

---

# Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction

---

Christina Wadsworth

Stanford University  
Stanford, CA  
cwads@cs.stanford.edu

Francesca Vera

Stanford University  
Stanford, CA  
fvera@cs.stanford.edu

Chris Piech

Stanford University  
Stanford, CA  
piech@cs.stanford.edu

## Abstract

Recidivism prediction scores are used across the USA to determine sentencing and supervision for hundreds of thousands of inmates. One such generator of recidivism prediction scores is Northpointe's Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) score, used in states like California and Florida, which past research has shown to be biased against black inmates according to certain measures of fairness. To counteract this racial bias, we present an adversarially-trained neural network that predicts recidivism and is trained to remove racial bias. When comparing the results of our model to COMPAS, we gain predictive accuracy and get closer to achieving two out of three measures of fairness: parity and equality of odds. Our model can be generalized to any prediction and demographic. This piece of research contributes an example of scientific replication and simplification in a high-stakes real-world application like recidivism prediction.

## 1. Introduction

Machine learning models and other data-based algorithms are being used increasingly in decision-making processes that affect individual lives. No longer is accuracy the only concern when developing models – fairness must be taken into account as well. Criminal risk assessment (Angwin et al., 2016), salary prediction (BBC, 2018), and loan approvals (Swarns, 2015) are examples of cases where existing societal biases against a certain gender or race can be perpetuated by machine learning or other algorithms and existing discrimination.

Much of recent discussion surrounding fairness has revolved around Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a risk assessment and recidivism prediction score developed by Northpointe and

used widely across the United States to affect sentencing and supervision on an individual level. As it stands, COMPAS is essentially a “black box” since Northpointe has not publicly released their algorithm, only hinting at important factors and the use of part of a 137-question survey that includes sections such as: Criminal History, Social Environment, Criminal Personality, and Criminal Attitudes (Angwin et al., 2016). Angwin et al. (2016) published a report on ProPublica that claimed COMPAS is biased against black inmates based on a study conducted using data from Florida inmates. COMPAS performs similarly in terms of accuracy for white and black inmates, but the errors COMPAS makes indicate discrimination against black inmates, according to Angwin et al. For example, a black inmate who does not re-offend is more likely to be classified as “high risk” than a white inmate who does re-offend. Black inmates in general are almost three times more likely to be classified as “high risk” than white inmates. Northpointe responded to the ProPublica analysis by arguing that COMPAS satisfies the fairness metric of calibration (Dieterich et al., 2016).

Bias against black inmates can be learned in neural networks that predict recidivism as well. Even if race is not an input feature, other features are correlated with race. For example, an inmate's number of priors and previous time in jail are correlated with race because black inmates are more likely to be jailed for the same crimes as white inmates and are arrested at a higher rate for less serious crimes (Williams, 2016; Fenton, 2016).

**Contribution** By adding an adversary to a network that predicts recidivism, we counteract racial biases found in criminal history datasets. Our adversary penalizes our recidivism prediction network if race is predictable from the recidivism prediction. Our model is generalizable to almost any prediction and any demographic.

## 2. Related Work

*Fairness Definitions* The three types of fairness often used in fairness research, which are also used in this paper, are: demographic parity, equality of odds, and calibration. Hardt

et al. (2016) and Kleinberg et al. (2016) provide definitions for these types of fairness and discuss their trade-offs. In the context of predicting recidivism, they are defined:

**Parity** A score  $S = S(x)$  satisfies parity if the proportion of individuals classified as high-risk is the same for each demographic.

**Equality of Odds** A score  $S = S(x)$  satisfies equality of odds if the proportion of individuals classified as high-risk is the same for each demographic, when true future recidivism is held constant. White and black inmates that do recidivate should have the same proportion of high risk classification.

**Calibration** A score  $S = S(x)$  is calibrated if it reflects the same likelihood of recidivism irrespective of the individual's demographic. In this application, black inmates who are classified as high risk should have the same probability of true recidivism as white inmates classified as high risk.

