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

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[5] argues that these scores are biased against Black and Latino inmates, and significant prior work has considered remedies. In this project, we will explore debiasing an augmented COMPAS dataset and try to improve existing methods with respect to specific fairness metrics while minimizing loss of model accuracy.

1 Previous Work

Related research has identified three main categories of debiasing techniques: pre-processing, in-processing, and post-processing[9]. Pre-processing algorithms focus on removing problematic features and modifying weights before any training occurs. For example, reweighing[2] calculates sample weights to equalize mean class labels. Post-processing techniques address biased model output, usually by modifying thresholds or equalizing rates of false positives and negatives[6]. This approach is especially practical if the model and its input are not accessible[9]. Finally, in-processing debiases the learning process itself, typically by adding a bias penalty[3] analogous to l2 regularization. More recent approaches have used adversarial networks[10] 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[9] provides implementations for many of the algorithms and metrics we cite, and the corresponding paper[8] was invaluable as a resource for exploring previous work. We also use sk-learn[1] for basic ML models and tasks.

2 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 $\phi(x_i)$

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

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

We explored a node embedding for crime descriptions, since we thought this might expose new information with a bias correlation. However, our sample size is too small for generic NLP approaches. We still intend to incorporate them in some way, likely with sentiment analysis.

3 Problem Definition

3.1 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[1] to predict non-violent recidivism. Then, we developed simple proof of concept debiasers using methods from IBM AI Fairness 360[9] and 'Race' as the protected variable. We also replicated and modified previous COMPAS analysis by ProPublica[5] using Reweighing, a pre-processing method. This was all done at the start of this project to validate our basic infrastructure and problem statement.

3.2 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, it still illustrates that perfect accuracy and fairness cannot be achieved simultaneously here.

3.3 Evaluation Metrics

We evaluate fairness using the following metrics: Disparate Impact, Average Odds Difference, Ranked Group Fairness, and Statistical Parity Difference. We expect to discuss our results in terms of these properties. We have also explored more complicated metrics, some of which are software-implemented [9]. Our work so far satisfies these metrics with limited loss in accuracy, but analysis

Metric	Description
Disparate Impact	$\Pr[\hat{y}=1\|u]/\Pr[\hat{y}=1\|v]$
Statistical Parity Difference	$\Pr[\hat{y}=1\|u]-\Pr[\hat{y}=1\|v]$
Average Odds Difference	Compare True and False positive rates for u and v
Ranked Group Fairness[7]	KL-divergence of $Pr[u]$ and $Pr[u\|\hat{y}\geq p]$ over $p\in(0,1)$

Table 2: Fairness Metrics (v is privileged, u is unprivileged)

indicates persisting bias issues. We intend to include more nuanced metrics to quantify these.

4 Adversarial Debiasing

4.1 Model

Our model consists of a multi-layer main neural network *N* whose primary objective is to predict recidivism $\hat{Y}$. We add an adversarial neural network *A* that penalizes our main recidivism neural network if demographic information can be predicted from the logit.

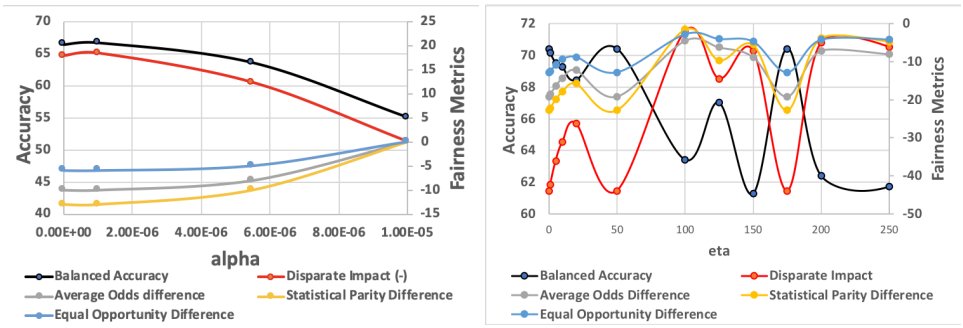

Figure 1: Metrics over varying α for Adversarial Debiasing and η for Bias Regularization

Both the main and adversary NN contain two hidden layers of size 20 with tanh activation followed by fully connected layer. The logits of the neural network *N* are fed to adversary *A*. Goal is for neural network *N* to predict $\hat{Y}$ i.e. recidivism accurately and for adversary *A* to predict demographic information *D* poorly. We use binary cross-entropy with logits losses for *N* and *A*. The losses of *N* and *A* are referred to as *Ly* and *Ld* respectively. To train *A*, we back-propagate *Ld* through *A*. Since we need *N* to be good at predicting $\hat{Y}$ while hiding *D*; we train our model N with the following loss function: $L = Ly - \alpha\, Ld$. Adversarial debiasing model was trained using tensorflow framework[4]. Adam optimizer was used for loss minimization.

4.2 Results and (Project) Future Work

We trained our model with the following hyperparameters: learning rate = 0.0001, # of epochs = 1000, α varying from 0.0 to 0.00001 (Figure 1). It was observed that adversary NN's accuracy was 81%; i.e. the adversary represents demographic information accurately. Also, we expect that *Ld* is independent of α. Furthermore, we satisfied all the fairness metrics when α was increased to about 0.00001; we expect small values of α to reduce bias since this encourages *N* to maximize *Ld*.

