jhc21

1. Introduction and Model Decisions

Fairness in AI is a multifaceted issue that is the subject of ongoing work on its definitions and its impact on the models produced. Fairness is highly aligned with the interpretability and the transparency of AI systems, and of the process with which those models were created. The idea of fairness in AI is a fast growing and wide ranging active field of research; every year more practitioners question the implications of using AI to aid in the decision-making process. This report aims to perform a high level study on some commonly interrogated datasets, with the aim of digging into the data and discovering the benefits and drawbacks of using certain metrics and preprocessing methods.

1.1. Modelling Decisions

For the purposes of this task a logistic regression model was used from the scikit learn package. This model was chosen as the relatively small size of the dataset does not warrant an SVM, and classification accuracy's are sufficiently good after fitting the model. This means that the hyperparameter to tune is the value of C in the model, i.e. the inverse of regularization strength for logistic regression. The numbers reported in all tables show the accuracy and fairness when evaluated against a held back test set after the 'best' model by hyperparameter C is selected from evaluation across K folds on the training data. The metric used to evaluate fairness is Equality of Opportunity difference which measures the relative acceptance rates of the qualified applicants from each of the groups (privileged and unprivileged). For the purposes of this experiment, the 'sex' feature was used throughout as it demonstrated to give the most steady results (particularly in the case of the German dataset which is significantly smaller than the adult one). 'race' was a reasonable alternative feature although it is inherently non binary and therefore the preprocessing module reduces this to 'white' and 'non-white', it was decided that using 'sex' would be the clearest way to analyse algorithmic bias. In our experiments, to obtain more reliable estimates of accuracy and fairness, we repeatedly split each dataset into a train (70%) and test (30%) set 5 times and report the average statistics for accuracy and fairness.

2. Task 1 - Standard Model

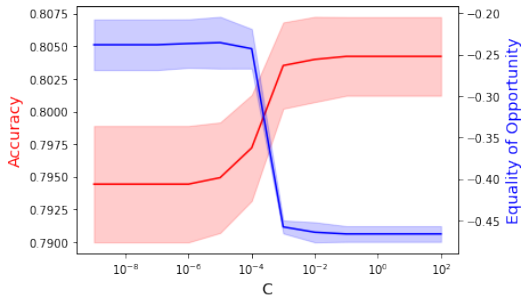
We can see the effect of adjusting this value on the standard model for the adult dataset in figure 1a. Smaller values specify stronger regularization and therefore result in more generalized models. As we can see from the figure there is a clear trade off between accuracy and fairness - as the accuracy increases it hurts the fairness. Increasing the value of C results in a less generalized model and this pushes the equality opportunity difference closer to 0 (bias in neither direction). Table 1 shows this effect as the most fair model has a low C value, and lower accuracy than the most accurate one. This is the result we would expect, an in depth study into a similar trade-off can be seen in figure 2 (Zafar et al. 2017) [5], as models are optimised on fairness the accuracy reduces. The german dataset however has a different pattern of results when running exactly the same algorithm. This can be seen in figure 1b and table 2. The fairness is generally closer to zero which means it is more fair. However it seems to slightly increase with the hyperparameter, when we would expect a trade off with accuracy (decrease). This could be down to the fact that the dataset is significantly smaller (700 datapoints compared to 34189) and has fewer features (11 compared to 18) than the Adult dataset and so it may that the results aren't statistically significant. The other point could be that the german dataset is not biased in the same way or we are looking at the wrong feature to judge bias. A case study into FICO scores (scores to predict credit worthiness in the USA) by Hardt et al.(2016) [1] demonstrate that for a non binary feature (race), the majority group (white) is classified more accurately than any other. For the purposes of similar comparison only sex was monitored for this dataset, and so no direct comparison can be made. However further analysis should be performed to dig deeper into why we see this effect were it in the scope of this report.

