Debiasing Recidivism Predictions

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

May 5, 2019

1 Introduction

Recent AI research demonstrates that bias-agnostic approaches can cause greater negative impacts than overtly prejudicial systems. This is particularly concerning for 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[4] shows that these scores are biased against Black and Latino inmates, and large amounts of analysis 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.

2 Previous Work

Related research has identified three main categories of debiasing techniques: pre-processing, in-processing, and post-processing[8]. Pre-processing algorithms focus on removing problematic features and modifying weights before any training occurs. For example, reweighing[2] increases weights for positive outcomes and decreases them for negative ones to match the dataset mean. Post-processing techniques consider biased model output, usually by modifying thresholds or equalizing rates of false positives and negatives[5]. This approach is especially practical if the model and its input are not accessible[8]. Finally, in-processing attempts to debias the learning algorithm itself, typically by applying a bias penalty[3] analogous to l2 regularization. More recent approaches have used adversarial networks[9] 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. Most importantly, IBM AI Fairness 360[8] provides implementations for many of the algorithms and metrics we cite. The corresponding paper[7] was also invaluable as a resource for exploring previous work. We also use sk-learn[1] for basic ML models, such as logistic regression, and tasks, such as cross-validation.

3 Dataset and Features

We leverage a dataset (Table 1 with over 10,000 criminal defendants in Broward County, Florida. This includes two recidivism predictions from COMPAS software, which outputs decile scores for [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 $\phi(x_i)$ 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	Crime ID	Crime Description	COMPAS	Two Year Recid
1	Male	69	Other	0	F	Aggravated Assault...	3	No
3	Male	34	African-American	4	F	Felony Battery...	4	Yes
10	Female	39	Caucasian	0	M	Battery	1	No

Table 1: Subset of COMPAS Nonviolent Data Fields

Some categorical and text features are not explored much by previous work, and have no simple map to numerical representations. For example, Crime IDs have too many unique labels for one-hot encoding. It may be useful to cluster these, perhaps using sparse K-means from class. We may also compute a node embedding for crime descriptions, since this could expose new information with a bias correlation.

4 Baselines and Oracles

As this dataset is highly popular for studying the bias topic, our primary baseline will be the best performance we can achieve with existing software and methods. We will also present simple accuracy/fairness baselines for COMPAS scores alone and each considered debiasing method. Finally, we plan comparisons to existing work, especially analysis performed by ProPublica[4].

Our dataset is labeled, so a perfect oracle merely observes the y values. We also compute an oracle for the accuracy/fairness tradeoff curve by debiasing and training with y values as input. There is a significant gap between our baselines and this oracle, which we do not expect to close since our predictive power is limited. However, it helps to clearly illustrate that perfect accuracy and fairness cannot be achieved simultaneously here.

We will report out on baselines and oracles in a similar format to Table 3.

5 Evaluation Metrics

We will evaluate fairness using the following metrics: Disparate Impact, Average Odds Difference, Ranked Group Fairness, Statistical Parity Difference and Theil Index. We expect to discuss our results in terms of these properties. We have also explored more complicated metrics, some of which are software-implemented [8]:

Metric	Description
Disparate Impact	$\Pr[\hat{y} = 1 \mid u] / \Pr[\hat{y} = 1 \mid v]$
Statistical Parity Difference	$\Pr[\hat{y} = 1 \mid u] - \Pr[\hat{y} = 1 \mid v]$
Average Odds Difference	Compare True and False positive rates for u and v
Theil Index	Generalized entropy index
Ranked Group Fairness[6]	KL-divergence of $Pr[u]$ and $Pr[u \mid \hat{y} \geq p]$ over $p \in (0, 1)$

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

6 Scope

For this project, our primary goal is to develop incremental improvements and alternatives for two in-processing techniques. Specifically, we have some ideas for extending bias regularization[2] and adversarial debiasing[9] to comprehensive metrics like ranked group fairness[5]. We intend to compare our results to a pre-processing technique with the same objective[6].

7 Early Progress

We mapped the simple (numerical, binary, and low cardinality categorical) data to a feature set. Using this, 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[8] and 'Sex' as the protected variable. We also replicated and modified previous COMPAS analysis by ProPublica[4]. This was all done to validate our basic infrastructure and problem statement.

Dataset	Debiasing Method	Category	Accuracy	Disparate Impact %	Average Odds Difference
COMPAS Score Only	None	None	65%	49.69%	25%
Simple Features	None	None	68%	51.11%	22%
Simple Features	Reweighing[2]	Pre-	67%	3.74%	-2%
Oracle	None	None	100%	47.6%	26%
Oracle	Reweighing	Pre-	95.7%	0	0

Table 3: Debiasing over Sex 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] ProPublica. (May 2016). How we analyzed the compas recidivism algorithm, [Online]. Available: `ttps://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

[5] 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`.

[6] 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`.

[7] 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`.

[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. 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`.

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