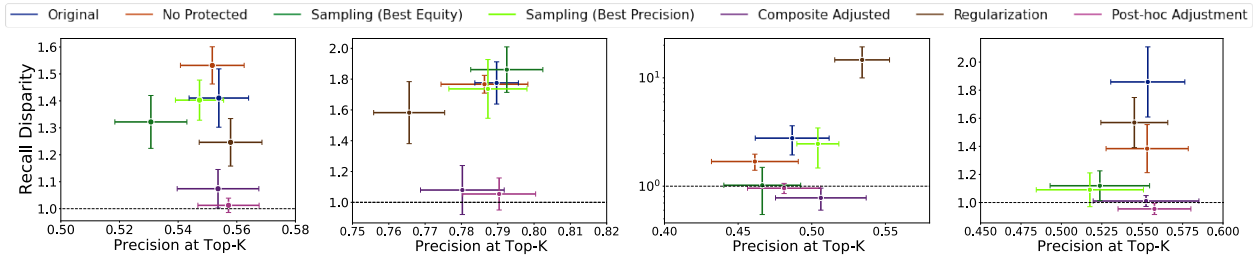
Figure 5 shows the results of applying these six strategies to the training data in each of the four policy settings. Although resampling of the training data had an impact on the models in many of the problem settings, there was a considerable inconsistency in the results both across settings and sampling strategies. In the Education Crowdfunding and Student Outcomes settings, many (but not all) of the strategies showed improvements in model fairness, while none of the strategies yielded fair results in the Housing Safety or Inmate Mental Health settings. Interestingly, this pattern reflects the results observed when removing the protected attribute described above, suggesting that both strategies may be accomplishing the same thing by effectively telling the model not to treat subgroup membership as important. Note, in particular, that in the Education Crowdfunding setting, sampling strategies 2, 3, 5, and 6 show improvements and each of these strategies equalizes the label distribution across the protected and non-protected subgroups.

Two more general patterns in Figure 5 do seem of note: First, over- and under-sampling approaches to the sample sampling strategy appear to yield similar results, suggesting that, at least in these four policy contexts, decreasing the total number of training examples through undersampling did not have an appreciable impact on model performance. Second, strategy 4 yielded particularly variable results, ranging from little impact to large disparities in either direction (note that in the Education Crowdfunding setting, both over- and under-sampling for strategy 4 resulted in no predicted positives in the protected class, yielding infinite disparities, so this strategy is omitted from the graph). However, this result might not be too surprising in light of the fact that this is the only strategy considered here where we adjusted the label distribution among the protected subgroup (to 50/50) without changing the distribution of non-protected subgroup. Depending on the baseline distribution, of course, this might (or might not) tell the model to see the protected attribute (or correlated features) as particularly important as a predictor of the outcome.

Taken together, these results suggest that **sampling of the training data can have an impact on disparities in the resulting models' predictions, but that these effects are both context and parameter dependent**. Without an obvious or consistent pattern for how a given sampling strategy will translate into changes in fairness metrics in a given modeling context, model developers are left to conduct a search over different values of these sampling parameters to explore this space in their setting. Even so, **there does not seem to be strong empirical evidence that a fairness-enhancing solution will be found in a particular context, or at what cost to model accuracy**.

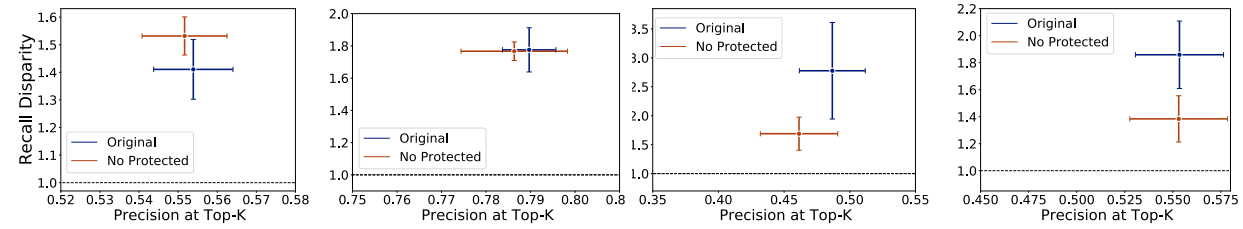
4.4 Effect of In-Processing Methods

The in-processing method considered here was proposed by Zafar and colleagues [75]. Models were trained with a con-



