Robustness of Fairness: An Experimental Analysis

Serafina Kamp[1], Andong Luis Li Zhao[2], and Sindhu Kutty[1]

[1] University of Michigan, Ann Arbor MI 48103, USA
{serafibk,skutty}@umich.edu
[2] Northwestern University, Evanston IL 60208, USA
andong@u.northwestern.edu

Abstract. Machine learning algorithms are increasingly used in making decisions with significant social impact. However, the predictions made by these algorithms can be demonstrably biased; oftentimes reflecting and even amplifying societal prejudice. Fairness metrics can be used to evaluate the models learned by these algorithms. But how robust are these metrics to reasonable variations in the test data? In this work, we measure the robustness of these metrics by training multiple models in two distinct application domains using publicly available real-world datasets (including the COMPAS dataset). We test each of these models for both performance and fairness on multiple test datasets generated by resampling from a set of held-out datapoints. We see that fairness metrics exhibit far greater variance across these test datasets than performance metrics, when the model has not been derived to be fair. Further, socially disadvantaged groups seem to be most affected by this lack of robustness. Even when the model objective includes fairness constraints, while the mean fairness of the model necessarily increases, its robustness is not consistently and significantly improved. Our work thus highlights the need to consider variations in the test data when evaluating model fairness and provides a framework to do so.

Keywords: classification · fairness · bootstrap sampling · robustness

1 Introduction

Machine learning methods use data-driven algorithms for automatic pattern recognition and prediction. Traditionally, the objective of these algorithms has been to optimize for performance metrics such as accuracy, which essentially measures the model's ability to make correct predictions about previously unseen data. These learned predictors can then be used to make decisions with significant societal impact. For instance, among other applications, machine learning is used in automated judicial review [2] and facial recognition for law enforcement [27].

Since machine learning models detect and learn from historical patterns in data, they can end up picking up and amplifying societal biases. Several recent

results show that the predictions based on these models can be demonstrably biased; for instance, automated facial analysis algorithms show significant accuracy differences across both race and gender [8] while music recommendation algorithms show gender bias in promoting artists [19]. The prevalence of these issues and the concerns they raise are well-documented, not only in machine learning literature, but also in the popular press [31, 32].

The degree of unfairness exhibited by these models can be captured by metrics that are widely accepted in machine learning literature [13, 21]. Typically, the fairness of the model can be evaluated by measuring it against test data. But how *robust* are these metrics to small perturbations in the data? Does the degree of robustness vary across models and application domains? And can we quantify the degree of *unfairness* across different *sub-populations*?

Fairness can be measured either as *individual* or *group* fairness. Group fairness metrics quantify how the model's predictions fare across different subgroups, often with an emphasis on subgroups that have been historically discriminated against. For instance, consider a model used to predict the probability of recidivism to determine whether or not to release a defendant for parole. This model may show different levels of predictive performance across different races. For one such model (used in US courts to predict recidivism), it has been shown that the probability of predicting a reoffence is greater for African American defendants than it is for Caucasian American defendants even when considering only those individuals who actually go on to reoffend [2]. One way to quantify this inequity in prediction is using the *equality of opportunity* fairness measure [21].

In this work, we are interested in measuring the reliability of fairness metrics when applied to a learned model. In particular, we first learn a predictive model using training data and then measure both its performance and fairness on test data. Crucially, rather than testing on a *single* held-out dataset, we measure fairness across variations in the testing data by generating multiple instances of this held-out dataset using bootstrap sampling [16, 17]. Effectively, bootstrap sampling uses the empirical distribution of the resampled data as a surrogate for the true distribution of datapoints. This allows us to measure the variation in both prediction error as well as fairness. We also explore the difference in both the mean and variance of both performance and fairness metrics across two different datasets with different semantic notions of socially disadvantaged groups: by *race* and by *age*. We show that fairness metrics are *less robust* (i.e., exhibit *significantly* more variance) than performance metrics under various underlying models; including models that use post-processing to achieve fairness. We also see that, typically, protected groups are the most affected by this lack of robustness.

1.1 Related Work

There has been significant work in quantifying fairness and designing techniques for achieving it [28, 37, 25, 33] as well as in understanding the implications of using fair predictors in practice [39]. The prevalence of bias in fields as wide-ranging as Natural Language Processing [36, 7], vision [8], ad-placement [41] and

health [1] have led to domain-specific analyses on bias detection and consequent work on both building and evaluating fairer datasets [4, 43].