Adult Dataset		
	$AccurateModel(1)$	$FairModel(2)$
$Accuracy$	0.805	0.798
EOD	-0.436	-0.219
C	0.001	1e-05

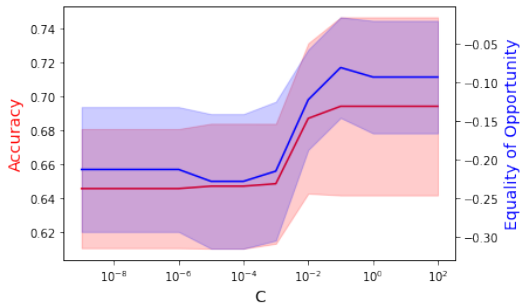
Table 1: Standard Model, Adult Dataset

German Dataset		
	$AccurateModel(1)$	$FairModel(2)$
$Accuracy$	0.697	0.66
EOD	0.157	0.15
C	0.01	1e-09

Table 2: Standard Model, German Dataset



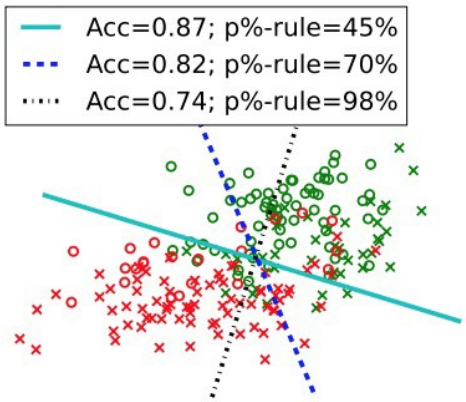
(a) Adult Dataset



(b) German Dataset

Figure 1: Standard model - accuracy and fairness (measured as equality of opportunity difference)

Figure 2: trade-off between accuracy and demographic parity on a linear classification problem (Zafar et al. AISTATS2017)[5]



3. Task 2 - Reweighed

The model is reweighed using the aif360 'Reweighed' module that draws its technique from F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, 2012. This is a type of preprocessing and we can see the results in figures 4a and 4b. This computes the weights for all combinations of classification (Y) and sensitive attribute (A) and applies this to the training dataset. The first thing to note is the significant improvement in fairness for all models across all hyperparameters, in the adult dataset the value gets very close to zero (see table 3) and we cannot see the same accuracy/fairness trade off as we did before the reweighing in the figure. It makes sense that the curve isn't the same shape as we have manipulated the training data by making those points with a higher weight used more often and vice versa with a lower weight. Therefore we can expect to see an increase in accuracy as the model is made less generalized but the changes in fairness are less significant. We can also note an overall loss in accuracy from the standard model in table 1 for both the fair and accurate models. This is relatively small though (0.014) and so we can infer that reweighing is a reasonable tradeoff. For the german dataset we see more of this trade off, but once again the numbers are significantly lower. This could be because the dataset is much smaller, we have fewer data points and so the reweighing takes on less significance. In fact the accuracy increases for both models slightly and fairness improves significantly.

An alternative or complementary prepossessing technique we could have used is a discriminitive clustering approach, i.e. cluster training data into K clusters, such that the probability of being assigned to the cluster k is independent of the sensitive feature A. This could be achieved through applying a Variational Fair Auto Encoder [3] or an adversarial approach (Madras et al., Learning adversarially fair and transferable representations, ICML 2018)[4]. The idea is that we want to learn a new representation Z such that it removes any information that the sensitive attribute is correlated with while preserving the information of X as much as possible. Figure 3 demonstrates this mapping graphically. The classification task can thus use a "cleaned" data representation and produce results that preserve fairness.

Adult Dataset		
	$AccurateModel(3)$	$FairModel(4)$
$Accuracy$	0.791	0.788
EOD	-0.035	-0.033
C	0.0001	1e-09

Table 3: Reweighed Model, Adult Dataset

4. Task 3 - Criterion Evaluation

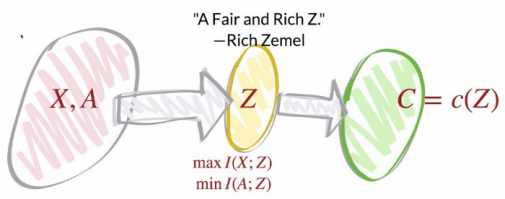
In order to balance the two metrics a metric based on F1 score was chosen:

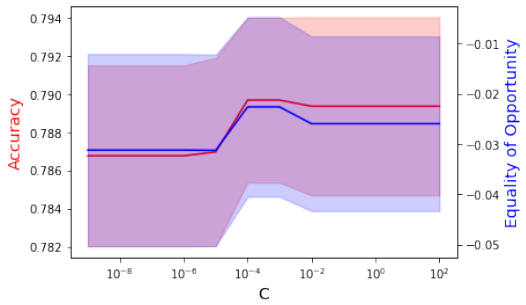
$$F = (1+\beta^2) * \frac{accuracy * (1 - |fairness|)}{(\beta^2 * accuracy) + (1 - |fairness|)}$$