*Adversarial Fairness* Past research has explored the use of adversarial networks to achieve fairness. Beutel et al. (2017) used an adversary on a shared hidden layer to satisfy parity for salary prediction, and Ganin et al. (2016) used a domain classifier and reverse gradient on a hidden layer to remove domain correlation. Recently, Zhang et al. (2018) used a predictor and adversary with an additional projection term to satisfy parity or equality of odds with word embeddings and predicted salary. Zhang et al. are the only existing work that uses the output layer of their predictor as input to the adversary, which was a departure from the conventions we noted earlier. No published adversarial research thus far has looked at recidivism. We replicate and simplify Zhang et al.'s model to explicitly demonstrate how adversarial models that achieve fairness can have real-world applications.

*Fairness and COMPAS* There is much debate among researchers who have studied COMPAS as to which definitions of fairness should be satisfied for recidivism predictions. Corbett-Davies et al. (2017) argue that because black inmates are arrested at a higher rate than white inmates, there should be some disparity of recidivism prediction for white and black inmates. Other researchers investigate COMPAS in relation to priors since black inmates have more priors (Chouldechova & G'Sell, 2017). We challenge these notions: if black people are more likely to become incarcerated in the US than white people when controlling for criminal behavior (Bridges & Crutchfield, 1988), black inmates should not be punished for our biases with harsher recidivism predictions.

*Methods to Improve Fairness on COMPAS* Past research has attempted to improve racial fairness within COMPAS according to the metrics of false positive and false negative differences across black and white inmates. Hardt et al. (2016) used post-processing on standard logistic regressor that picks different thresholds for different groups and at times adds randomization. Zafar et al. (2017) first tried

to enforce equal False Positive and False Negative rates for groups by introducing penalties for misclassified data points during training. They then optimized by using the covariance between sensitive attributes and distance between feature vectors of misclassified samples and classifier boundary. Bechavod et al. (2017) used a logistic regressor while simultaneously penalizing unfairness by penalizing the difference in the average distance from the model's decision boundary across values of the protected attribute. They use an Absolute Value Difference (AVD) model and a Squared Difference (SD) model. We compare our model to these existing methods that aim to improve fairness for COMPAS. Other research has been done to debias COMPAS, focusing on debiasing input features. Lum et. al. propose univariate transformations of input features to achieve independence from the protected demographic, then use random forest to predict unbiased outputs (2017). Johndrow et. al. apply this model to recidivism prediction (2016). Ludrow et. al. transform the input dataset so that predictability of the protected demographic is impossible before classification. We contribute to this work by creating a model that optimizes for an unbiased output, as opposed to optimizing for unbiased input features. Additionally, we present separate models to satisfy multiple definitions of fairness.

### 3. Adversarial Model

#### 3.1. Model Structure

We start with a multi-layer neural network  $N$  that outputs a probability of recidivism, denoted  $\hat{Y}$ .  $\hat{Y}$  is our primary prediction objective, and  $N$  is our baseline. We now want our output  $\hat{Y}$  to satisfy demographic parity or equality of odds for demographic  $D$ . Even if  $D$  is not an input feature to our neural network, it may be correlated with other features, from which the network can learn a bias.

**Demographic Parity Model:** We input the *logit* from  $N$  (the unnormalized predicted recidivism probability, i.e. just before the sigmoid) to an adversarial neural network  $A$  that learns to classify demographic  $D$ . If  $\hat{Y}$  is biased for demographic  $D$ ,  $A$  should learn to have a high accuracy because the *logit* will be highly predictive of  $D$ . Our goal is for neural network  $N$  to predict  $\hat{Y}$  accurately and for  $A$  to predict  $D$  poorly.

**Equality of Odds Model:** We can input the *logit* as well as the true recidivism value  $Y$ . Here the goal is for there to be no difference in  $\hat{Y}$  across demographic, given true recidivism value  $Y$ , which will be satisfied when  $A$  cannot predict  $D$  with high accuracy given  $Y$  and the *logit*.