There is no single agreed-upon measure of fairness since different contexts may require different criteria of measurement, including exogenous concerns like privacy-preservation [5, 6, 42]. In fact, so-called "impossibility theorems" show that some measures of fairness cannot be simultaneously satisfied [26, 12]. However, while there is no consensus measure of fairness, some tests for evaluating group fairness that have gained widespread acceptance include *demographic parity* [9], *equalized odds* and *equal opportunity* [21]. While we focus primarily on *equal opportunity* in this work, we also use *equalized odds* to derive a fair predictor.

There is an inherent tradeoff between the *performance* of a model, typically measured by metrics such as accuracy, and the *fairness* of the model, usually measured by how the predictor differs across different subgroups [29]. Achieving fairness in a predictive model can be framed by explicitly optimizing for fairness [18, 10], as constrained optimization problems [12, 21, 44, 14] and as conflicting objective functions [13].

Recent work has analyzed the effects of statistical and adversarial changes in the data distribution. Some of this work has focused on deriving fair models when there is a distributional shift in the data [38], when strategically acting adversaries inject errors in the data [11] or when the data is perturbed to negatively impact a particular subgroup [30, 3]. A survey of industry practitioners highlights the need to understand the practical implications of using fairness metrics [23].

In the present work, we study the *equal opportunity* fairness metric by focusing on the following research questions:

RQ1. For a given model, is the *equal opportunity* fairness metric a reliable measure of fairness? Does it show stability across reasonable fluctuations in the test data?

RQ2. How does the variation of the fairness metric compare with that of more traditional performance metrics?

RQ3. How much do different choices of models and features affect the robustness of the fairness metric? Is the robustness of the fairness measure affected by post-processing a model to satisfy fairness constraints? Further, does optimizing for a stronger notion of fairness affect the robustness of weaker notions of fairness?

RQ4. If we measure the effects of unfairness on different subgroups, do we see the same effects repeated across different datasets and models?

The rest of this paper is organized as follows: we cover background and a brief overview of our framework in Section 2. In Section 3 we provide details on the framework as well as the methodology for conducting our experiments. We also provide a description of the datasets and metrics used. We provide both numerical results and plots as well as an analysis of our results in Section 4. We conclude with a summary and directions for future work in Section 5.

2 Preliminaries and Overview

To learn a predictive model, we use logistic regression both with and without an ℓ_2 regularizer [22]. This involves solving the following optimization problem:

$$\min_{\bar{\theta},b} \quad C\sum_{i=1}^{n}\log(\exp(-y_i(x_i^T\bar{\theta}+b))+1)+\frac{1}{2}||\bar{\theta}||_2^2$$

where (x_i, y_i) are labeled training datapoints, $\bar{\theta}$, b are the learned parameters, and C is a hyperparameter that controls the degree of regularization.

Each datapoint has a corresponding binary label $\in \{0, 1\}$. For instance, in the COMPAS dataset (see Section 3.1) each datapoint corresponds to an individual and a label of 1 indicates an individual who re-offends within two years. The features that distinguish historically disadvantaged groups are called *sensitive attributes* and the groups themselves are called *protected groups* [21]. Each datapoint includes a sensitive attribute $z \in \{0, 1\}$ that indicates their membership in a protected group. We train the base classifier both including and excluding these sensitive attributes.

We use group fairness measures to evaluate the fairness of the predictor returned by the algorithm. In this work, we focus primarily on analyzing the equal opportunity fairness metric [21], which enforces equal true positive rates (TPR) across different groups. A formal definition of the fairness metrics used is given in Section 3.2. To achieve this fairness measure, the predictor is post-processed by solving a constrained optimization program with the constraints specifying the fairness conditions [21, 35, 34].

To evaluate a model, we rely on test data that is held out during the training process. However, datasets are only samples of the "true" data distribution; thus although they may be representative of the original distribution, there is a degree of uncertainty associated with these measures. We use the resampling technique of bootstrap sampling to generate multiple instances of the test dataset. We describe the resampling process in further detail in Section 3.3.

3 Framework and Experimental setup

We will now describe in detail our framework for evaluating the robustness of fairness metrics across uncertainty in test data. Prior work has cast the uncertainty in the training data using a Bayesian model [13]. However, we use a resampling approach to design experiments to study the empirical effects of this uncertainty on test data. The *robustness* of a metric is inversely related to the amount of variation we see in this measure across multiple instances of a given dataset; a robust metric should show minimal variance across sampling variations. We empirically analyze the robustness of the *equal opportunity* by measuring its variance across two datasets drawn from different application domains with the aim of measuring the persistence of our results across multiple learned models.