We plan to optimize the hyper-parameters of the model (namely learning rate, number of hidden units in *N* and *A*, and debias rate α). The adversarial debiasing approach was able to satisfy our fairness metrics. However, we need to explore approaches to maintain the initial accuracy, since it decreased significantly when we reduced bias (Table 3). Our current model also lumps Hispanic and African American individuals together as one underprivileged class. We plan to build a more complete model that will debias for multiple variables (such as in Section 5), perhaps with multiple adversaries.

5 Bias Regularization

For initial work, we used existing optimization code[3] and an AIF360[9] wrapper. This required lib modifications, which are not currently in the repository. We also found that the bias function only actually permits a single sensitive variable. For now, we resolved this by concatenating them (if sex,race are sensitive, then 'White,Male' becomes 'White-Male').

5.1 Model

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

$$\underset{w}{\operatorname{argmin}} \sum_i L(f(x_i), y_i) + \lambda ||w||_2 + \eta R(f(X), Y) \tag{1}$$

We used logistic loss for L, since this simple model worked well. Also, this model debiases on a per-class basis, instead of bucketing sensitive variables into 'privileged' and 'unprivileged'. For now, we report on our fairness metrics as a weighted average over underprivileged classes.

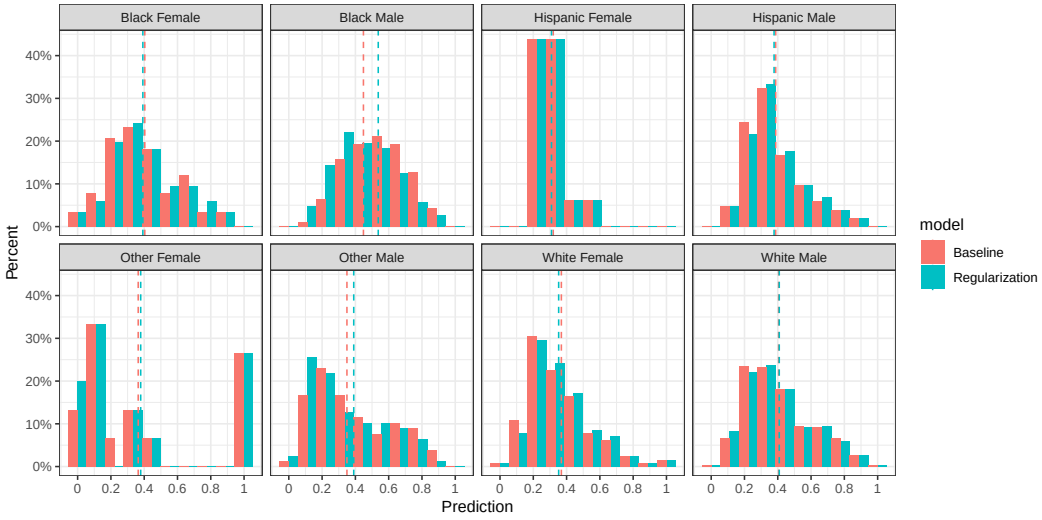

Figure 2: Prediction Distributions by class and model

5.2 Results and (Project) Future Work

Increasing η to about 100 nearly satisfied all the fairness metrics (Table 3). Curiously, increasing it further seemed to consistently hurt more than it helped (Figure 1. This is probably because the bias function is nonlinear, and is more likely to run into local minima at large magnitudes. This amounts to saying that $Pr(y|s = s_i) \approx Pr(y)$ 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 black males disproportionately receive high (≥ 0.8) scores (Figure 2) despite a class mean near the population after debiasing.

To fix this, we plan to add a bias penalty on discretized sublevel sets of $\hat{y}$. 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 would have corresponding penalty:

$$R(a, s_i) = \eta g(Pr(\hat{y}|s = s_i, \hat{y} \leq a), Pr(\hat{y})|\hat{y} \leq a)) \tag{2}$$

We can update this with SGD by caching $Pr(\hat{y}|\hat{y} \leq a)$ for each (a, s_i). In our case, race and sex on 10 deciles would create 5*2*10=100 variables, which should not be prohibitive. If successful, this lets us debias for $||Pr(\hat{y}|s = s_i, \hat{y} \leq a|| = Pr(\hat{y}|\hat{y} \leq a)$. If not, we will try other ideas.

6 Summary Table

Algorithm	Debiasing	Accuracy	Disparate Impact %	Avg. Odds Difference
Baseline, COMPAS only	None	65%	49.69%	-25%
Baseline	None	68%	61.42%	-21%
Reweighing	Yes[2]	67%	38.93%	-14%
Adversarial	None (alpha=0)	66.4%	17.85%	-10%
Adversarial	Yes[10] (alpha=1E-5)	55.1%	0%	0%
Regularization	No (eta=0)	70.4%	43.9%	-19.2%
Regularization (Race & Sex)	Yes[3] (eta=125)	64.9%	2.04%	-4.2%
Oracle	None	100%	47.6%	26%
Oracle	Reweighing	95.7%	0	0

Table 3: Debiasing over Race Variable

References

[1] 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.

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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