Debiasing Recidivism Predictions

Jonathan Hollenbeck, Manish Pandit and Amita C. Patil
Department of Computer Science
Stanford University
Stanford, CA 94305
(jonoh, manish7, amita2)@stanford.edu

Abstract

We employed various methods to debias an augmented recidivism dataset from ProPublica[8]. This dataset contains COMPAS scores, which quantify the likelihood that criminal defendants will reoffend. We augmented this with a crime description clustering, using a universal sentence encoder[13]. Next, we debiased against two variables, race and sex, using prior work[7] with reweighing[3] as a baseline. Finally, we extended existing methods for a binary (privileged/unprivileged) variable, Adversarial Debiasing[14] and Bias Regularization[4], to work for multiple categorical variables. Our results were reasonably successful for a binary (high/low risk) recidivism classification, but we were unable to handle more complicated thresholds. Thus, our models are still biased – in particular, African-American Males are disproportionately likely to receive a very high (> 0.8) risk prediction.

1 Introduction

Recent AI research demonstrates that bias-agnostic approaches can cause greater negative impacts than overtly prejudicial systems. This is particularly concerning in difficult or high stakes prediction settings, since biased models often disproportionately punish and deny benefits to at-risk groups. For example, COMPAS scores from North-pointe Inc predict nonviolent and violent recidivism for hundreds of thousands of inmates across the United States. Judges leverage this to determine sentencing and parole, potentially long after the initial evaluation. Research[7] argues that these scores are biased against African-American and Hispanic inmates, and significant prior work has considered remedies.

There is no objectively correct concept of fairness. Bias could be naturally defined as a difference in outcomes given the same prediction, marginalized over some sensitive variable. For example, if the average outcome for Female/Male is 0.6/0.8 when the prediction is 0.7, this would indicate a bias against females. However, for this project, we define unfairness as a difference in outcomes between a class and the mean, even when such discrepancies are well calibrated. We think it is more interesting and reasonable to frame the bias problem as a tradeoff between model accuracy and concrete fairness metrics, maximizing accuracy on one extreme and equality of outcomes on the other. That said, a reasonable person could argue otherwise, especially for violent recidivism.

2 Previous Work

Related research has identified three main categories of debiasing techniques: pre-processing, in-processing, and post-processing[12]. Pre-processing algorithms focus on removing problematic features and modifying weights before any training occurs. For example, reweighing[3] calculates sample weights to equalize mean class labels. Post-processing techniques address biased model

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

output, usually by modifying thresholds or equalizing rates of false positives and negatives[9]. This approach is especially practical if the model and its input are not accessible[12]. Finally, in-processing debiases the learning process itself, typically by adding a bias penalty[4] analogous to l_2 regularization. More recent approaches have used adversarial networks[14] by penalizing the main predictor if a second model can guess protected labels.

During our survey, we observed that combining these approaches is quite rare, since the choice of algorithm and metric generally reflects a specific goal. If the result is already contextually optimal, adding another debiasing layer must be counterproductive. Limited model access and large datasets can also preclude certain techniques.

We lean on software for most of our preliminary code. IBM AI Fairness 360[12] provides implementations for many of the algorithms and metrics we cite, and the corresponding paper[11] was invaluable as a resource for exploring previous work. We also use sk-learn[2] for basic ML models and tasks.

3 Dataset and Features

We leverage a dataset (Table 1) with over 10,000 criminal defendants in Broward County, Florida. This includes two decile scores from COMPAS software, which correspond to likelihood of [any] recidivism and violent recidivism after the defendant competes a survey. It also contains a set of 52 numerical, binary, categorical, and text features, which we map into a numerical input vector x_i for record i. Our y value (output) is known, and labels whether or not the defendant reoffended within a two year window.

id	Sex	Age	Race	Priors	Description	COMPAS	Two Year Recid
1	Male	69	Other	0	...Assault	3	No
3	Male	34	African-American	4	Felony Battery	4	Yes
10	Female	39	Caucasian	0	Battery	1	No

Table 1: Subset of COMPAS Nonviolent Data Fields

COMPAS decile scores display clear disparities along racial lines. For example, African-American defendants are far more likely to receive a high decile score (Figure 1).