We use the COMPAS dataset [2] which has been used widely in machine learning literature to study fairness. We also run these empirical analyses on the South German Credit dataset [20]. Fundamentally, these datasets differ both in the social context (one was collected in the US in 2013-2014 and the other in Southern Germany in 1973-75). The historically disadvantaged groups in the two cases were also different (*race-based* discrimination vs. *age-based* discrimination). More details about these datasets are provided in Section 3.1.

Following prior work [28], the features that distinguish the traditionally privileged vs. disadvantaged groups are referred to as *sensitive attributes*. To understand the effects of the sensitive attributes on the learned model, we train the ML algorithm both with and without the sensitive attributes. We also investigate the robustness of both the fairness and performance metrics for different levels of model complexities by studying the effects of regularization.

By analyzing these metrics across two datasets and across different instantiations of test data, for different features and model complexities, including or ignoring fairness constraints, we are better able to assess the robustness of these metrics and the generalizability of these results. We describe the datasets, the metrics, and the methodology in further detail below.

3.1 Datasets

We use both the COMPAS dataset [2] and the South German Credit (SGC) dataset [15] for our analyses.

The COMPAS dataset contains 6150 datapoints with 8 features. The features include demographic information such as age, race, and sex as well as criminal history information such as priors, juvenile offences, and degree of current crime. When assuming a binary sensitive attribute, the dataset is restricted to Caucasian American and African American defendants; given the bias inherent in the dataset, African American defendants are considered to be the protected group. The binary-valued label indicates whether or not the individual has reoffended within two years after being released from prison.

The SGC dataset [15] contains 1000 datapoints with 20 features. The features of this dataset include demographic information such as age, sex, and marriage status, financial standing information such as credit history, savings account amount, and homeowner status, and, finally, information about the requested loan such as loan amount, purpose of loan, and duration of loan. Consistent with prior work [20, 24], we use age as the sensitive attribute for this dataset where an age of 25 years or younger are considered the protected group. The outcome for this dataset is a binary variable indicating whether or not the loan contract has been fulfilled after the duration of the loan.

3.2 Metrics

Accuracy For a given model, we measure its performance using accuracy [22]:

$$Acc = \frac{1}{N}\sum_{i=1}^{N}[[\hat{y} = y]]$$

where $\hat{y}$ is the outcome predicted by the model, y is the true outcome and N is the number of samples we are evaluating.[3]

Equality of Opportunity and Equalized Odds While there is no single agreed upon way to measure fairness, one metric that has semantic relevance for the datasets we consider and is widely accepted is *equal opportunity* [21]. A predictor is said to satisfy equal opportunity if and only if

$$\Pr(\hat{y} = 1|z = 1, y = 1) = \Pr(\hat{y} = 1|z = 0, y = 1)$$

where z is a sensitive attribute. For the COMPAS dataset, this can be interpreted as requiring the predictor to be agnostic to race for individuals who reoffend. For the SGC dataset, equal opportunity means that the probability of predicting a loan default should not change based on an individual's age for those individuals who repaid their loan.

We also consider a model where the predictor is modified to satisfy the stricter measure of *equalized odds* [21], that additionally enforces equal false positive rates. Formally, equalized odds requires the following to hold:

$$\Pr(y = 1|z = 1, y = a) = \Pr(y = 1|z = 0, y = a) \quad \forall a \in \{0, 1\}$$

Degree of Fairness and Direction of Unfairness We also measure the extent to which a model deviates from equality of opportunity. We define the *degree of fairness* of the predictor as:

$$1 - |\Pr(\hat{y} = 1|z = 1, y = 1) - \Pr(\hat{y} = 1|z = 0, y = 1)|$$

The range of this measure is the unit interval $[0, 1]$; a higher value indicates a fairer model.

To identify the subgroup against which a predictor is biased, we defined the *direction of unfairness* as

$$\text{sign}[\Pr(\hat{y} = 1|z = 1, y = 1) - \Pr(\hat{y} = 1|z = 0, y = 1)]$$

For example, in the COMPAS dataset, $z = 1$ indicates an African American defendant and $z = 0$ indicates an Caucasian American defendant. So, a *positive* direction of unfairness corresponds to unfairness towards the protected group (in this case, African American defendants).

