Investigating discrimination bias in predictive modelling

Sara Thiringer   Jonathan Rittmo

26/10/2020

Background: The Problem

In recent years, many scandalous examples have shown that statistical models trained on large amounts of data can "act" discriminatory. Examples include:

- Adds of high-income jobs being shown less frequently to women, presumable becasue they've been predicted to be less interested or suitable[1]
- Black people's health status being underestimated, leading to inappropriate healtch care measures[2]
- Black people begin predicted a higher risk for crime recidivism, leading to higher penalties[3]

[1] Datta, Tschantz, and Datta (2015)
[2] Obermeyer et al. (2019)
[3] ProPublica (2016)

Project Aims

- How can we quantify fairness in order to be able to evaluate algorithmic fairness?
- What methods are available to increase algorithmic fairness? In what type of situations do they apply? (i.e. In what kind of situations can we expect them to be successful?)

Background: Why Discrimination Bias?

- Correlation between outcome y and protected charateristic x_p
- Correlation between important predictors $\boldsymbol{x}_i$ and protected carachteristic x_p
- Under/over sampling of groups with protected carachteristic x_p

Possible Solutions

Pre-Processing	Training	Prediction
Resampling	Penalty	Threshold
Mapping	Model bias	adjustments
Altering labels	Tuning for fairness	Alter predictions

We've chosen to work with resampling and threshold adjustment.

Possible Goals

Demographic parity

$$P(Y = 1|X = 1) = P(Y = 1|X = 0)$$

Equalized odds

$$P(G = 1|X = 0, Y = 1) = P(G = 1|X = 1, Y = 1)$$

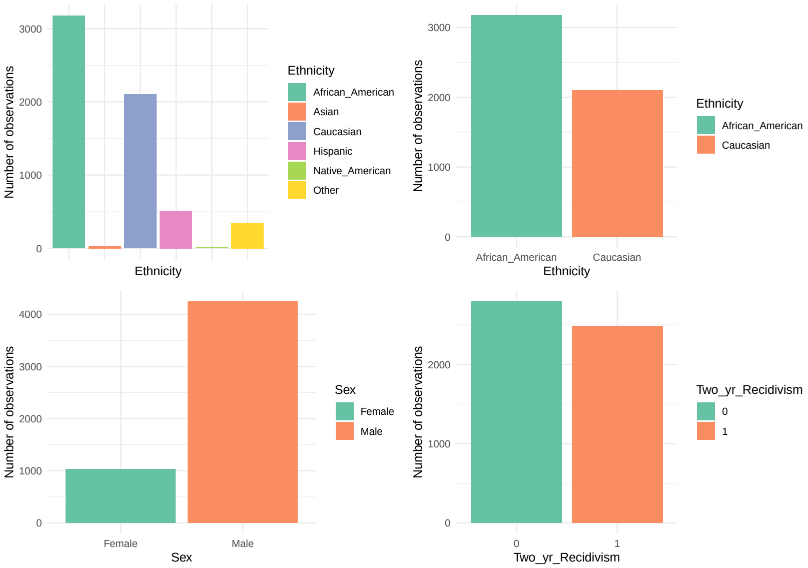
Data

The COMPAS[4]-data

Variable	Type	Values
Two_yr_Recidivism	Factor	1 / 0
Number_of_Priors	Numerical	Mean (sd) : 3.5 (4.9)
Above45	Factor	1 / 0
Below25	Factor	1 / 0
Misdemeanor	Factor	1 / 0
Ethnicity	Factor	African_American / Caucasian
Sex	Factor	Female / Male