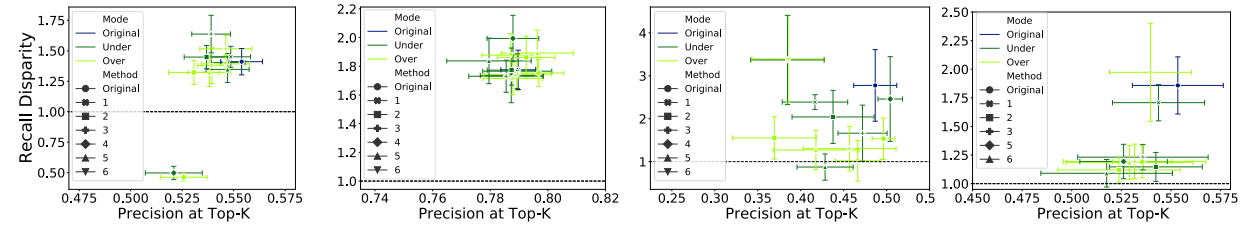
(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 3: Results from the different fairness-enhancing strategies considered here across the four policy settings, showing the relationship between model accuracy (as measured by precision@k) on the x-axis and fairness (as measured by recall disparities) on the y-axis. Ideal models would have high values of precision@k and be near a disparity value of 1.0. Note that the y-axis in (c) is on a log scale based on the large variation in performance across methods (performance for each method is shown separately on a linear scale in the figures below). Error bars show 95% confidence intervals over validation sets.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 4: Effect of removing the protected attribute from machine learning modeling on model accuracy (precision@k) and fairness (recall disparities). Error bars show 95% confidence intervals over validation sets.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 5: Results from resampling of training data for machine learning on model accuracy (precision@k) and fairness (recall disparities). Each of the six strategies from Table 1 was performed either via under-sampling (dark green) or over-sampling (light green). Error bars show 95% confidence intervals over validation sets.

straint to equalize the false negative rate[4] between the protected and non-protected group in each setting, with results shown in Figure 6. In general, in-processing failed to appreciably improve the fairness of the models in any of the four settings, reducing disparities only slightly in three settings and making them appreciably worse in the fourth (Student Outcomes).

Importantly, these results should not be seen as an inherent critique of either Zafar's method specifically or in-processing in general, but rather as a mismatch between the currently available methods using this approach and the common context of resource-constrained problem settings in which a given number of highest-risk entities must be selected for an intervention. In-processing methods generally add a fairness constraint to a classifier that optimizes for overall accuracy around an implicit threshold of 0.5 (or best-partitioning decision boundary). To select the "top k" for intervention, we naively threshold the resulting score (or, equivalently, shift the decision boundary) to yield only k highest-risk predicted positives. Of course, the fairness constraints used during model training applied to the original boundary, not the shifted one. As such, it is not surprising that Zafar's method here failed to improve fairness of these models when applied to a "top k" setting, even if it might perform well in settings without such a constraint. Perhaps for this reason, other methods such as Microsoft's Fair Learn [13] only provide predicted class labels without a continuous score, but unfortunately those methods are also poorly suited to the "top k" setting where a small subset of k individuals would need to be randomly chosen from the predicted positive class at considerable cost to accuracy/precision.[5]

4.5 Effect of Model Selection

Results of applying fairness-aware model selection in these contexts are shown in Figure 7. Here, several of the settings suggest an often considerable trade-off between fairness and accuracy, with constraints that put more weight on fairness in the model selection process yielding sizable decreases in precision@k (note that the range of the x-axes for these graphs is generally much wider than for the results of using other methods). For instance, in the Education Crowdfunding setting, disparities could be removed through the model selection process, but at the expense of loosing nearly half of the model's precision. In other cases, even large fairness constraints in the model selection process failed to remove disparities effectively: even when sacrificing a large amount of precision in the Inmate Mental Health context, the resulting models still showed considerable disparities of 1.27 on average. Likewise, in the Housing Safety context, fairness-aware model selection failed to reduce the disparities in these models regardless of constraint type or size. Notably, across all four contexts, similar results could be obtained by placing either a soft constraint on the largest acceptable disparity or a hard constraint on the largest acceptable decrease in precision@k to improve fairness (represented by different colors in Figure 7).