We compare the variance in the degree of fairness with the variance of accuracy across multiple models and instances of datasets. In the next section, we describe the methodology we use to measure this variance.

3.3 Methodology

We learn twelve different models on the training data to evaluate their effects on both the mean and variance of fairness and performance metrics:

[3] We use [[]] to denote the Iverson bracket which returns a value of 1 if the predicate contained within is true and 0 otherwise.

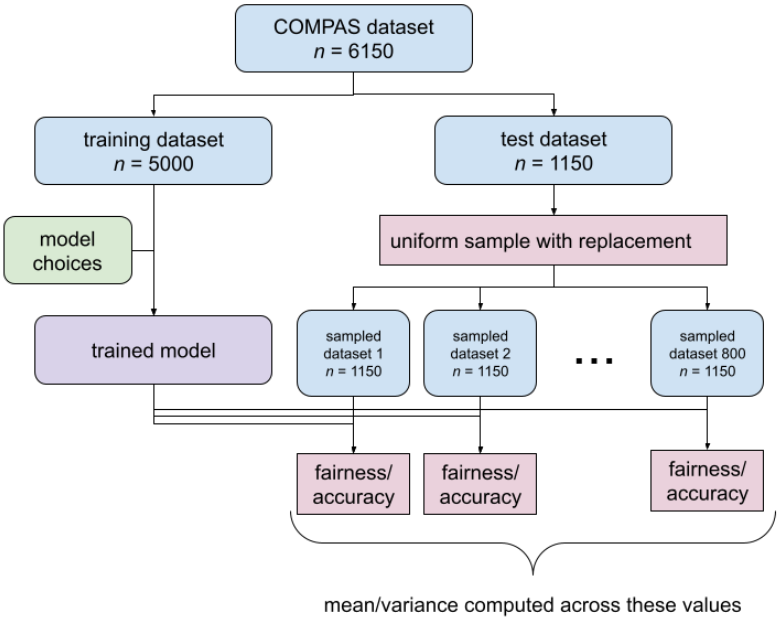


Fig. 1: Schematic illustrating our framework for measuring robustness of performance and fairness metrics (for the COMPAS dataset)

- Logistic regression
  - with and without an ℓ_2-norm regularizer.
  - including and excluding sensitive attributes while training.
- In addition to the above four models, we learn modified models by post-processing each above model to separately satisfy
  - equality of opportunity.
  - equalized odds.

In order to split the datasets into training and held-out sets, we first randomly shuffle each dataset. While splitting the datasets, we ensure the original proportion of positive to negative examples is preserved in the training and held-out set. For models trained with regularization we used 5-fold cross-validation to choose the hyperparameter that determines how much we penalize model complexity.

For the COMPAS dataset, we trained each model on 5000 points and held out 1150 for evaluation. For the South German Credit dataset, we trained each model on 600 points and held out 400 for evaluation. Specifically, for each dataset, we trained four models with and without regularization and with and without race and age, respectively, in the input feature vector. Then we applied post-processing for fairness constraints to train a total of twelve models.

We evaluate the performance and fairness of each model on multiple test datasets generated from the held-out dataset using bootstrap sampling. Each

sample set was the same size as the held-out set and was created by uniformly picking a point from the held-out set with replacement. We created 800 such sample datasets for each evaluation and then measured accuracy and degree of fairness on each sample dataset as described in Section 3.2. A schematic of this approach is shown in Figure 1. Note that both the degree of fairness and performance measures are defined on the unit interval; for both a higher value is more desirable.

We compute both the mean and variance of the degree of fairness and accuracy measures. We then compare the variance of these metrics over these 800 datasets in multiple ways.

- First, we numerically compute the variance achieved by these metrics and tabulate it for comparison across all twelve models (see Tables 1 and 2).
- Next, we plot the values of both metrics for each of the bootstrap sampled datasets (see Figure 3). For visual consistency, we plot fairness along the horizontal axis and performance along the vertical axis for all plots in that figure. We also use the same scale for both axes. A larger spread along a particular axis, therefore, indicates a larger variance along that metric.
- Then, we plot a histogram of both metrics for a visual representation of the distribution of these measures (see Figure 2 for the plots for two models on the COMPAS dataset. Due to space contraints, additional figures have been omitted.).
- Lastly, we translate both measures from the $[0, 1]$ to the $(-\infty, +\infty)$ interval by first centering to 0.5 mean and then applying the logit function to the values so obtained[4]. We see that the mapped values broadly follow a normal distribution. We then compute the variance of these mapped values and apply the F-test [40] to determine the significance of the difference in variances with high confidence[5].