[4] Correctional Offender Management Profiling for Alternative Sanctions

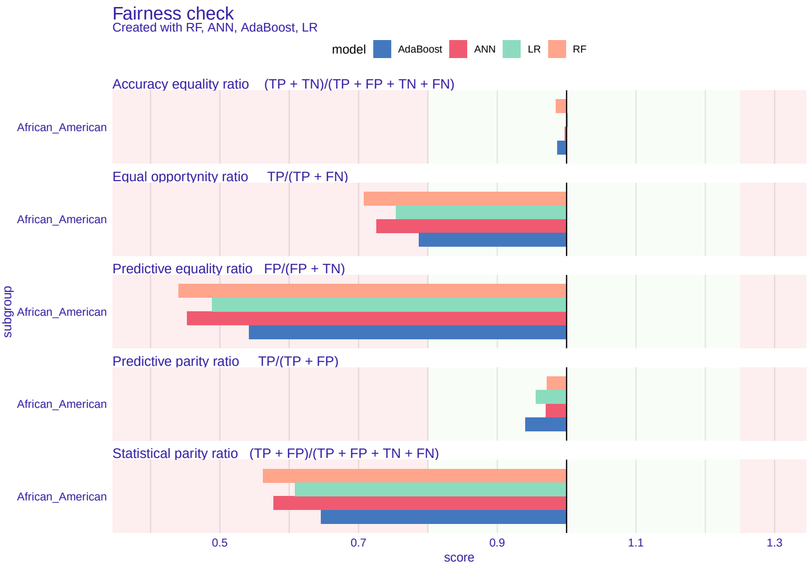
Descriptives

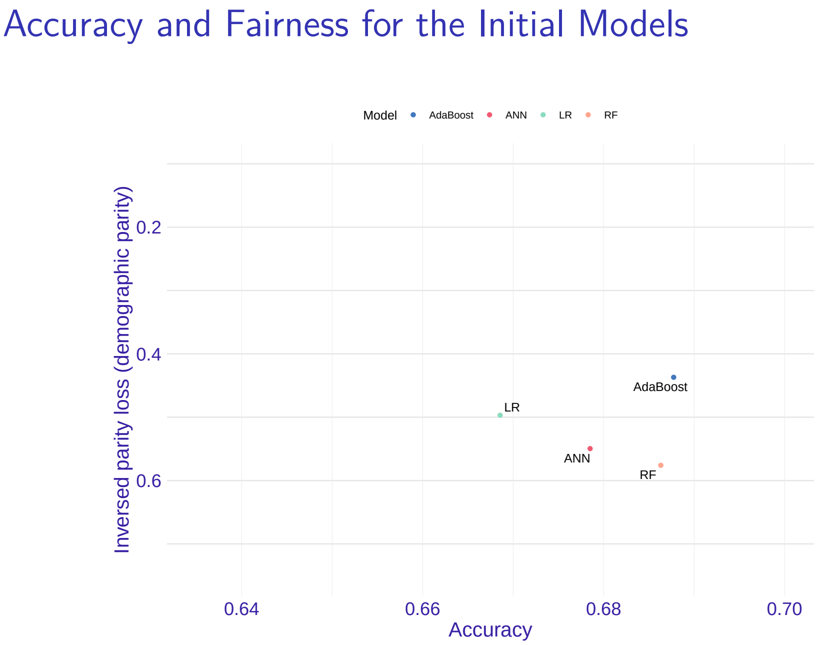


Models

Model	**Tuning**
Random Forest	Predictors at each split
Artificial neural net	Number of hidden nodes
Logistic ridge regression	Penalisation
K-nearest neighbour (left out)	Number of neighbours
AdaBoost	Predictors at each split

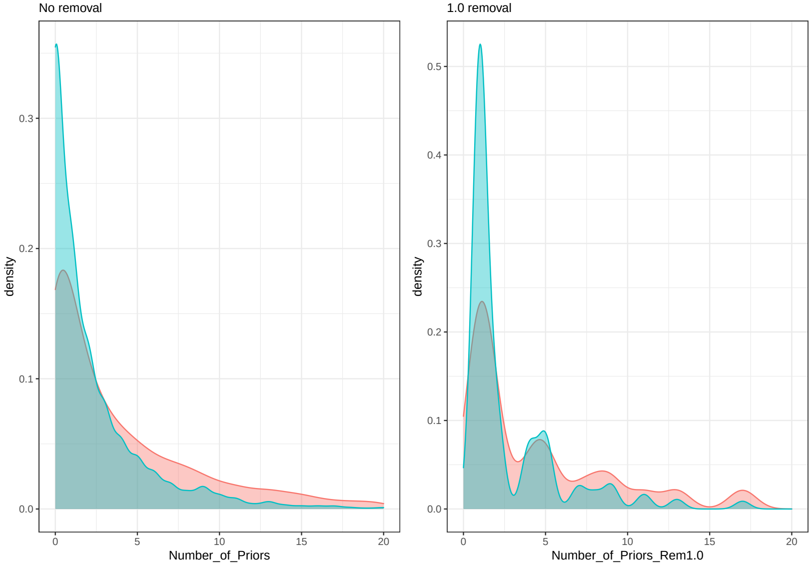
Evaluation of the Initial Models




Accuracy and Fairness for the Initial Models
Model
AdaBoost
ANN
LR
RF
Inversed parity loss (demographic parity)
0.2
0.4
0.6
AdaBoost
LR
ANN
RF
0.64
0.66
0.68
0.70
Accuracy

Disparate Impact Removal

Removes differences between groups while preserving it within the groups. Example: Number of priors.



Resampling

Using undersampling and oversampling to even out inequalities between the depraved and privileged groups having positive and negative outcome attributes respectively.

Table 4: Joint distribution of Ethnicity and Recidivism

	African_American	Caucasian
0	1514	1281
1	1661	822

Note: 1 = Recidivism

Uniform Resampling

Aim: Make the joint distribution of `Ethnicity` and `Two_yr_Recidivism` uniform by duplicating some observations and removing others.