Although on the surface, these results suggest some semblance of the "Pareto Frontier" one might anticipate could reflect an inherent trade-off between fairness and accuracy, it is important to keep in mind that the nature of this frontier is highly dependent on the model grid over which this selection process is taking place (that is, other model type/hyperparameter combinations may perform better on one or both metrics). Likewise, other approaches at improving model fairness (such as the other methods explored here) may expand this frontier and allow for considerably less drastic trade-offs between fairness and accuracy.

4.6 Effect of Post-Hoc Adjustments

Figure 8 shows the results of post-hoc adjustments to equalize TPR across groups by choosing separate, group-specific thresholds. Across all four policy settings, this approach consistently improved the fairness of the models, entirely removing the disparities in most cases. Notably, this improved fairness was achieved with negligible cost in terms of model accuracy in all four settings. While this lack of fairness-accuracy trade-off is somewhat surprising on its face, the "top k" setting likely plays a role here as well. With limited resources relative to needs, there are many ways to choose k individuals for intervention with equally high precision, making it possible to swap some high-risk individuals from one group with those from another in order to improve fairness without appreciably reducing accuracy. To the extent that any small trade-offs may exist when making these adjustments, they seem to be dominated by variation over time in the generalization performance of the models, yielding consistently fair adjusted models without sacrificing accuracy (for a more detailed discussion of the lack of trade-offs with this approach, see our recent work in [64]).

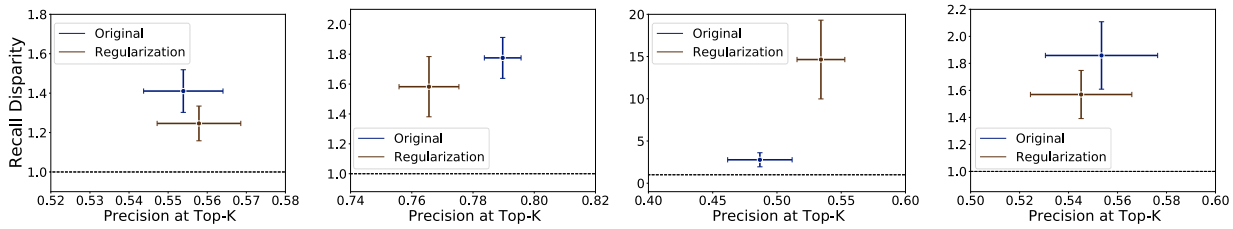
4.7 Effect of Composite Models

The final approach explored here follows Dwork's proposal [25] to build composite models, either through separate model selection or fully decoupled training for each subgroup. Figure 9 presents the results of these two strategies in each of the policy settings. In general, we find these approaches to perform quite well, consistently reducing the disparities across all four settings. As noted above, because the uncalibrated scores of these group-specific models cannot be assumed to be comparable, we combined the models across groups by making use of the same process of choosing TPR-equalizing thresholds as we used to make post-hoc adjustments to single models. As such, the fairness improvements seen here might either be a result of the composite strategy itself or the method for choosing thresholds, which, as seen above was itself very successful in reducing disparities here. However, if selecting (or training) separate models was appreciably improving model performance for the subgroups, we might hope to see increases in the overall accuracy of the composite models relative to the post-hoc adjusted ones in Figure 8, but the results here do not lend evidence to support this hypothesis. While there may be a slight improvement in precision@k for the composite model in the Student Outcomes setting, the difference in that setting is far from statistically significant and accuracy of the composite models in other settings is nearly identical to or somewhat lower than that of the post-hoc adjusted models.

To disentangle the effects of the composite modeling strategy itself from the TPR-equalizing group-specific thresholds used here, other strategies for choosing and combining the models across subgroups could be explored, although these

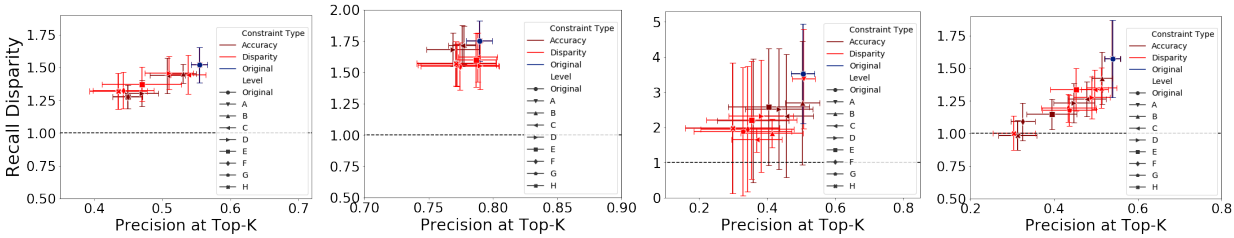
[4]Note that $FNR = 1 - TPR$, so this constraint is equivalent equalizing TPR across groups.