We provide plots of accuracy vs. degree of fairness for each sample. We also provide the variance and mean of each of these metrics across the test sets. We describe our results in the next section.

4 Results

4.1 Variance of Fairness and Performance Metrics

As shown in Tables 1 and 2, we note that variances in degree of fairness are higher than for accuracy. As an example, Figure 2a shows visually that the spread of accuracy and degree of fairness can vary significantly. In fact, we show that difference in variance is statistically significant for various significance levels. We transform the data to the real number line using the logit function and

[4] Datasets with unit fairness were withheld in the F-test analysis due to numerical issues. However, these accounted for less than 1.5% of all 800 sample datasets.

[5] While the independence assumption does not strictly hold, the F-test gives us one more means of comparison.

Table 1: Mean (and variance) values in percentage for **accuracy** and **degree of fairness** for the COMPAS dataset reported for Logistic regression (LogReg); postprocessing for equal opportunity (EqOpp) and equalized odds (EqOdds); L2 indicates regularization.

	No Sensitive		Sensitive	
Data set	Accuracy	Deg of Fairness	Accuracy	Deg of Fairness
LogReg	62.16 (2.17)	78.32 (20.64)	62.73 (2.06)	61.50 (20.54)
LogReg + L2	61.97 (2.07)	77.37 (19.68)	62.37 (2.12)	65.01 (20.59)
EqOpp	58.72 (2.12)	96.54 (6.36)	56.59 (1.91)	96.17 (8.09)
EqOpp + L2	58.66 (2.00)	96.05 (7.62)	56.37 (1.97)	95.34 (10.84)
EqOdds	58.62 (2.09)	96.29 (7.77)	56.97 (2.20)	95.68 (10.18)
EqOdds + L2	58.68 (1.97)	96.56 (6.62)	56.92 (2.05)	95.25 (11.11)

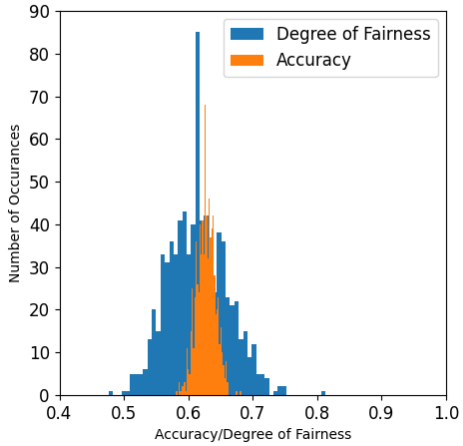
Table 2: Mean (and variance) values in percentage for **accuracy** and **degree of fairness** for the SGC dataset reported for Logistic regression (LogReg); post-processing for equal opportunity (EqOpp) and equalized odds (EqOdds); L2 indicates regularization.

	No Sensitive		Sensitive	
Model	Accuracy	Deg of Fairness	Accuracy	Deg of Fairness
LogReg	77.42 (4.53)	88.97 (35.38)	77.30 (4.07)	85.07 (43.70)
LogReg + L2	77.20 (4.54)	91.94 (27.45)	78.53 (4.00)	88.07 (34.54)
EqOpp	75.76 (4.36)	91.98 (27.71)	73.59 (4.10)	92.14 (29.90)
EqOpp + L2	75.22 (4.55)	94.56 (16.90)	74.44 (3.98)	94.72 (16.33)
EqOdds	74.77 (4.31)	91.55 (28.80)	73.66 (4.14)	92.04 (30.34)
EqOdds + L2	74.15 (4.61)	94.27 (18.47)	74.51 (4.01)	94.95 (14.87)

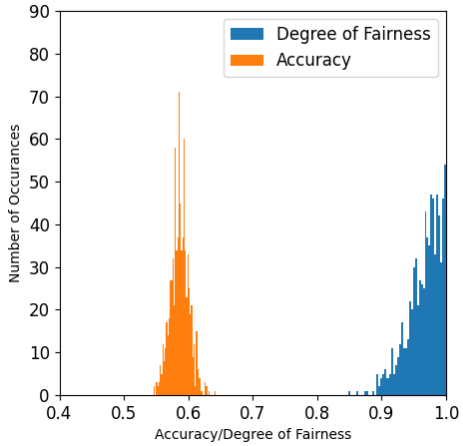
apply the F-test to this transformed data (see Section 3.3 for details). Table 3 reports these values for the logistic regression base classifier with regularization trained on data with sensitive attributes both before and after post-processing for fairness constraints[6]. This indicates that the fairness metric of equal opportunity is not as robust as accuracy across the sampled test sets.