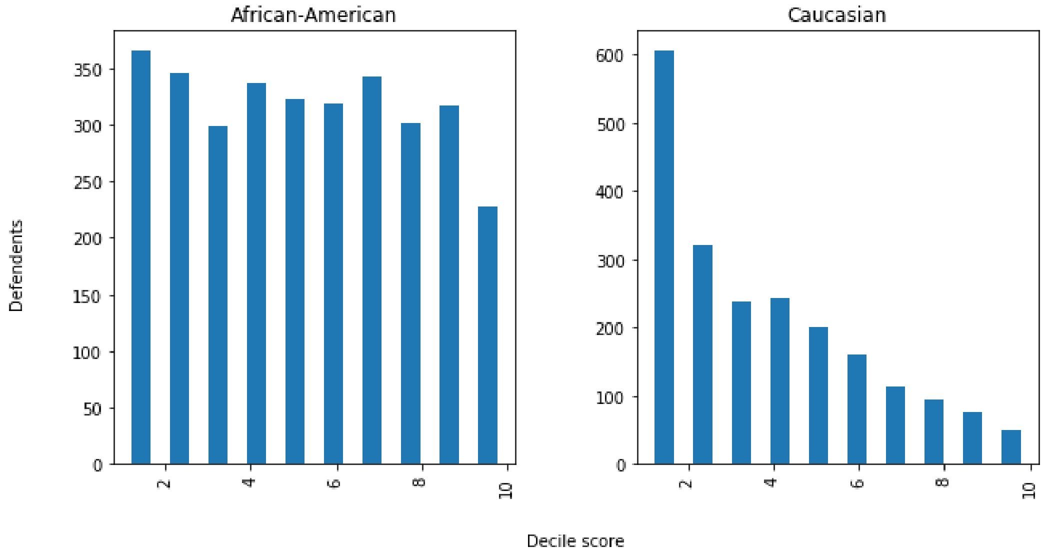


Figure 1: Racial Bias in nonviolent COMPAS deciles

All algorithms struggled with low population classes, since we do not have enough samples to check generalization. For example, there were 31 Asian Males, so the expected sizes of the train/test sets are 25/6. For this reason, we merged classes with fewer than 50 samples (Asian Males/Females, Native American Males/Females) to ’Other’. We acknowledge poor handling of small classes as a major flaw, and preferred to avoid this step.

3.1 Crime Description Clustering

In the dataset, crime degrees are explicitly differentiated by a binary variable (felony/misdemeanor). Since this is fairly low information, we tried to augment this with unsupervised learning on the crime description field, which contains a short, technical description of the crime. First, we computed a node embedding, using TF-IDF[1] sparse vectors and kmeans with k=5. However, this generic NLP approach failed to capture meaningful clusters, probably because our sample size is too small. Next, we used a universal sentence encoder[13] to convert the charge descriptions into 512-dimension semantic vectors, then ran kmeans with k=5 to cluster charges. Analysis shows that these clusters are meaningful, and can be approximately summarized as: other, sex/minors, drugs, vehicles, and violence (Table 2, Figure 2).

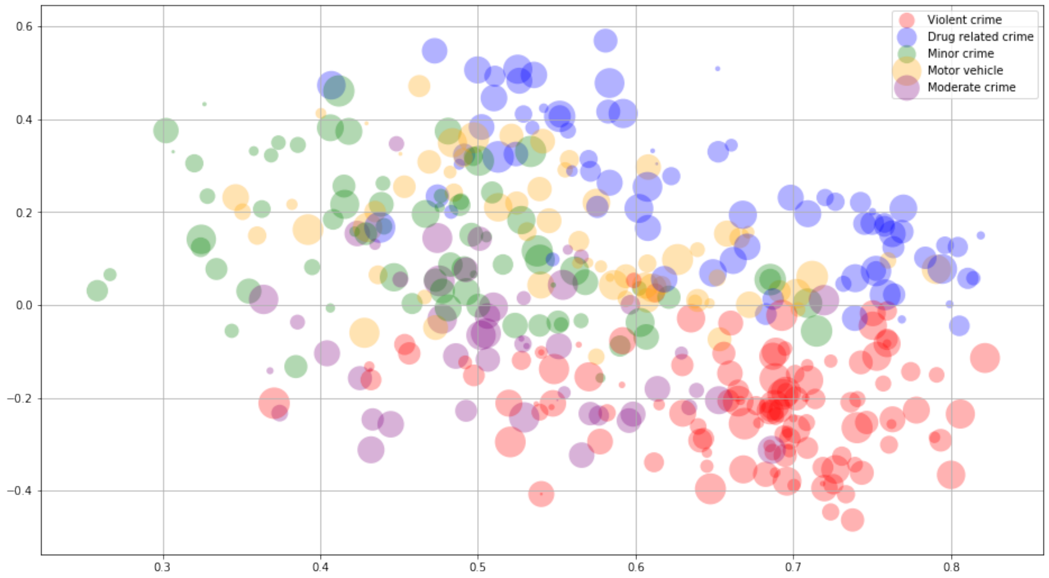


Figure 2: Crime Description clusters

There are lots of exceptions to this and descriptions (violent sex crimes while on drugs) that should map to multiple categories. There are also instances where the clustering is clearly inappropriate, such as "Violation Of Boater Safety Id" mapping to the "Violent Crime" cluster. Finally, we note that the degree of crime varies significantly within these categories. For example, trafficking 150kg of cocaine and purchase of cannabis map to the same cluster (#2).