[5]We explored this package in particular in the Education Crowdfunding setting in a recent tutorial presented at the 2020 KDD and 2021 AAAI conferences [68].



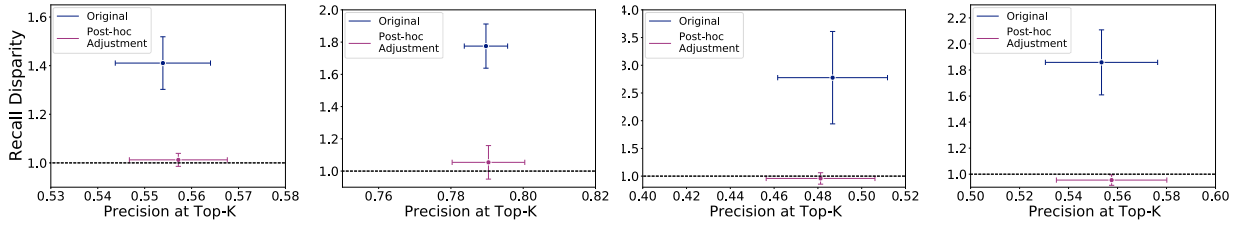
(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 6: Results of using the in-processing method proposed by Zafar and colleagues to perform fairness-constrained optimization during model training on model accuracy (precision@k) and fairness (recall disparities). Error bars show 95% confidence intervals over validation sets.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 7: Effect of fairness-aware model selection on model accuracy (precision@k) and fairness (recall disparities). Model selection was performed either by setting a maximum acceptable disparity and choosing the model with the best precision@k among these (Disparity Constraint) or setting a maximum acceptable decrease in precision@k and choosing the lowest-disparity model among these (Accuracy Constraint). For each type, eight levels of constraint were explored (labeled A-H in the figure, from least to most weight on fairness). For Disparity Constraints, these are: A: 5.0, B: 2.0, C: 1.5, D: 1.3, E: 1.2, F: 1.1, G: 1.05, H: 1.0; for Accuracy Constraints, these are: A: 0.0, B: 0.05, C: 0.10, D: 0.15, E: 0.2, F: 0.25, G: 0.5, H: 0.6. Error bars show 95% confidence intervals over validation sets.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 8: Effect of post-hoc adjustments on model accuracy (precision@k) and fairness (recall disparities). Separate score thresholds were chosen for the protected and non-protected subgroups to equalize recall across the groups. Error bars show 95% confidence intervals over validation sets.

are complicated somewhat by the requirement of the "top k" setting here that a total of k entities is selected across groups. The score threshold yielding the desired number of entities will vary considerably across pairs of models, so the appropriate cut-off at which to evaluate model performance for one subgroup depends on what models it will be combined with for other subgroups. As a preliminary experiment, we explored a simplified strategy in which we selected models for each subgroup based on their performance among the same number of highest-risk individuals that would have been selected from a single (non-composite) model. These group-specific models were combined and then the top k individuals with the highest scores in the composite model were chosen with a single score threshold.[6] In these initial experiments, composite models with a single score threshold failed to improve on either the accuracy or fairness of the original models, lending support to the conclusion that the improvements observed in Figure 9 are likely driven by the TPR-equalizing thresholds used to combine the models across subgroups.