#### 3.2. Model Training

Our goal is for neural network  $N$  to predict  $\hat{Y}$  accurately and for  $A$  to predict  $D$  poorly. If we achieve this, our model will be outputting an unbiased, accurate  $\hat{Y}$ . More specifically, we use binary cross-entropy losses for  $N$  and  $A$ , which we refer to as  $L_y$  and  $L_d$ , respectively. To train

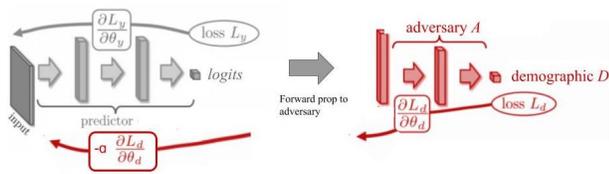


Figure 1. Diagram of our adversarial model structure.

$A$ , we back-propagate  $L_y$  through  $A$ . However, we need to train  $N$  to be good at predicting  $\hat{Y}$  and bad at predicting a *logit* that is highly correlated with  $D$ . If we subtract  $L_d$  from  $L_y$ ,  $N$  will be encouraged to maximize  $L_d$ , which will produce a *logit* that cannot be used to predict race and  $\hat{Y}$  values that are closer to achieving parity. We train our model  $N$  with the following loss function:

$$L = L_y - \alpha * L_d$$

### 4. Experiments and Results

**Data** We apply our adversarial model to recidivism prediction. To do so, we used public criminal records data from Broward County, Florida that was compiled and published by ProPublica. Much of recidivism research in the past two years has been conducted on this dataset. The dataset also includes COMPAS scores for Broward County inmates, so we are able to compare our results to the performance of COMPAS. Our training set is size 8230 and our test set size 2213. We only use data from white and black inmates, of which 41% represents white inmates and 59% black inmates. Beutel et al. (2017) showed that a more obviously skewed distribution on demographic  $D$  can affect how helpful the adversary is. We found that despite our slight skew, the adversary was just as effective.

**Model** Predictor  $N$  has 2 256-unit ReLU hidden layers. Adversary  $A$  has a single 100-unit ReLU hidden layer. We used a learning rate of  $e^{-4}$ , binary cross entropy loss, a sigmoid output layer, an Adam optimizer, and an alpha value of 1. To settle on these hyper parameters, we tuned a number of hidden layers and hidden layer size for both the predictor and adversary, alpha, and learning rate. Figure 2 shows tuning for number and size of hidden layers for  $N$ . When tuning alpha, we wanted to maximize  $L_d$ , then minimize  $L_y$ . This allows for fairness to be satisfied first, then for our predictor  $N$  to be as accurate as possible.

**Metrics** To evaluate accuracy, we use area under ROC curve. We also define metrics that can be used to compare demographic parity and equality of odds:

$$\begin{aligned} \text{High Risk Gap: } & |HighRisk_{white} - HighRisk_{black}| \\ \text{False Positive Gap: } & |FP_{white} - FP_{black}| \\ \text{False Negative Gap: } & |FN_{white} - FN_{black}| \end{aligned}$$

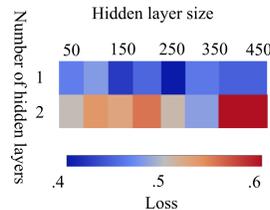


Figure 2. Model structure tuning.

If High Risk Gap is zero, demographic parity is satisfied. False Positive Gap and False Negative Gap are used to assess equality of odds. If both are zero, equality of odds is satisfied. Also included are conditional probability graphs to compare evaluations on the fairness metric calibration.

**Experiments** Outlined are the models used to conduct experiments on our COMPAS dataset for recidivism prediction: **Recidivism Prediction:** We trained a regular recidivism predictor without any type of adversary to compare our adversarial models to a baseline and to confirm that bias is perpetuated in machine learning models.

