Danmarks
Tekniske
Universitet

Responsible AI
Project 1 Fairness

Malthe Jelstrup - s184291
Christian Ole Nielsen - s204131

September 16, 2022

1 Introduction

This report will be a diagnostic analysis of bias present in the Catalan juvenile recidivism dataset, this is followed by an attempt at bias mitigation and lastly another analysis is performed to understand the bias still present or introduced through the bias mitigation techniques[1]. The BiasBalancer package will be utilized for the diagnostic analysis and visualisations, the Fairlearn package will be used for its bias mitigation techniques and code for the analysis is available on github[2][3][4].

2 Diagnosis & Mitigation

The above mentioned dataset follows all young offenders who in 2010 ended their current involvement with the juvenile justice system in Catalonia. The data describes whether the persons committed any further crimes up until the end of 2015. The data used in this analysis is a randomly chosen subset of the original data.

2.1 Diagnosis

A model was trained, that attempts to predict whether a person will reoffend based on the available data. To function as a baseline for this project the model is a simple MLPClassifier from the sklearn package, with default settings with the exception of maximum iterations increased to 4000 and early stopping increased to 800, the prediction threshold was set to the proportion of reoffenders in the training dataset, as seen below in equation 1; where **#A** is the number of reoffenders in the training set and **N** is the size of the training set.

$$\text{pred}_i = \begin{cases} 0 & \text{pred score}_i < \dfrac{\#A}{N} \\ 1 & \text{pred score}_i >= \dfrac{\#A}{N} \end{cases} . \tag{1}$$

Utilizing the BiasBalancer package it becomes apparant that the model is unfair in more ways than one illustrated in Figure 1 below. Even though independence is highlighted as the largest problem in the "Unfairness barometer" that is not necessarily a desired criterion in this particular case, given that simply equalizing the rate $\mathbb{P}\{\hat{Y} = 1\}$ would mean equalizing the recidivism rate across gender, which may not be true. Instead

separation was chosen as the criterion to attempt to satisfy through mitigation. This means equalizing the error rates $\mathbb{P}\{\hat{Y} = 0|Y = 1\}$ and $\mathbb{P}\{\hat{Y} = 1|Y = 0\}$ across genders. That essentially incentives reduction of errors for both sensitive groups simultaneously.

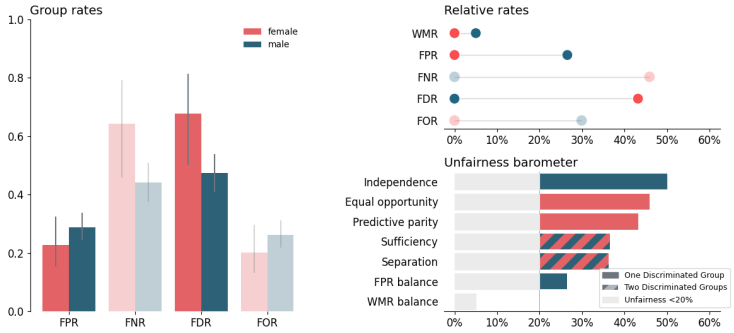


Figure 1: Overview of unfairness pre mitigation

2.2 Mitigation

As previously mentioned the mitigation technique used in this project is from the Fairlearn package[3]. This package includes a function **ThresholdOptimizer** that works as a post-processing step to adjust a predictors output from training data. It also has parameters to choose a constraint to satisfy during the optimization, in this case separation. The resulting model predictions with the new thresholds can then be analysed in the same manner as above, to identify how effective the mitigation strategy was. Figure 2 below shows significant improvements in separation and as a by product also in independence. However, this naturally comes with an almost equally significant loss in sufficiency, due to the two types of fairness being mutually exclusive.