Once we post-process for fairness constraints, we see that, as expected, mean degree of fairness improves. We also note that the variance in degree of fairness reduces significantly, especially for the COMPAS dataset (see Table 1). This can be seen visually in Figure 2b, where for comparison the spread in accuracy is indicated as well. We note, however, that the differences in variance of degree of fairness and accuracy are still statistically significant for all models, with the variance of degree of fairness always being higher than that of accuracy. We see in Table 3 that the F-test value is much larger than the f-critical value for the

[6] While we do not report results on all models due to space constraints, the omitted results are similar to reported values



(a) LogReg, no regularization, with sensitive attribute



(b) EqOdds, regularized, no sensitive attribute

Fig. 2: Histogram showing the difference in mean and variance of degree of fairness and accuracy scores for different models on the COMPAS dataset. Figure 2a includes scores for logistic regression without regularization including sensitive attributes. Figure 2b includes scores for logistic regression with regularization but without sensitive attributes and post-processing for equalized odds fairness constraint.

Table 3: F-test for statistical significance of the difference between performance and fairness variances reported for Logistic regression (LogReg); postprocessing for equal opportunity (EqOpp) and equalized odds (EqOdds). All models include sensitive attributes and a regularizer term. ✓ indicate that the ratio is higher than the F critical value, thereby indicating that the difference is statistically significant

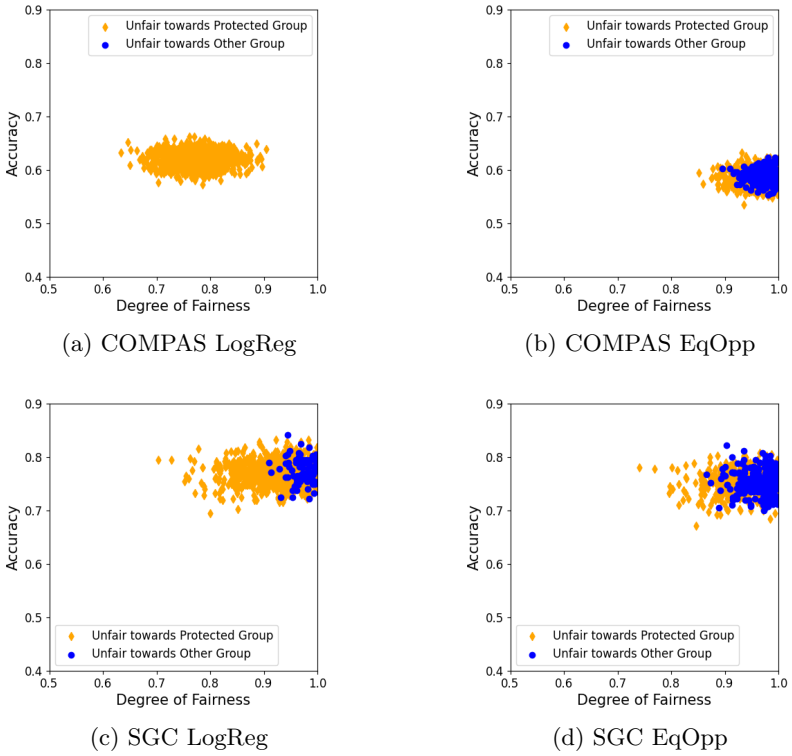
			$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.001$
Data set	Variances	Ratio	1.1234	1.1488	1.2446
COMPAS	LogReg	9.853	✓	✓	✓
SGC	LogReg	9.188	✓	✓	✓
COMPAS	EqOpp	5.542	✓	✓	✓
SGC	EqOpp	9.188	✓	✓	✓
COMPAS	EqOdds	8.793	✓	✓	✓
SGC	EqOdds	8.753	✓	✓	✓

number of observations, thus indicating high confidence that the variances are in fact significantly different.

When comparing the effect of incorporating different fairness constraints, we note that both equalized odds as well as equality of opportunity yield fairly similar results for degree of fairness. Typically, we observe that means of degrees of fairness of models with post-processing for fairness constraints are within at most 1% of each other. We also observe that in most cases equality of opportunity and equalized odds have comparable magnitudes of degree of fairness variance. However, in the case of unregularized base classifiers, equality of opportunity has a smaller degree of fairness variance; a likely explanation for this lies in our measure of degree of fairness which explicitly checks for deviation from the equality of opportunity measure.