**Adversarial Models:** We then trained two variants of the adversarial recidivism predictor. One adversary accepted the *logit* as input and the other accepted the *logit* and true recidivism value  $Y$ . We tried an additional variation of the model that accepted a hidden layer as input instead of the *logit*. This model was less stable and required more hyper parameter tuning to see a decent result. As such, we continued most of our research with our models that accept the *logit* because the adversary was just as powerful and less finicky. In our results section, we present the model that accepted just the *logit* as input as our adversarial model.

| MODEL                        | HIGH RISK GAP | FN GAP      | FP GAP      |
|------------------------------|---------------|-------------|-------------|
| COMPAS SCORES (OUR TEST SET) | 0.18          | 0.22        | 0.17        |
| OUR RECIDIVISM MODEL         | 0.21          | 0.27        | 0.15        |
| OUR CHOSEN ADVERSARIAL MODEL | <b>0.02</b>   | <b>0.02</b> | <b>0.01</b> |

Table 1. Comparison of High Risk Gap, False Negative Gap (FN Gap), and False Positive Gap (FP Gap), across models.

| MODEL                        | AUC  |
|------------------------------|------|
| COMPAS SCORES (OUR TEST SET) | 0.66 |
| OUR RECIDIVISM MODEL         | 0.72 |
| OUR CHOSEN ADVERSARIAL MODEL | 0.70 |

Table 2. Comparison of AUC for ROC curves across models.

**Results** Our regular recidivism predictor is similarly biased against black inmates with respect to parity and equality of odds, which suggests that biases can be learned and perpetuated by machine learning models. However, the results in Table 1 show that our chosen adversarial model is

| MODEL                                     | ACCURACY | FP GAP      | FN GAP      |
|---|----------|-------------|-------------|
| COMPAS SCORES (OUR TEST SET)              | 0.68     | 0.17        | 0.22        |
| OUR RECIDIVISM MODEL                      | 0.70     | 0.15        | 0.27        |
| OUR CHOSEN ADVERSARIAL MODEL              | 0.70     | <b>0.01</b> | <b>0.02</b> |
| BEHAVOD ET AL. AVD PENALIZERS (2017)      | 0.65     | <b>0.02</b> | <b>0.04</b> |
| BEHAVOD ET AL. SD PENALIZERS (2017)       | 0.66     | <b>0.02</b> | <b>0.03</b> |
| BEHAVOD ET AL. VANILLA REGULARIZED (2017) | 0.67     | 0.20        | 0.30        |
| ZAFAR ET AL. (2017)                       | 0.66     | <b>0.03</b> | 0.11        |
| ZAFAR ET AL. BASELINE (2017)              | 0.66     | <b>0.01</b> | 0.09        |
| HARDT ET AL. (2016)                       | 0.65     | <b>0.01</b> | <b>0.01</b> |

Table 3. Comparison of accuracy, False Positive Gap (FP Gap), and False Negative Gap (FN Gap) across models.

much closer than COMPAS to satisfying parity and equality of odds. Further, from Table 2, we see that our adversarial model has improved accuracy compared to COMPAS scores. On the other hand, Figure 3b demonstrates that COMPAS satisfies calibration, whereas some bias is evident in our model (Figure 3a), especially at the threshold cutoff 0.5, which is the cutoff for recidivism prediction – although our bias is slight.

When comparing feature importance for our recidivism prediction baseline model and our adversarial model, we notice that the adversarial model relies more heavily on 6 of the top 10 most important features (Figure 4). This indicates that the adversarial model depends on more holistic information while the regular recidivism model is mostly dependent on charge degree, age, and priors. Recall that black inmates have more priors because black inmates are jailed for crimes more often than white inmates are (Fenton, 2016). Additionally, charge degree distributions are different across races. It is possible that relying on these three features contributes to the racial bias found in the baseline model.

We further assess our model in the context of other work done to debias COMPAS. Most existing work compares false positive and false negative values. As shown in Table 3, our adversarial model is just as good at debiasing with regards to race as state of the art work done on COMPAS is.

Overall, our model outperforms COMPAS scores in terms of accuracy and comes close to satisfying parity and equality of odds. We also achieve false positive and false negative differences that are on par with the best of fairness research done on COMPAS. However, we note that our model is slightly more biased than COMPAS with regards to calibration.

### 5. Case Study

The quantitative COMPAS scores impact real individuals. In a case study on two Broward County inmates, we compare COMPAS predictions to the results of our adversarial model and investigate the stories of Joe and Bob.