Table 5: Joint distribution of Ethnicity and Recidivism, uniform resampling

	African_American	Caucasian
0	1332	905
1	1184	803

Note: 1 = Recidivism

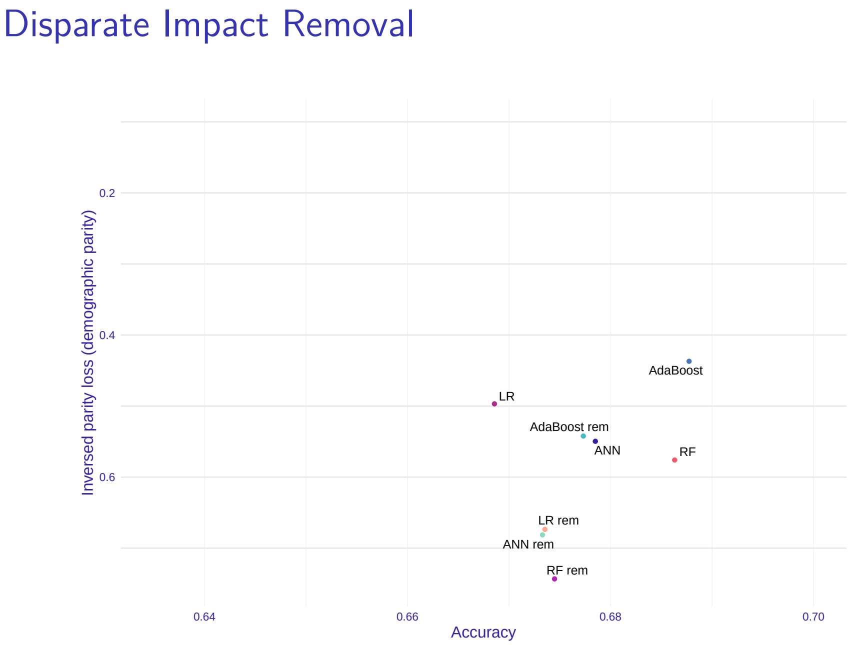
Preferential Resampling

Unequal probability sampling where probabilities are determined by fitting a logistic regression model on the outcome variable. Borderline observations are skipped or duplicated more often. Result is the same as for uniform.

Table 6: Joint distribution of Ethnicity and Recidivism, preferential resampling

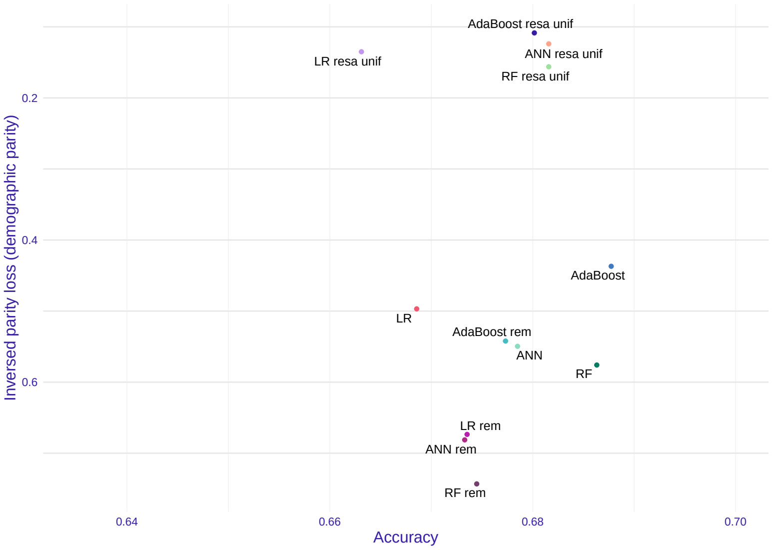
	African_American	Caucasian
0	1332	905
1	1184	803