In most of the problem settings considered here, the composite models with and without fully-decoupled training performed similarly, but the Housing Safety context provides a notable exception (Figure 9(c)). Although the composite approach performs well in this setting, the decoupled strategy shows a considerable loss in precision as well as overshooting the necessary adjustment to achieve a fair result (ending up with bias in the opposite direction). Notably, the Housing Safety dataset is considerably smaller than the others used here, with an order of magnitude fewer entities than the next-largest setting. As observed in [25], one potential disadvantage to decoupled model training is that the smaller number of training examples might degrade model performance, particularly if there are common patterns in the data that could be learned across groups. We would, of course, expect this issue to be exacerbated as the overall number of available examples decreases. Likewise, performing model selection on relatively small subgroups might be prone to over-fitting, choosing an overly-optimistic model specification whose performance reverts to a lower mean when measuring generalization performance on a future validation set. Such over-optimistic performance estimates for one subgroup could also affect the recall-balancing thresholds chosen across groups, leading to relatively too many individuals being chosen from one group and yielding disparities in the final composite model as well.

5. DISCUSSION

The goal of the current work was to build on the extensive recent work in algorithmic fairness by comparing how the wide variety of proposed fairness-enhancing approaches and methods perform in the context of real-world problems in high-stakes policy problems. While our aim was not to comprehensively include every existing approach, we sought to sample a wide range of techniques applied at different phases of the machine learning pipeline by pre-processing of the input data, in-processing during model training, and post-processing of trained models. Similarly, we focus here on resource-constrained assistive policy contexts where the optimization problem reflects a "top k" setting and we argue TPR disparities are an appropriate fairness metric (reflecting a concept of "equality of opportunity" as discussed in [32, 65]). While some of the results described here might not generalize beyond this problem setting, we note that it is very commonly encountered in high-stakes decisions across education [4, 49], healthcare [60], criminal justice [34], social services [10], as well as many other contexts [48, 73, 22, 7, 28], and has been the most common formulation encountered in the more than 100 projects we have been involved in applying machine learning to social good problems with government and non-profit partners.

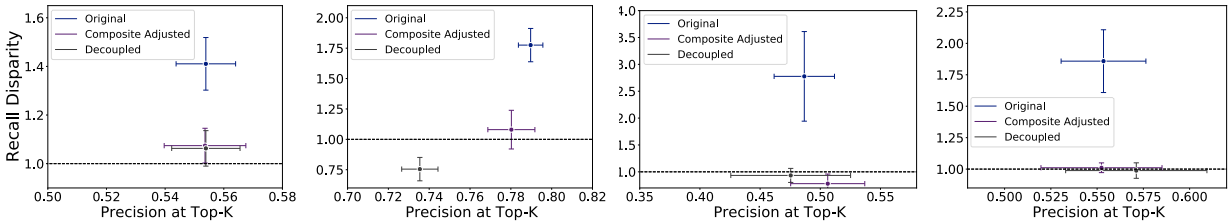
In this setting, **pre-processing methods showed decidedly mixed and inconsistent results**, with both sampling and omitting the protected attribute improving fairness in some contexts but not others. This inconsistency is perhaps not entirely surprising given the wide range of potential contributors to disparities at any stage of the machine learning pipeline [63], only some of which we might expect these pre-processing methods to address. Unfortunately, it seems unclear *a priori* whether these strategies will be effective in a given context (or, with sampling, what approach will work), making them unreliable as a fairness-enhancing approach.

Similarly, **we found little success with removing disparities through in-processing**, but, as noted above, existing methods to add fairness constraints in the process of model training seem particularly poorly suited to the "top k" setting. In principle, developing new in-processing methods better suited to the "top k" setting should be feasible, but poses particular technical challenges. As other work developing methods for this setting (without fairness constraints) has observed, the loss function in this setting is, in general, not only non-convex, but discontinuous (as disjoint regions of the parameter space yielding exactly k predicted positives must be connected by regions yielding either more or fewer than k) [51]. To our knowledge, no methods presently exist that seek to improve fairness through in-processing for "top k" models, but we believe this could be an interesting future research direction.

By contrast, we found **consistent success across the four problems with eliminating disparities using post-hoc methods.** Across all four policy settings considered here, these improvements in model fairness could be accomplished without a corresponding trade-off in accuracy. Although such trade-offs are often assumed to be an inherent aspect of reducing disparities in machine learning models [26, 76, 16] making this result somewhat surprising, the resource-constrained nature of these policy settings may contribute to the lack of trade-off as discussed above. Further, the consistency with which fair predictions could be obtained without cost to accuracy across the settings considered here may have important implications for policymakers and machine learning practitioners, reinforcing the moral imperative to ensure the fairness of models deployed in similar contexts (see our recent work in [64] for a more detailed discussion of these policy implications).