The effects of incorporating fairness constraints on accuracy have been previously observed [29]. This is corroborated in our experiments as we observe a trade-off between accuracy and degree of fairness. In all cases, adding a fairness constraint reduced overall accuracy; however, the effect on its variance was typically minimal and inconsistent in direction indicating that adding fairness constraints does not seem to effect stability of the performance measure. Amongst models that were optimized for fairness, we notice that their mean accuracy is quite similar, being within at most 1% of each other's performance. This can be explained by the relationship between the fairness constraints and the degree of fairness measure. Another important trend we note is that higher mean degree of fairness generally corresponds to lower degree of fairness variance.

The effects of both including sensitive attributes in training the model, and adding a regularization term in the objective function, are mixed. The best performing models for accuracy are logistic regression models with access to sensitive attributes; however, these are often among the worst performing with respect to the mean and variance of degree of fairness. We also note that reg-



(a) COMPAS LogReg (b) COMPAS EqOpp

(c) SGC LogReg (d) SGC EqOpp

Fig. 3: Scatter plot for degree of fairness and accuracy. Orange diamonds indicate unfairness towards protected group, blue dots indicate unfairness towards the other group. Plots shown for the COMPAS and SGC datasets for Logistic regression (LogReg); postprocessing for equal opportunity (EqOpp) trained with regularization and without sensitive attributes.

ularization has a significant effect on variance of degree of fairness especially when post-processing for fairness in the SGC dataset (Table 2) as compared to the COMPAS dataset (Table 1). This can be likely explained by the difference in sizes of the two datasets.

A notable case is when we use a logistic regression model and fairness post-processing with access to sensitive attributes in the COMPAS dataset, which we can see in Figure 2b. In this case, the mean accuracy is roughly 56%, which is only slightly better than naively predicting the most common label in the dataset (which would give roughly 53% accuracy). This might indicate that there are degenerate cases of fairness where predictions are equally uninformative for different subgroups, potentially because the solution space is too restricted by regularization and fairness constraints.

4.2 Direction of Unfairness

In addition to looking at the general trends of fairness, we also explore the direction of unfairness in these models. In Figure 3, we show a scatter plot of the 800 bootstrapped sampled test datasets (for both SGC and COMPAS datasets) along the accuracy and degree of fairness axes. We also indicate which group the model is unfair towards. We can clearly see that generally the models are unfair towards the protected groups. Fairness constraints help shift the entire distribution to more fair outcomes, but we still see that most of the unfairness is to the detriment of protected groups. The plots for other models are omitted due to space constraints, but they show similar results as well.

5 Conclusions and future work

In this paper, we have provided a framework for evaluating the robustness of fairness metrics across uncertainty in test data. To do this we resample test data using bootstrap sampling and compute both the mean and variance of degree of fairness and accuracy. This allows us to compare the variations across these metrics for different learning models. We train logistic regression model for binary classification with and without a regularizer, as well as with and without sensitive attributes. We also post-process these models to separately satisfy two separate fairness constraints. We evaluate these twelve models separately on 800 bootstrapped test datasets to measure the variability as well as the mean of both a performance metric and a fairness metric. We show that the equality of opportunity fairness metric is less robust to variations in the test data than the accuracy performance metric. We highlight that current post-processing methods for improving fairness can affect mean fairness and reduce fairness variance; by and large, however, the variance of fairness still remains significantly higher than that of performance. We show that variance in model fairness is typically to the detriment of protected groups, making fairness variance analysis an important part of developing robust and fair machine learning models.

This lays the groundwork for further exploration of the robustness of fairness across other learning models, including those that incorporate a notion of fairness in their objective. Additionally, we are also interested in whether these effects will persist across other fairness metrics and datasets. In particular, we are interested in exploring other group fairness metrics as well as individual fairness metrics. We are also interested in studying the effects of in-processing learning methods for fairness on its variance. We leave these questions for future work.