cluster	size	example1	example 2	description
0	466	Attempted Burg/struct/unocc	Defrauding Inkeeper	Other/Moderate Crime
1	230	Aggrav Child Abuse-Causes Harm	Live on Earnings of Prostitute	Sex & crimes relating to minors
2	1179	Traffick Amphetamine 28g	Posession of Paraphernalia	Drug related crime
3	610	Failure To Pay Taxi Cab Charge	Expired DL More Than 6 months	Motor Vehicle
4	3687	Aggravated Assault	Felony Battery	Violent Crime

Table 2: Crime Description Clusters

4 Problem Definition

4.1 Baselines

As this dataset is highly popular for studying the bias topic, our primary baseline is the best performance we can achieve with existing software and methods. We also present simple accuracy/fairness baselines for COMPAS scores alone and each considered debiasing method.

Specifically, we applied logistic regression[2] to predict non-violent recidivism. We tried a few other methods available in sklearn, such as SVM, Neural Nets, and XGBoost, but ended up with very similar results ($\pm 1\%$) on test data. Then, we developed simple proof of concept debiasers using methods from IBM AI Fairness 360[12] and race as the protected variable. We also replicated and modified previous COMPAS analysis by ProPublica[7] using Reweighing, a pre-processing method. This was all done at the start of this project to validate our basic infrastructure and problem statement.

4.2 Oracles

Our dataset is labeled, so a perfect oracle merely returns the y values. We also compute an oracle for the accuracy/fairness tradeoff curve by debiasing with y values as input. There is a significant gap between our baselines and this oracle, which we do not expect to close since no previous work claims accuracy over 70%, even without debiasing. However, this serves to illustrate the degree to which perfect accuracy and fairness cannot be achieved simultaneously.

4.3 Evaluation Metrics

We evaluate fairness using the following metrics: Disparate Impact, Statistical Parity Difference, Average Odds Difference, Equal Opportunity Difference, and Ranked Group Fairness. We discuss our results in terms of these properties. We have also explored more complicated metrics, some of which are software-implemented [12].

Metric	Description
Disparate Impact	$x = \Pr[\hat{y} = 1 \mid u] / \Pr[\hat{y} = 1 \mid v]$. DI $= -(max(x, 1/x) - 1)$
Statistical Parity Difference	$-\lvert \Pr[\hat{y} = 1 \mid u] - \Pr[\hat{y} = 1 \mid v] \rvert$
Average Odds Difference	Compare True and False positive rates for u and v
Equal Opportunity Difference	Difference in True Positive rates for u and v
Ranked Group Fairness[10]	KL-divergence of $Pr[u]$ and $Pr[u \mid \hat{y} \geq p]$ over $p \in (0, 1)$

Table 3: Fairness Metrics (v is privileged, u is unprivileged)

As a note, all our fairness metrics are ≤ 0. Some of them are naturally the opposite, but we found it very confusing to equate positive metrics decreasing with negative metrics increasing (especially graphically). They are also highly correlated for our dataset, so focusing on one is generally sufficient.

When handling multiple classes, we computed these metrics by taking a weighted (by class size) average of the fairness metrics for each class. For example, one of these calculations would flag Hispanic Males as 1 and all other classes as 0, then calculate disparities between these binary classes.

5 Adversarial Debiasing

Goodfellow et al. [5] pioneered the technique of using multiple networks with competing goals to force one network to "deceive" another, applying this method to the problem of creating real-life-like pictures. In [14] and [15], the authors demonstrate use of this technique to reduce bias in recidivism and word embeddings while maintaining performance on certain tasks. Our adversarial model is inspired by these prior ideas.