Given the success of these post-hoc adjustments across models and settings, we also investigated whether applying these adjustments on top of the pre-processing and in-processing strategies explored here could remove any residual (or newly

[6]Note that this approach will likely yield a different number of individuals in each group than was in process of selecting the group-specific models, so the assumption being made here is that these differences will be small enough that the model performance among each subgroup will not depart appreciably from what was used during selection.



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Figure 9: Effect of composite model approaches on model accuracy (precision@k) and fairness (recall disparities). Composite models were developed by choosing separate models for each subgroup either with or without decoupled training. Error bars show 95% confidence intervals over validation sets.

introduced) disparities from those methods. Consistently, post-hoc adjustment by choosing thresholds to equalize TPR yielded more fair results, even when applied in combination with other strategies that failed to improve fairness themselves. While this result suggests a robustness of this strategy, we did not observe any improvement in model performance by combining post-hoc adjustments with other strategies, so we do not see any advantage to doing so in practice. Finally, we should note the importance of considering the broader context in which a machine learning model will be applied. While the work here has focused on improving the fairness of a model's predictions, doing so is only one step in the process of ensuring outcomes of the broader sociotechnical system are themselves equitable. In most policy contexts, these models are deployed in a manner intended to inform the decision-making process of a human expert such as a doctor, case worker, or school administrator, rather than being fully autonomous. As such, fairness in a model's recommendations is not necessarily a guarantee that interventions will be allocated fairly, depending on how and when these humans in the loop follow or override them. Further, the interventions themselves may not be equally effective for everyone. For instance, additional after-school tutoring might be difficult to attend for students who have work or family obligations in the afternoons, or programs offered only in English might not effectively serve individuals for whom it is not their first language. Likewise, when the labels themselves are measured in inaccurate and disparate ways, such as using arrests as a proxy for crime commission [5, 21, 53, 47, 31], measures of fairness that take these labels as "ground truth" will fail to capture these underlying disparities. Understanding the implications for fairness at each stage of the process — from label definition through modeling to decisions and interventions — is essential to understanding and mitigating biases in deployed machine learning systems that impact people's lives. The work here explores one key aspect of this process, but machine learning practitioners and the policymakers who deploy and act on the systems they build must be cognizant of these broader contextual aspects as well.

6. SUMMARY AND FUTURE WORK

In the present study, we explored the performance of several proposed fairness-enhancing methods on reducing bias and enhancing fairness in general and improving equality of opportunity (as measured by TPR disparities) in particular across four real-world policy contexts. Among the methods considered, we found that post-hoc adjustments to model scores by choosing TPR-equalizing group-specific score thresholds was capable of removing disparities without loss of accuracy in all four settings. Most directly, our results have implications for practitioners building and deploying machine learning systems in similar resource-constrained policy contexts for whom this post-hoc approach should be both straightforward to implement and likely to improve the fairness of their models. For the machine learning research community, we believe this work highlights the importance of evaluating new methods on real-world problems, in particular demonstrating a gap with how well-suited current in-processing methods are to this "top k" setting.

Although we focus here on characteristics of machine learning problems commonly encountered in high-stakes policy contexts, it will be important extend this work to other policy settings, particularly those for which other bias metrics beyond TPR disparity are of interest. In particular, we hope to understand whether the consistent improvements of the post-hoc adjustments employed here will generalize to other fairness metrics, especially those which are not guaranteed to be monotonically increasing or decreasing with the model score. Additionally, in all the settings considered here, the sensitive attribute was known exactly, but this is not always the case. Unfortunately, many of the approaches considered in this study (such as sampling, composite models, and post-hoc adjustment) cannot be directly applied where there is uncertainty around group membership, and more work will be required to both extend these methods to those contexts as well investigate the performance of methods that are inherently better-suited to them (such as those described in [18, 55]). Finally, although we sought to sample a range of fairness-enhancing methods across pre-, in-, and post-processing approaches, many more methods have been proposed than we could incorporate in the present work and continuing to extend upon these findings with additional methods will be an interesting avenue for future work.

7. ACKNOWLEDGEMENTS

This project was partially funded by the C3.AI Digital Transformations Institute. We would also like to thank the Data Science for Social Good Fellowship fellows, project partners, and funders as well as our colleagues at the Center for Data Science and Public Policy at University of Chicago for the initial work on projects that were extended and used in this study.