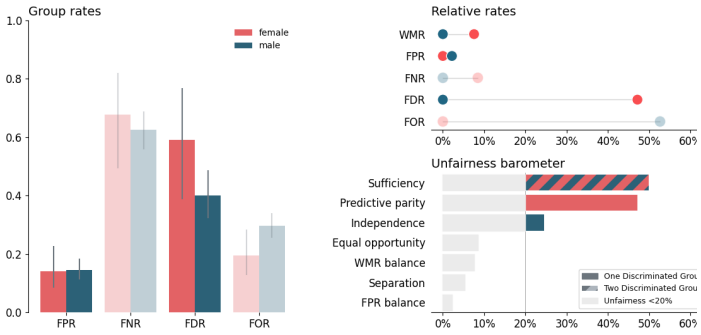
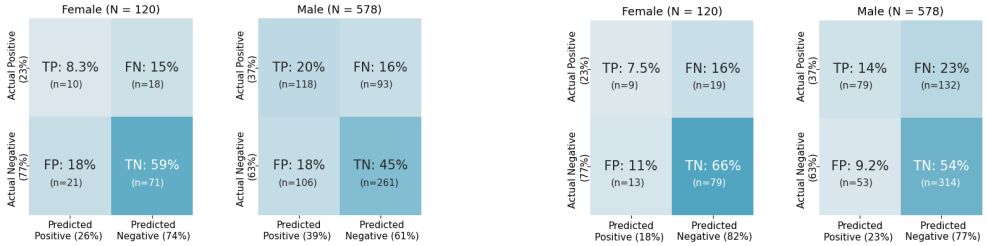


Figure 2: Overview of unfairness post mitigation

2.3 Performance

To evaluate model performance and compare pre and post mitigation performance the resulting confusion matrices will be evaluated. When it comes to the confusion matrix, the goal with the implemented mitigation strategy was to bring the true positive and false positive rates closer together between the groups, ideally make them equal. This goal was achieved to some extend, as can be seen in figure 3, seeing as the true positive rates are closer together and the false positive rates are not much further apart between the groups. However, this comes at the cost of a slightly lower true positive rate and a higher false negative rate, but conversely a lower false positive and higher true negative rate. From this it can be derived that, the model predictions now tend more towards false than true. In this particular case it could be argued that a false positive impacts an individual and society more strongly than a false negative would, which in turn would make the prediction trend a positive overall. As motivation behind the choice of which fairness criterion to aim to satisfy, the claim was made that it "[...] essentially incentives reduction of errors for both sensitive groups simultaneously". Looking at the wrong predictions both per individual group and aggregate, this claim appears to hold. The false positive and false negative predictions pre mitigation sum to $21 + 18 + 106 + 93 = 238$ whereas summing the same prediction categories for the post mitigation confusion matrix results in $13 + 19 + 53 + 132 = 217$, an overall net reduction in wrongly predicted individuals in the validation set.



(a) Pre mitigation performance

(b) Post mitigation performance

Figure 3: Confusion matrix for model prediciton

3 Conclusion

The bias mitigation technique had its intended effect of satisfying the separation criterion relatively well, and caused the theorized additional effect of lowering false predictions in both groups. The post mitigation predictions suffers from lack of sufficiency, which is to be expected given that the mitigation was an attempt to satisfy separation which is mutually exclusive with sufficiency. As well as causes the intended fairness effects, introducing the bias mitigation efforts, resulted in predictions that would be preferable given the context, due to the type of wrong predictions, where false positives are arguably significantly more damaging than false negatives.

References

[1] M. F. P. B. F. F. Ă. R. C. J. M. E. A. B. M. B. L. I. Marta Blanch Serentill, Manel Capdevila Capdevila, "Recidivism in juvenile justice," 2017.

[2] E. Z. Caroline Amalie Fuglsang-Damgaard, "Responsible ai project 1." `https://github.com/elisabethzinck/Fairness-oriented-interpretability-of-predictive-algorithms`, 2022.

[3] F. Org, "Fairlearn," 2022.

[4] M. A. L. J. Christian Ole Nielsen, "Fairness oriented interpretability of predictive algorithms." `https://github.com/Stinth/ResponsibleAI`, 2021.