German Dataset		
	$AccurateModel(3)$	$FairModel(4)$
$Accuracy$	0.707	0.703
EOD	0.023	0.017
C	0.01	1.0

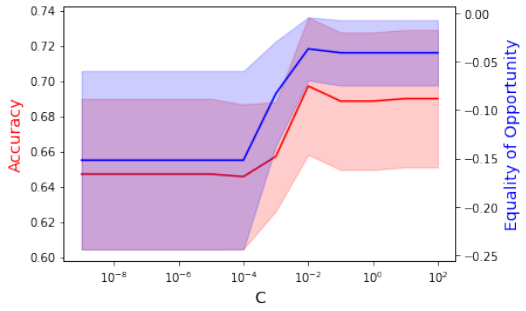
Table 4: Reweighed Model, German Dataset

Figure 3: Preprocessing illustration Zemel et al. ICML2013 [6]





(a) Adult Dataset

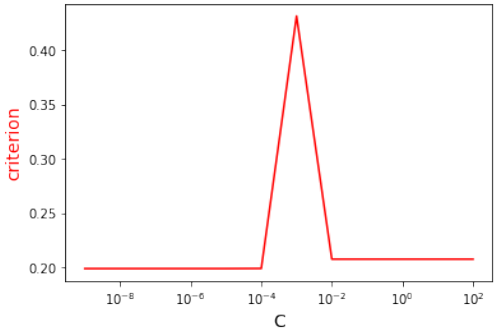


(b) German Dataset

Figure 4: Reweighed model - accuracy and fairness (measured as equality of opportunity difference)

This has the effect, of being able to choose a β where fairness is considered β times as important as accuracy. We are able to tune this value but for the purposes of this we can give equal importance to both by setting β to 1. We do 1-—fairness— in order to account for the fact that close to 0 is good and that we want a high score to be good. The behaviour we hope to see is displayed in figure 5 for the german dataset and standard model as a value peaks with a good tradeoff. However adjusting beta means that this is different for all datasets/models and is too engineered, and so this behaviour is not seen throughout. However it does provide a rational way of getting the best of fairness and accuracy.

Figure 5: Custom F score plotted against C value, German unweighed dataset)[5]



In table 5 we see that when comparing the 'best' model according to the F score against previously selected models for the adult data set, for the standard models we select the same hyperparameter as the 'fair' model. This would denote that the increase in fairness outweighs the increase in accuracy. For the reweighed models, we select the same model as the 'accurate' model when selecting by criterion, implying that the difference in accuracy is more important once we reweigh. This is as we would expect according to the previous section in which we decided that reweighing does not see a significant change between the accurate and fair models in table 4a. In the German dataset the opposite models are selected which lines up with the reverse behaviour we see throughout the German analysis curves.

We can also infer that the process of applying weights to the training data gives a model with a better balance of accuracy and fairness as well as being more fair generally. This aligns with the conclusions drawn from Kamiran et al. (2012)[2] in which the fairness is significantly improved at little cost to the accuracy after reweighing. We should therefore always be reweighing models for classification.

5. Task 4 - Remove Feature

In order to take the analysis further it was decided to remove the sensitive feature (sex) from the datasets. The results as seen in table 6 demonstrate a significant benefit from doing this with regards to the fairness. At the same time we see an overall reduction in accuracy (from models (5) and (6)) after removing the feature which implies that the model was relying on using this feature to classify. For both datasets, both unweighed and reweighted we see an

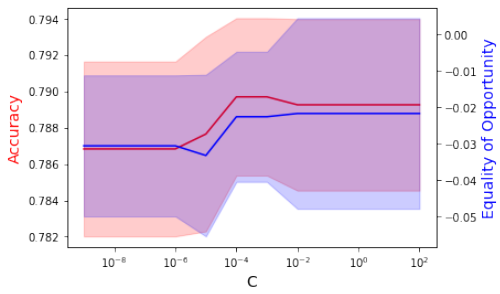
Adult Dataset		
	Standard(5)	*Reweighed*(6)
Accuracy	0.798	0.791
EOD	-0.219	-0.035
C	1e-05	0.0001
F	0.789	0.869
German Dataset		
	Standard(5)	*Reweighed*(6)
Accuracy	0.697	0.703
EOD	-0.157	-0.017
C	0.01	1.0
F	0.763	0.82