Our model (Figure 3) consists of a two-layer main recidivism neural network *NN* whose primary objective is to predict recidivism $\hat{y}$ from a set of input features *X*. We add two adversarial networks Adv_R and Adv_S that penalize our main recidivism neural network if demographic information, such as race Z_R and sex Z_S, can be predicted from the logits of the main *NN*. The two hidden layers of the main *NN* each have 256 hidden units with tanh activation, followed by one fully connected layer. Logits of the neural network *NN* are fed as input to the adversaries: Adv_R, Adv_S. Both the adversarial networks contain one hidden layer with tanh activation. Adv_R predicts six different race classes present in the data-set and has 100 hidden units. The second adversarial network Adv_S predicts sex and has 10 hidden units. The goal is for neural network *NN* to predict *ŷ* i.e. recidivism accurately and for adversaries Adv_R, Adv_S to predict demographic information Z_R, Z_S poorly. We begin by modifying weights θ_y to minimize loss L_Y(*ŷ*,*y*) using batch gradient descent. Suppose the adversaries have a loss L_Z(*ẑ*,*z*) and weights θ_z. Then, every K^{th} epoch, we update θ_z to minimize L_Z(*ẑ*,*z*). We modify θ_y as per the expression given below:

$$\theta_y = \theta_y - \eta \frac{\partial L_Y}{\partial \theta_y} - \alpha_R \frac{\partial L_{Z_R}}{\partial \theta_{z_R}} - \alpha_S \frac{\partial L_{Z_S}}{\partial \theta_{z_S}} \tag{1}$$

Here η (learning rate), α_R, α_S and *K* are hyper-parameters that were tuned to optimize fairness/accuracy. Since we subtract L_{Z_R} and L_{Z_S} from L_Y, *NN* is encouraged to maximize L_{Z_R} and L_{Z_S}. Therefore, the logits cannot be used to predict race or sex and we can achieve unbiased recidivism prediction. We used binary cross-entropy with logits losses for recidivism and sex predictors, *NN* and Adv_S. We used sparse softmax cross-entropy with logits loss for Adv_R, since it predicts multiple race classes. The model was trained using tensorflow [6], and an ADAM optimizer was used for loss minimization.

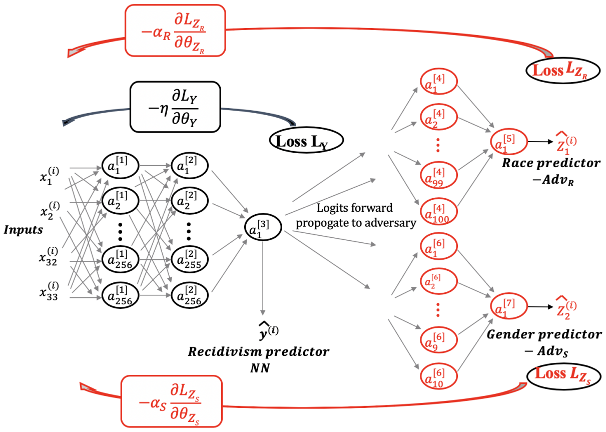


Figure 3: Model of adversarial neural network

6 Bias Regularization

For initial work, we used existing optimization code[4] and an AIF360[12] wrapper. This required lib modifications, which are not currently in the repository. We also found that the bias function only permits a single sensitive variable, which we resolved by concatenating them (if sex,race are sensitive, then 'Caucasian,Male' becomes 'Caucasian-Male'). This formulation creates a number of classes exponential in the number of variables, so further work likely requires a different approach.

Bias regularization adds a penalty for bias in $\hat{Y}$, analogous to penalizing model size with l_2 regularization. Thus, our objective is (where $\hat{Y} = f(X, w)$ and R is the bias function).

$$\underset{w}{\text{argmin}} \sum_i L(f(x_i, w), y_i) + \lambda ||w||_2 + \eta R(f(X, w), Y) \tag{2}$$

For bias function implementation, we used python code provided by the author[4], with slight modifications as described above. This formulation is identical to equation 11 in this paper, restated in our notation for clarity. Also, some technical details (similar to Laplace Sampling) are important to prevent $\pm\infty$ terms.

$$R(f(X), Y) = \sum_i \sum_{y \in \{0,1\}} [y\sigma(f(x_i, w)) + (1 - y)(1 - \sigma(f(x_i, w))]log \frac{Pr[\hat{y} = y|s_i]}{Pr[\hat{y} = y]} \tag{3}$$

For intuition, this adds a large penalty when the prediction goes in the same direction as the bias, and a reward in the opposite case. If we predict $\sigma(f(x_i, w)) = 0.99$ for a class with twice the average rate of $y = 1$, our penalty will be $0.99 * log(2/1) + 0.01 * log(1/2) \approx 0.68$. Conversely, if our prediction was 0.01, we would receive a negative penalty (reward) of 0.68. This helps SGD converge to an optimum where class rates equal the average rate, and there is thus no penalty [$log(1/1) = 0$].