Despite having only 1 prior and 2 charges before COMPAS, Joe, a 55 year old black inmate, received a COMPAS score of 8 out of 10, labeling him as “high risk” and likely to

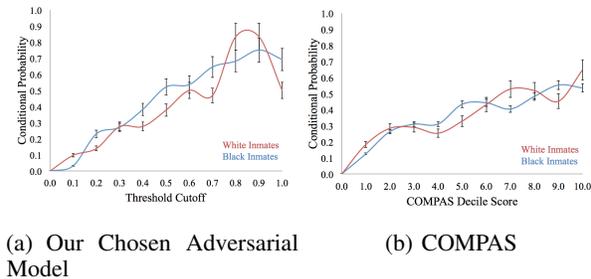


Figure 3. Conditional probability across models.

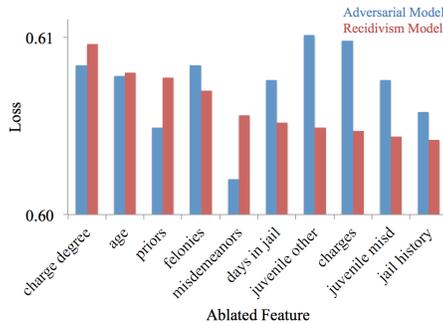


Figure 4. Feature importance.

recidivate. Our adversarial model disagreed by making a prediction of 0.05 after the sigmoid layer. Presently, Joe has no charges since the COMPAS score was assigned and has not recidivated.

In contrast, despite having 13 priors and 24 charges before COMPAS, Bob, a 27 year old white inmate, received a COMPAS score of 5 out of 10 – a “medium risk” score that disagrees with our prediction of 0.84 – a score indicating Bob will recidivate. Since receiving his COMPAS score, Bob has been charged 6 times, including for disorderly intoxication and aggravated battery (an act of violent recidivism). Had our model been used, perhaps his fate would have been different due to a higher level of supervision. It is important to take into account how COMPAS and other similar models are being used to affect real lives on an individual level. A general racial bias towards black people means that individual black people are being punished.

### 6. Conclusion

We have shown that it is possible to reduce the bias of a machine learning model that is trained on demographically biased data. We also demonstrate a general method for training unbiased models that can enforce constraints for multiple definitions of fairness. Our adversarial models are less biased than the original COMPAS scores and our recidivism prediction baseline, but the models still outperform COMPAS in terms of accuracy, providing the first piece of research on recidivism and COMPAS that achieves this with adversarial learning.

## Acknowledgments

Thanks to Ramtin Keramati for his helpful advice and Jerry Cain for his continued support of this research.

## References

- Gender pay gap: Men still earn more than women at most firms. BBC, 2018.
- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. ProPublica, 2016.
- Bechavod, Y. and Ligett, K. Learning fair classifiers: A regularization-inspired approach. *CoRR*, abs/1707.00044, 2017.
- Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.
- Bridges, G. and Crutchfield, R. Law, social standing and racial disparities in imprisonment. *Social Forces*, 1988.
- Chouldechova, A. and G'Sell, M. Fairer and more accurate, but for whom? *FATML Poster Session*, 2017.
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230, 2017.
- Dieterich, W., Mendoza, C., and Brennan, T. Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpointe, 2016.
- Fenton, S. Black people much more likely to be jailed over criminal offences than white people, research suggests. The Independent, 2016.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *JMLR*, 17:1–35, 2016.
- Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.
- Johndrow, J. and Lum, K. A statistical framework for fair predictive algorithms. *arXiv preprint*, 2016.
- Kleinberg, J. M., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016.
- Lum, K. and Johndrow, J. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint*, 2017.
- Swarns, R. L. Biased lending evolves, and blacks face trouble getting mortgages. The New York Times, 2015.
- Williams, T. Marijuana arrests outnumber those for violent crimes, study finds. The New York Times, 2016.
- Zafar, M. B., Valera, I., Rodriguez, M. Gomez, and Gummadi, K. P. Fairness beyond disparate treatment &#38; disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW '17*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017. ISBN 978-1-4503-4913-0. doi: 10.1145/3038912.3052660.
- Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018.