8. REFERENCES

[1] Data on donorschoose. https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose/data. Accessed: 2020-06-23.

[2] Statistics on donorschoose. https://www.donorschoose.org/about. Accessed: 2020-06-23.

[3] S. Aghaei, M. J. Azizi, and P. Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426, 2019.

[4] E. Aguiar, H. Lakkaraju, N. Bhanpuri, D. Miller, B. Yuhas, and K. L. Addison. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, LAK '15, page 93–102, New York, NY, USA, 2015. Association for Computing Machinery.

[5] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine Bias. Technical report, ProPublica, 5 2016.

[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23, 2016.

[7] S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[8] M. N. Atwell, R. Balfanz, J. Bridgeland, and E. Ingram. Building a grad nation: Progress and challenge in raising high school graduation rates. annual update 2019. *Civic*, 2019.

[9] M. J. Bauman, K. S. Boxer, T.-Y. Lin, E. Salomon, H. Naveed, L. Haynes, J. Walsh, J. Helsby, S. Yoder, R. Sullivan, et al. Reducing incarceration through prioritized interventions. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–8, 2018.

[10] M. J. Bauman, R. Sullivan, C. Schneweis, R. Ghani, K. S. Boxer, T.-Y. Lin, E. Salomon, H. Naveed, L. Haynes, J. Walsh, J. Helsby, and S. Yoder. Reducing Incarceration through Prioritized Interventions. In *Proceedings of the Conference on Computing and Sustainable Societies (COMPASS)*, pages 1–8, New York, New York, USA, 2018. ACM.

[11] C. R. Belfield and H. M. Levin. *The price we pay: Economic and social consequences of inadequate education*. Brookings Institution Press, 2007.

[12] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[13] S. Bird, M. Dudık, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *URL https://www. microsoft. com/en-us/research/uploads/prod/2020/05/Fairlearn_whitepaper. pdf*, 2020.

[14] A. J. Bowers, R. Sprott, and S. A. Taff. Do we know who will drop out? a review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *The High School Journal*, 96(2):77–100, 2012.

[15] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[16] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3995–4004, 2017.

[17] S. Caton and C. Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[18] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[20] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[21] A. Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, 6 2017.

[22] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR, 2018.

[23] D. Dua and C. Graff. UCI machine learning repository, 2017.

[24] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[25] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133. PMLR, 2018.

[26] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[27] E. Fuller Torrey, A. D. Kennard, D. Eslinger, R. Lamb, and J. Pavle. More Mentally Ill Persons Are in Jails and Prisons Than Hospitals: A Survey of the States. Technical report, Treatment Advocacy Center and National Sheriffs' Association, 2010.

[28] E. L. Glaeser, A. Hillis, S. D. Kominers, and M. Luca. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review*, 106(5):114–18, 2016.

[29] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7):1445–1459, 2012.

[30] M. Hamilton. People with complex needs and the criminal justice system. *Current Issues in Criminal Justice*, 22(2):307–324, 2010.

[31] B. E. Harcourt. Risk as a Proxy for Race: The Dangers of Risk Assessment. *Federal Sentencing Reporter*, 27(4):237–243, 2015.

[32] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

[33] H. Heidari, C. Ferrari, K. P. Gummadi, and A. Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *arXiv preprint arXiv:1806.04959*, 2018.

[34] J. Helsby, S. Carton, K. Joseph, A. Mahmud, Y. Park, A. Navarrete, K. Ackermann, J. Walsh, L. Haynes, C. Cody, et al. Early intervention systems: Predicting adverse interactions between police and the public. *Criminal justice policy review*, 29(2):190–209, 2018.

[35] H. Holtzen, E. G. Klein, B. Keller, and N. Hood. Perceptions of physical inspections as a tool to protect housing quality and promote health equity. *Journal of health care for the poor and underserved*, 27(2):549–559, 2016.

[36] A. Howard and J. Borenstein. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24(5):1521–1536, 2018.

[37] M. Huzra. What do teachers spend on supplies, 2015.

[38] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[39] V. Iosifidis, B. Fetahu, and E. Ntoutsi. Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1375–1380. IEEE, 2019.