As an extension, we tried to add a bias penalty on discretized sublevel sets of $\hat{y} = f(x_i, w)$. For example, if one of the sublevel sets is $\hat{y} \leq a$ and g is a function (related to an appropriate information metric), we have corresponding penalty:

$$R(a, s_i) = \eta g(Pr(\hat{y}|s = s_i, \hat{y} \leq a), Pr(\hat{y})|\hat{y} \leq a)) \tag{4}$$

We then update this using SGD by caching $Pr(\hat{y}|\hat{y} \leq a)$ for each (a, s_i) to debias on $||Pr(\hat{y}|s = s_i, \hat{y} \leq a|| = Pr(\hat{y}|\hat{y} \leq a)$. However, this approach ran into serious problems with local optima that we were unable to resolve.

7 Results and Analysis

Algorithm	Debiasing	Accuracy %	Disparate Impact %	Avg. Odds Difference %
Baseline, COMPAS only	None	64.9	-49.7	-25.1
Baseline	None	68.3	-60.3	-26.1
Reweighing	Yes[3]	65.8	-38.9	-14.1
Adversarial	None (α=0)	69.6	-80.5	-29.0
Adversarial	Yes[14] (α=0.008)	62.8	-22.1	-15.0
Regularization	No (eta=0)	70.3	-39.0	-15.8
Regularization (Race & Sex)	Yes[4] (eta=50)	63.2	-6.9	-3.6
Oracle	None	100	-47.6	-26.1
Oracle	Reweighing	90.4	0	0

Table 4: Debiasing over Race Variable

7.1 Reweighing

The reweighing algorithm generalizes well for a single privileged and unprivileged class. For example, if we debias only on sex, or bundle 'African-American/Hispanic' and 'Everything Else', the results are reasonably good. However, if we include more classes, generalization is poor.

On inspection, it appears that the model we train for more classes mostly memorize quirks in the data and decreases average predictions across the board. Classes with large (positive and negative) disparity also tend towards predictions of 0.55 (the average). None of this translates well to test data, and we tend to see a 're-biasing'. The magnitude of this effect increases with the number of classes (Table 5).

Variables	Classes	Train/Test Accuracy %	Train/Test Disparate Impact %
sex only	2	67.1/66.2	-1.1/-2.4
race only	4	65.9/66.3	-2.4/-15.1
sex, race	8	64.8/65.8	-2.3/-38.9

Table 5: Reweighing

Fixing this is difficult because reweighing does not have bias hyperparameters; we are merely changing weights to equalize average class labels.

7.2 Adversarial Debiasing

We trained our model with the following hyperparameters: learning rate = 0.0001, # of epochs = 800 and *K* = 10. α_R and α_S were varied from 0.0 to 0.01, and thresholds for sigmoid activations of the main *NN* and Adv_S were tuned to optimize accuracy/fairness. We observed that adding crime description clusters to the feature set showed significantly higher bias toward African-American Males