References

1. Abebe, R., Goldner, K.: Mechanism design for social good. AI Matters **4**(3), 27–34 (Oct 2018)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. Propublica (2016)

3. Awasthi, P., Kleindessner, M., Morgenstern, J.: Equalized odds postprocessing under imperfect group information. In: Chiappa, S., Calandra, R. (eds.) The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]. Proceedings of Machine Learning Research, vol. 108, pp. 1770–1780. PMLR (2020)
4. Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., Venkatasubramanian, S.: It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. CoRR **abs/2106.05498** (2021)
5. Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org (2019), http://www.fairmlbook.org
6. Binns, R.: On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 514–524. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020)
7. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 4356–4364. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
8. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR, New York, NY, USA (23–24 Feb 2018)
9. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: 2009 IEEE International Conference on Data Mining Workshops. pp. 13–18 (2009)
10. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. **21**(2), 277–292 (2010)
11. Celis, L.E., Mehrotra, A., Vishnoi, N.K.: Fair classification with adversarial perturbations. CoRR **abs/2106.05964** (2021)
12. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
13. Dimitrakakis, C., Liu, Y., Parkes, D.C., Radanovic, G.: Bayesian fairness. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 509–516 (Jul 2019)
14. Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., Pontil, M.: Empirical risk minimization under fairness constraints. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 2796–2806. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
15. Dua, D., Graff, C.: UCI machine learning repository (2017), http://archive.ics.uci.edu/ml
16. Efron, B.: The bootstrap and modern statistics. Journal of the American Statistical Association **95**(452), 1293–1296 (2000)
17. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Springer (1993)
18. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 259–268. KDD '15, Association for Computing Machinery, New York, NY, USA (2015)

19. Ferraro, A., Serra, X., Bauer, C.: Break the loop: Gender imbalance in music recommenders. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. p. 249–254. CHIIR '21, Association for Computing Machinery, New York, NY, USA (2021)
20. Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 329–338. FAT* '19, Association for Computing Machinery, New York, NY, USA (2019)
21. Hardt, M., Price, E., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016)
22. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer Series in Statistics, Springer (2009)
23. Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., Wallach, H.: Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?, p. 1–16. Association for Computing Machinery, New York, NY, USA (2019)
24. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012)
25. Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A.: Algorithmic fairness. AEA Papers and Proceedings **108**, 22–27 (May 2018)
26. Kleinberg, J.M., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: Papadimitriou, C.H. (ed.) 8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA. LIPIcs, vol. 67, pp. 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2017)
27. MacCarthy, M.: Mandating fairness and accuracy assessments for law enforcement facial recognition systems. The Brookings Institution (2021), https://www.brookings.edu/blog/techtank/2021/05/26/mandating-fairness-and-accuracy-assessments-for-law-enforcement-facial-recognition-systems
28. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. CoRR **abs/1908.09635** (2019)
29. Menon, A.K., Williamson, R.C.: The cost of fairness in binary classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 107–118. PMLR, New York, NY, USA (23–24 Feb 2018)
30. Nanda, V., Dooley, S., Singla, S., Feizi, S., Dickerson, J.P.: Fairness through robustness: Investigating robustness disparity in deep learning. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 466–477. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021)
31. Noble, S.U.: Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press (2018)
32. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, USA (2016)
33. Parkes, D.C., Vohra, R.V., et al.: Algorithmic and economic perspectives on fairness. CoRR **abs/1909.05282** (2019)
34. Pleiss, G.: Code and data for the experiments in "On fairness and calibration" (2013)

35. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
36. Prabhakaran, V., Hutchinson, B., Mitchell, M.: Perturbation sensitivity analysis to detect unintended model biases. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 5739–5744. Association for Computational Linguistics (2019)
37. Rambachan, A., Kleinberg, J., Ludwig, J., Mullainathan, S.: An economic perspective on algorithmic fairness. AEA Papers and Proceedings **110**, 91–95 (May 2020)
38. Rezaei, A., Liu, A., Memarrast, O., Ziebart, B.D.: Robust fairness under covariate shift. Proceedings of the AAAI Conference on Artificial Intelligence **35**(11), 9419–9427 (May 2021)
39. Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., Liu, Y.: How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations. Artif. Intell. **283**, 103238 (2020)
40. Snecdecor, G.W., Cochran, W.G.: Statistical Methods. Wiley-Blackwell (1991)
41. Sweeney, L.: Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. Queue **11**(3), 10–29 (Mar 2013)
42. Tran, C., Fioretto, F., Van Hentenryck, P.: Differentially private and fair deep learning: A lagrangian dual approach. Proceedings of the AAAI Conference on Artificial Intelligence **35**(11), 9932–9939 (May 2021)
43. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 547–558. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020)
44. Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web. p. 1171–1180. WWW '17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2017)