[40] D. J. James and L. E. Glaze. Mental Health Problems of Prison and Jail Inmates. Technical report, US Department of Justice, Bureau of Justice Statistics, 2006.

[41] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR, 2020.

[42] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[43] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[44] T. Kehrenberg, Z. Chen, and N. Quadrianto. Tuning fairness by balancing target labels. *Frontiers in Artificial Intelligence*, 3:33, 2020.

[45] E. G. Klein, B. Keller, N. Hood, and H. Holtzen. Affordable housing and health: a health impact assessment on physical inspection frequency. *Journal of public health management and practice*, 21(4):368–374, 2015.

[46] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27, 2018.

[47] J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable Algorithms. *University of Pennsylvania Law Review*, 165:633–706, 2016.

[48] A. Kumar, S. A. A. Rizvi, B. Brooks, R. A. Vanderveld, K. H. Wilson, C. Kenney, S. Edelstein, A. Finch, A. Maxwell, J. Zuckerbraun, et al. Using machine learning to assess the risk of and prevent water main breaks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 472–480, 2018.

[49] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1909–1918, 2015.

[50] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1), 2016.

[51] L.-P. Liu, T. G. Dietterich, N. Li, and Z.-H. Zhou. Transductive optimization of top k precision. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1781–1787, 2016.

[52] P. Manisha and S. Gujar. A neural network framework for fair classifier. *arXiv preprint arXiv:1811.00247*, 10, 2018.

[53] S. G. Mayson. Bias In, Bias Out. *Yale Law Journal*, 128:2018–2035, 2019.

[54] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[55] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR, 2018.

[56] I. Morgan and A. Amerikaner. Funding gaps 2018: An analysis of school funding equity across the us and within each state. *Education Trust*, 2018.

[57] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[58] O. A. Osoba and W. Welser IV. *An intelligence in our image: The risks of bias and errors in artificial intelligence*. Rand Corporation, 2017.

[59] D. Pedreshi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

[60] E. Potash, J. Brew, A. Loewi, S. Majumdar, A. Reece, J. Walsh, E. Rozier, E. Jorgenson, R. Mansour, and R. Ghani. Predictive modeling for public health: Preventing childhood lead poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2039–2047, 2015.

[61] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

[62] A. Ramachandran, A. Kumar, H. Koenig, A. De Unanue, C. Sung, J. Walsh, J. Schneider, R. Ghani, and J. P. Ridgway. predictive analytics for retention in care in an urban hiv clinic. *Scientific reports*, 10(1):1–10, 2020.

[63] K. Rodolfa, P. Saliero, and R. Ghani. Bias and fairness. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, and J. Lane, editors, *Big data and social science*, chapter 13. CRC Press, 2020.

[64] K. T. Rodolfa, H. Lamba, and R. Ghani. Machine learning for public policy: Do we need to sacrifice accuracy to make models fair? *arXiv preprint arXiv:2012.02972*, 2020.

[65] K. T. Rodolfa, E. Salomon, L. Haynes, I. H. Mendieta, J. Larson, and R. Ghani. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 142–153, 2020.

[66] K. T. Rodolfa, E. Salomon, L. Haynes, I. H. Mendieta, J. Larson, and R. Ghani. Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 142–153, New York, NY, USA, 1 2020. ACM.

[67] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

[68] P. Saleiro, K. T. Rodolfa, and R. Ghani. Dealing with bias and fairness in data science systems: A practical hands-on tutorial. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3513–3514, 2020.

[69] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019.

[70] J. L. Skeem and C. T. Lowenkamp. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4):680–712, 2016.

[71] S. Verma and J. Rubin. Fairness Definitions Explained. *IEEE/ACM International Workshop on Software Fairness*, 18:7, 2018.

[72] P. Williams, W. Kendall, and M. Hooten. Model selection using multi-objective optimization. *arXiv preprint arXiv:1810.10669*, 2018.

[73] T. Ye, R. Johnson, S. Fu, J. Copeny, B. Donnelly, A. Freeman, M. Lima, J. Walsh, and R. Ghani. Using machine learning to help vulnerable tenants in new york city. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 248–258, 2019.

[74] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *26th International World Wide Web Conference*, pages 1171–1180, Perth, Australia, 2017. WWW.

[75] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 962–970, Fort Lauderdale, FL, 4 2017. PMLR.

[76] I. Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.