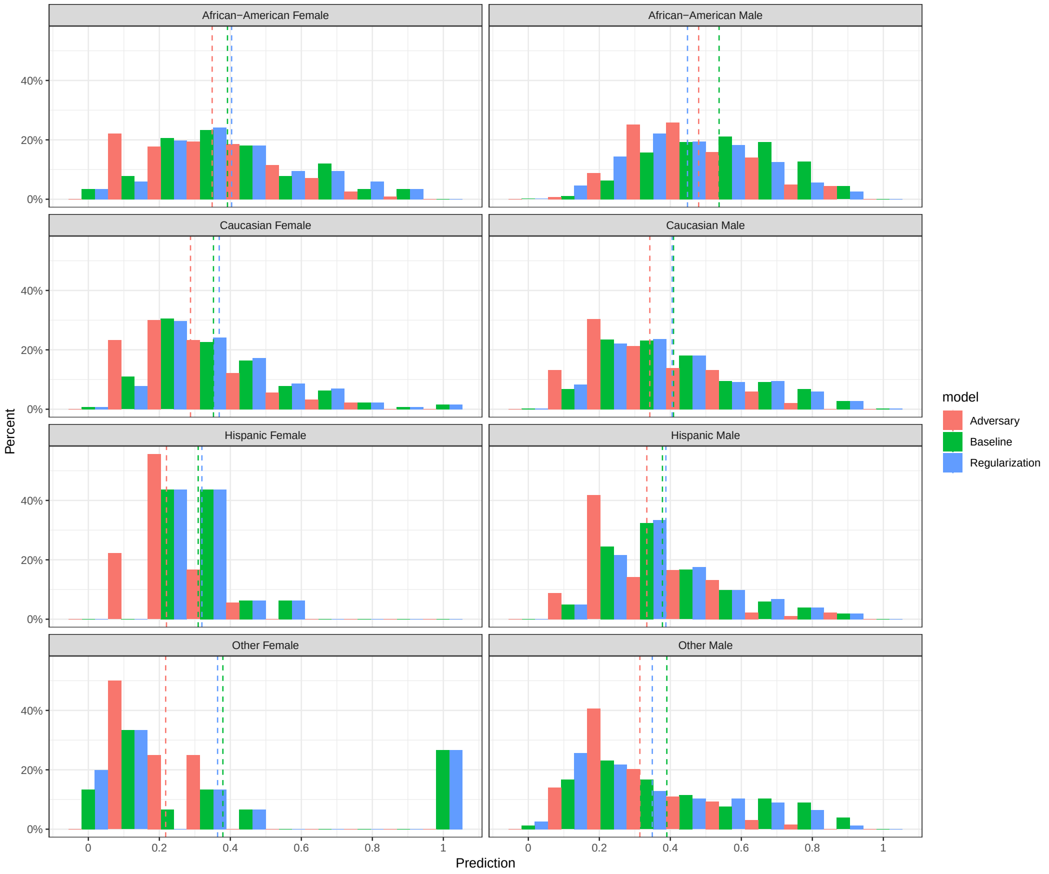


Figure 4: Prediction Distributions by Class and Model

for $\alpha = 0$ (Figure 5). We suspect this is because the crime description clusters provide additional useful information to our model, and therefore exploit inherent bias in the dataset more accurately. The best results from both models were similar, indicating that the adversarial debiasing approach improved fairness. However, to remove bias completely while retaining accuracy we may need to access additional features.

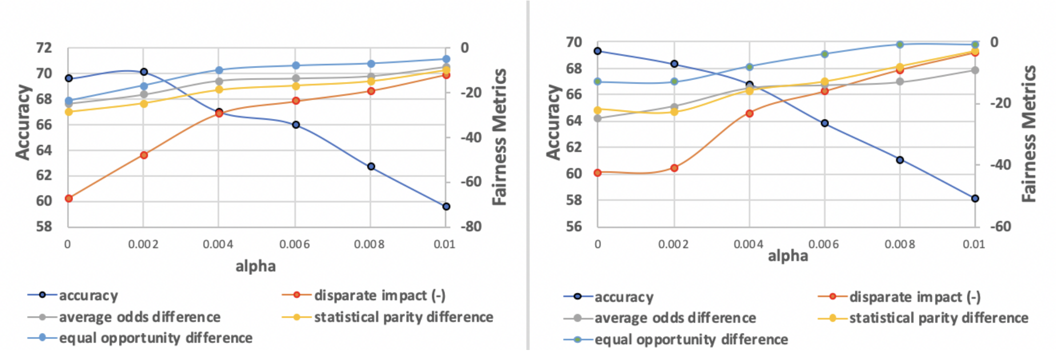


Figure 5: Training α's for adversarial debiasing, with and without crime description clusters

Without a bias penalty, our model's accuracy was 70%, which is inline with literature. As is evident from Table 6, bias against African-Americans is significant. Disparate impact is -80.5% and average odds difference is -29%. With adversarial debiasing, we were able to improve disparate impact by a factor of 4 and average odds difference by a factor of 2. Our model is able to reduce bias for all race/sex protected classes as per our defined measures of fairness, but we could not achieve zero bias, and fairness came an the expense of slightly reduced accuracy.

Demographic	Disparate Impact % (α = 0)	Disparate Impact % (α = 0.008)
African-American	-80.5	-22.1
Hispanic	-26.0	-12.7
Male	-39.1	-12.6
African-American+Hispanic Male	-72.0	-19.2

Table 6: Fairness metrics with and without adversary for various protected classes

7.3 Bias Regularization

Increasing η to about 50 nearly satisfied all the fairness metrics (Table 4). Increasing it further did not help much, and hurt accuracy a bit. Curiously, training was far smoother after we added the crime description clusters (Figure 6). This is probably because the bias function is nonlinear, so convergence to a global minimum is not guaranteed. Improving the information quality made it easier to find that minimum.