Table 5: Model selected on custom criterion, German Dataset

overall improvement in fairness when compared with models (5) and (6), this demonstrates that as the model is unable to see the protected attribute, it therefore cannot be biased against it when evaluating the equal opportunity difference, and so this is to be expected. In a similar way to reweighing this can be seen to reduce the fairness to a level at which the changes with hyperparameter are insignificant as it is so low. The values for fairness are very similar to that achieved after reweighing. The figures for the German dataset can be viewed in appendix A, displaying similar behaviour.

This is somewhat unexpected as research demonstrates that simply removing the protected attribute is insufficient. As long as the model takes in features that are correlated with gender even if gender is explicitly excluded, it wont do much good. This is known as the concept of redundant encoding of gender or the naive approach of 'fairness through unawareness' [1]. In fact taking this approach can prevent machine learning engineers from accessing the protected attribute they need to check whether or not they are biased. It should therefore be avoided. With more extensive modelling on a larger dataset with a larger set of features we would expect a similar result.

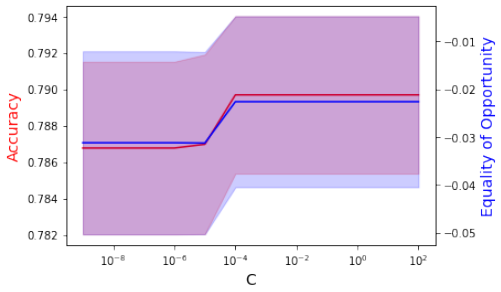
Figure 6: Standard Model, Adult Dataset, Feature Removed, Accuracy and Fairness



Adult Dataset		
	Standard(7)	*Reweighed*(8)
Accuracy	0.791	0.791
EOD	-0.035	-0.035
C	0.0001	0.0001
German Dataset		
	Standard(7)	*Reweighed*(8)
Accuracy	0.663	0.713
EOD	-0.133	-0.14
C	1e-09	0.001

Table 6: Model selected on custom criterion with feature removed, German Dataset

Figure 7: Reweighed Model, Adult Dataset, Feature Removed, Accuracy and Fairness



6. Conclusion

The main conclusion to draw is that there is an innate trade-off between accuracy and fairness when performing regression tasks with respect to sensitive features. The reweighing process did a lot to reduce this balance and according to the custom criterion we can achieve a better overall model by preprocessing the data in this way. The main limitation to all this analysis is that there is no clear possible way to make a model completely fair. In the much publicised case of ProPublica and Northpointe's debate over COMPAS's software, Propublica emphasised the fact that it consistently overestimated the number of black defendants who didn't re-offend and underestimated the risk of White defendants who did. Northpointe defended by saying that the model was equally accurate in its predictions of defendants reoffending whether Black or White. Both criteria cannot be satisfied at the same time unless the base rates are the same. This issue will always remain, but this report hopes to have drawn attention to some of the inherent biases in the datasets and the difficulties faced in unravelling them, in the hope that all models will be interrogated in a similar vein.

References

[1] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning, 2016. 1, 4

[2] F. Kamiran, A. Karim, S. Verwer, and H. Goudriaan. Classifying socially sensitive data without discrimination: An analysis of a crime suspect dataset. pages 370–377, 12 2012. 3

[3] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder, 2017. 2

[4] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations, 2018. 2

[5] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 20–22 Apr 2017. 1, 2, 3

[6] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. 3

A. Appendix

Figure 8: Standard Model, German Dataset, Feature Removed, Accuracy and Fairness

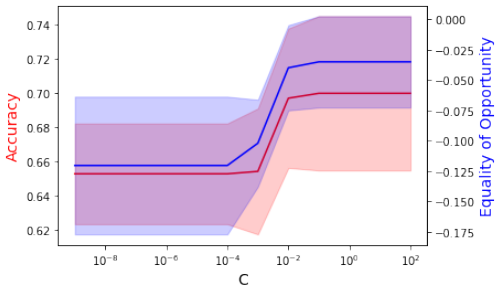


Figure 9: Reweighed Model, German Dataset, Feature Removed, Accuracy and Fairness