Note: 1 = Recidivism

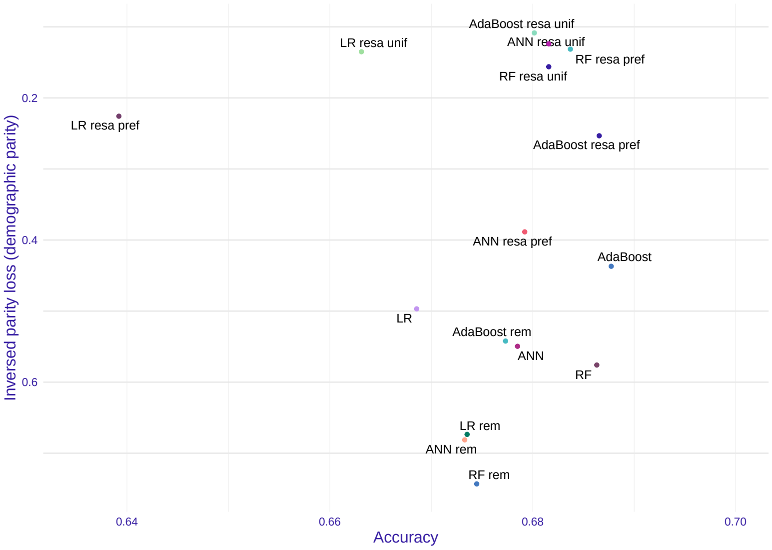


Disparate Impact Removal
Inversed parity loss (demographic parity)
0.2
0.4
0.6
Accuracy
0.64
0.66
0.68
0.70
AdaBoost
LR
AdaBoost rem
ANN
RF
LR rem
ANN rem
RF rem

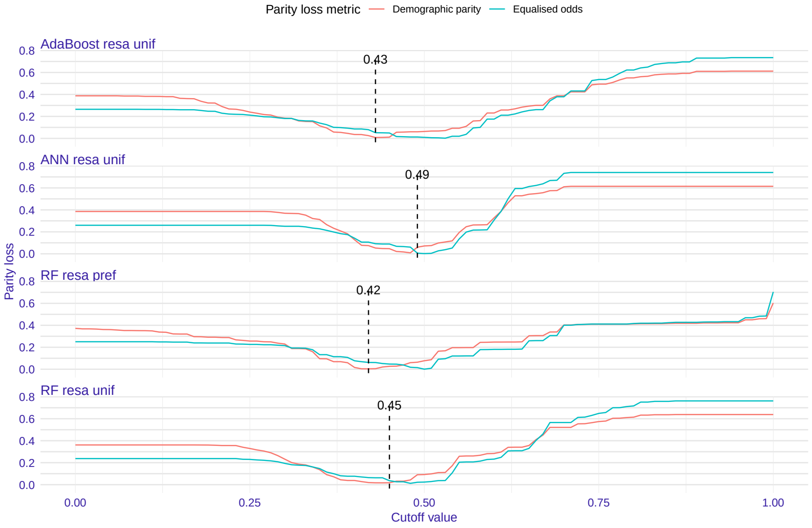
Uniform Resampling



Full Comparison

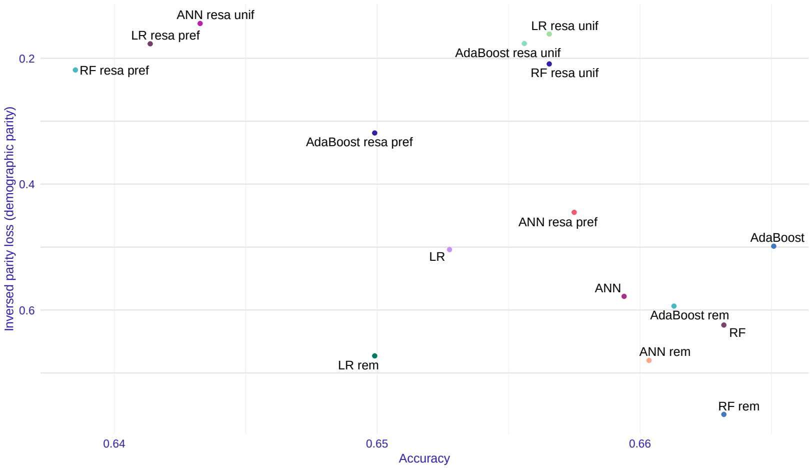


Threshold Adjustment



Comparing Models with Different Thresholds

Evaluation on Test Data - Comparison



Final Model Performance

- The resampling methods seems to have best effect on minimising parity loss while also preserving accuracy rates
- Random Forest looked most promising, but while tested on new data Logistic Regression, Artifical Neural Network and AdaBoost were equally strong competitors (all resampled)

Table 7: Performance

recall	precision	accuracy	auc
0.733	0.658	0.657	0.706

Final Conclusions

- It is possible to build models who satisfy some fairness criteria without a too large drop in accuracy
- Decisions for fair models include: fairness measure, evaluation metrics and choice of methods
- Less complex methods can be found amongst resampling and threshold adjustment
- As always, what type of data we are dealing with will largely impact the results of different methods (for example why disparate impact remover didn't work)
- Bias investigation simulteaneously adds and decreases complexity

References

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. 2015. “Automated Experiments on Ad Privacy Settings.” *Proceedings on Privacy Enhancing Technologies* 2015 (1): 92–112. https://doi.org/https://doi.org/10.1515/popets-2015-0007.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366 (6464): 447. https://doi.org/10.1126/science.aax2342.