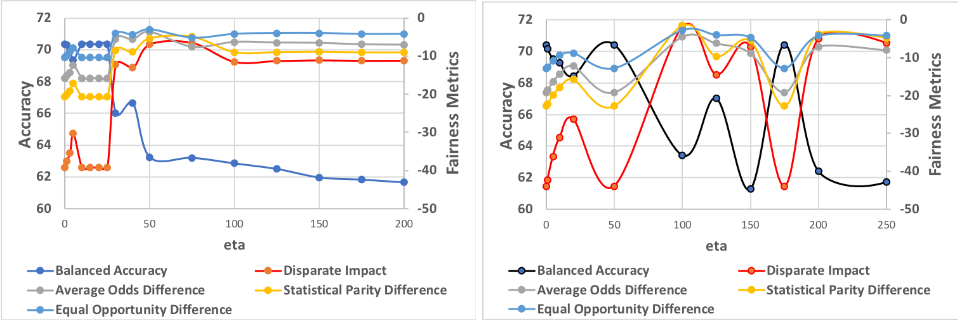


Figure 6: Training Bias Regularization, with and without Crime Description clusters

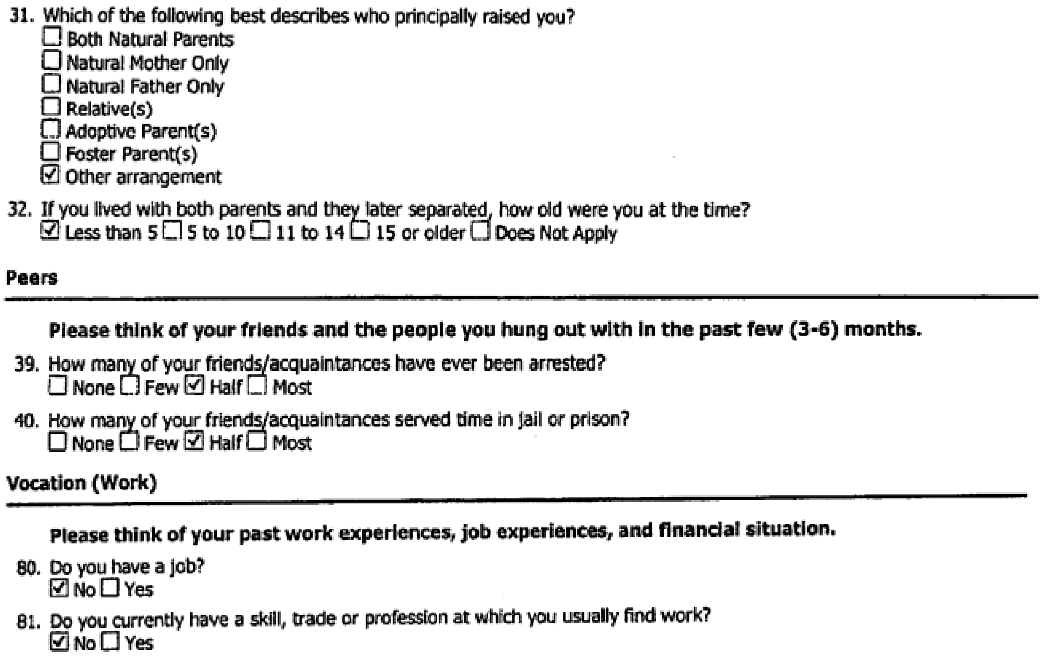
For $\eta = 50$, our average Disparate Impact was -6.9%. This means that that $Pr(y|s = s_i) \approx Pr(y)$ after debiasing, where s is the sensitive variable and s_i is in the domain of s. However, this fairness is not necessarily uniform given $\hat{y}$. In particular, we observed that African-American males disproportionately receive high (≥ 0.8) scores despite a class mean near the population after debiasing (Figure 4).

Class	Test Examples	DI % ($\eta = 0$)	Disparate Impact % ($\eta = 50$)
African-American Male	519	-65.9	-7.6
African-American Female	100	-25.9	-18.4
Hispanic Male	338	-15.9	-2.7
Hispanic Female	86	-23.4	-8.7
Caucasian Male	91	-19.9	-5.1
Caucasian Female	16	-30.7	-14.7
Other Male	56	-15.8	-0.2
Other Female	18	-21.9	-10.3

Table 7: Bias Regularization, $\eta = 50$

7.4 Other Analysis

Debiasing against specific variables increases reliance on others. For example, young males are particularly high risk to reoffend, so the model tends towards higher predictions for young males and lower predictions for older males when we debias on sex. There are also important latent variables, such as economic status and location. Some COMPAS questions[16] raise concerns about this, since they ask about job prospects, family history, and if many of your acquaintances have been arrested



31. Which of the following best describes who principally raised you?
□ Both Natural Parents
□ Natural Mother Only
□ Natural Father Only
□ Relative(s)
□ Adoptive Parent(s)
□ Foster Parent(s)
☑ Other arrangement

32. If you lived with both parents and they later separated, how old were you at the time?
☑ Less than 5 □ 5 to 10 □ 11 to 14 □ 15 or older □ Does Not Apply

Peers

Please think of your friends and the people you hung out with in the past few (3-6) months.

39. How many of your friends/acquaintances have ever been arrested?
□ None □ Few ☑ Half □ Most

40. How many of your friends/acquaintances served time in jail or prison?
□ None □ Few ☑ Half □ Most

Vocation (Work)

Please think of your past work experiences, job experiences, and financial situation.

80. Do you have a job?
☑ No □ Yes

81. Do you currently have a skill, trade or profession at which you usually find work?
☑ No □ Yes

Figure 7: COMPAS sample questions

(Figure 7). We anticipate that decreasing bias on race and sex increased bias for many of these factors.

Recidivism significantly varied within classes when broken out by crime description clusters (Table 8). This additional information may explain the improved model performance and training consistency, and rates were fairly consistent between train/test sets. However, we have not run any significance tests, so this should be interpreted as an observation.

Class/Crime Description Cluster	Other	Sex/Minors	Drugs	Vehicles	Violence
African-American Male	0.64	0.58	0.59	0.53	0.54
African-American Female	0.42	0.28	0.56	0.43	0.33
Caucasian Male	0.50	0.44	0.37	0.43	0.40
Caucasian Female	0.51	0.38	0.34	0.31	0.34
Hispanic Male	0.33	0.53	0.32	0.47	0.38
Hispanic Female	0.33	0.33	0.22	0.63	0.29

Table 8: Recidivism probability by sensitive class and crime cluster

8 Conclusion

Both our methods, Adversarial Debiasing and Bias Regularization, produced models that generalized to maintain accuracy while significantly reducing bias. Results significantly improve on existing baselines, but neither model is robust to changing thresholds, nor do they fully remove bias.

8.1 Github Link

`https://github.com/manishpandit/recidivism`

We prepared a notebook with exploratory data analysis and crime description clustering: `https://nbviewer.jupyter.org/github/manishpandit/recidivism/blob/master/recidivism.ipynb`

References

[1] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, pp. 11–21, 1972.

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

[3] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, Oct. 2012, ISSN: 0219-3116. DOI: 10.1007/s10115-011-0463-8. [Online]. Available: https://doi.org/10.1007/s10115-011-0463-8.

[4] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Machine Learning and Knowledge Discovery in Databases*, P. A. Flach, T. De Bie, and N. Cristianini, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 35–50, ISBN: 978-3-642-33486-3.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14, Montreal, Canada: MIT Press, 2014, pp. 2672–2680. [Online]. Available: http://dl.acm.org/citation.cfm?id=2969033.2969125.

[6] M. et al., *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: http://tensorflow.org/.

[7] ProPublica. (May 2016). How we analyzed the compas recidivism algorithm, [Online]. Available: ttps://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[8] ——, (May 2016). Machine bias, [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[9] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 5684–5693, ISBN: 978-1-5108-6096-4. [Online]. Available: http://dl.acm.org/citation.cfm?id=3295222.3295319.

[10] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "Fa*ir: A fair top-k ranking algorithm," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, ser. CIKM '17, Singapore, Singapore: ACM, 2017, pp. 1569–1578, ISBN: 978-1-4503-4918-5. DOI: 10.1145/3132847.3132938. [Online]. Available: http://doi.acm.org/10.1145/3132847.3132938.

[11] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *CoRR*, vol. abs/1810.01943, 2018. arXiv: 1810.01943. [Online]. Available: http://arxiv.org/abs/1810.01943.

[12] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, *AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*, Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.01943.

[13] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil, "Universal sentence encoder," *CoRR*, vol. abs/1803.11175, 2018. arXiv: 1803.11175. [Online]. Available: http://arxiv.org/abs/1803.11175.

[14] C. Wadsworth, F. Vera, and C. Piech, "Achieving fairness through adversarial learning: An application to recidivism prediction," Jun. 2018.

[15] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," *CoRR*, vol. abs/1801.07593, 2018. arXiv: 1801.07593. [Online]. Available: http://arxiv.org/abs/1801.07593.

[16] N. Inc. (). Sample risk assessment, [Online]. Available: https://assets.documentcloud.org/documents/2702103/Sample-Risk-Assessment-COMPAS-CORE.pdf.