Student Number: 244106

1. Introduction

With the advent of ever more powerful tools in artificial intelligence to automate real-world decision-making, the social and political urgency of ensuring that such methods retain not only a high level of accuracy but also produce fair and unbiased solutions is becoming increasingly apparent [1].

In general, the training process in machine learning aims to produce accurate models which can generalise well to new, unseen data. This can be controlled by a regularisation term which imposes some form of constraint on the model to prevent from overfitting to the data. When we underfit we get a biased solution, and when we overfit this means that our model is too complex and small changes to the data means that the solution changes greatly thereby leading to a large amount of variance and a poor ability to generalise. This is known as the bias-variance trade-off [2]. It therefore seems intuitive that fitting well to the data may accentuate any inherent bias present within the data and gives rise to the question of whether fairness be controlled to some extent by regularisation strength.

While algorithmic fairness methods have been developed to attempt to mitigate the inherent biases present within many datasets, improved fairness has classically been seen to be at the expense of accuracy [3][4]. However, the underlying reasons for the existence of such a compromise is by no means trivial, nor is it unavoidable. Indeed, recent work in [5] suggests that this trade-off arises as a result of mappings between outcome and unprivileged groups being noisier, reducing the separability of the data. The author's theoretical analysis shows that for a biased data distribution, there always exists an 'ideal distribution' for which a trade-off is not apparent when accuracy is measured with respect to such an adjusted distribution.

Given the uncertainty surrounding the subject, we therefore aim to investigate this apparent trade-off by comparing the effect of regularisation strength on both accuracy and fairness. We first investigate this in the context of standard machine learning models, before going on to analyse how this model selection strategy might change with fairness-based algorithms. Finally, we propose a selection criterion which allows us to obtain models with both high fairness and high accuracy.

2. Methods

We use two datasets from the *AIF360* library in order to perform our analysis. The first is the 'Adult' dataset (with 48842 data-points), which includes census data from 1994 and the prediction task is to classify whether a person earns over $50k p.a. or not. The second is the 'German' dataset (with 1000 data-points) and the task is to predict whether people have a good or a bad credit score based on a set of attributes. Both datasets include sensitive attributes: 'sex' in the Adult dataset and 'age' in the German dataset - both of which fall under the category of legally protected characteristics in many countries [6]. There is an inherent bias towards the more privileged group in both datasets for male and aged individuals respectively (Fig 1. Appendix B).

Given that both datasets involve a binary prediction class, we use a logistic regression model in which we varied the hyperparameter λ in order to explore the bias-variance trade-off discussed previously. Equation 1 describes the loss function that we are trying to minimise in order to discover the optimal weights for our model, with the added L2 regularisation bounding these weights while eq.(2) describes the gradient of the loss function with respect to the weights.

$$minimize\, w \sum_{n=1}^{N} - log(P(y_n | x_n, w)) + \frac{\lambda}{2}||W||^2 \quad (1)$$

$$\nabla_W E_{logistic\ regression} = \sum_{n=1}^{N} \frac{-y_n x_n}{1+exp(y_n W^T x_n)} + \lambda W \quad (2)$$

We measure accuracy as the proportion of correctly predicted labels while fairness is measured by the 'equality of opportunity' metric (EOP). EOP means that the classifier must make its positive predictions independent of the sensitive attribute and therefore we expect there to be no difference in true positive rates between the privileged and unprivileged groups. EOP is therefore most fair when closest to zero, while negative values indicate a bias towards the privileged group [7].

3.1. Can better generalisation correspond to fairer models?

The first task involved establishing whether better generalisation, as controlled by varying the regularisation strength, could correspond to fairer models. We use accuracy as an indication of the generalisability of each model to the data, and a process of 5-fold cross-validation in order to determine the best value of C (inverse of λ) to use.

As Fig.1 shows, we found that for our standard model, decreasing the regularisation strength (increasing the value of C) had the effect of increasing the model accuracy for both datasets. For both, this increase occurred between the

C values of 1E-05 until 1E-02 (with 2.95E-02 and above yielding the best accuracy in cross-validation for both). When selecting this value for training, the accuracy achieved was 80.42% and 70.34% for the Adult and German data respectively (see table 1 in Appendix A).

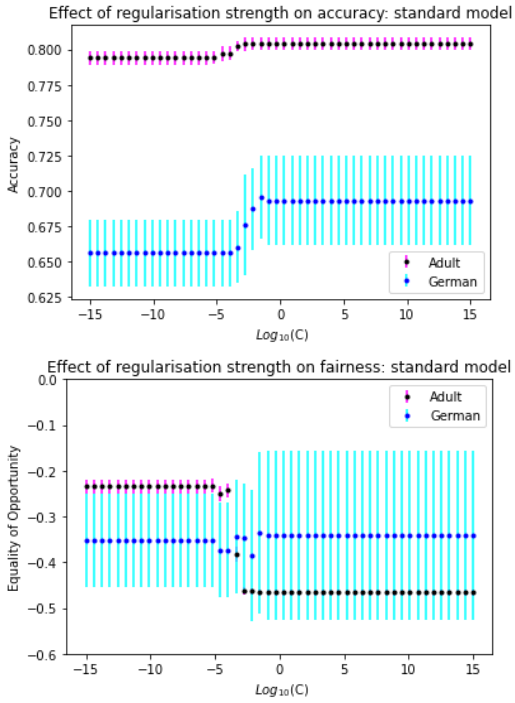

Figure 1. Standard model. Average of 5-fold cross-validation for a) accuracy and b) EOP for inverse regularisation strength (C) for the two datasets. Error bars indicate the standard deviation from the mean.

However, these low regularisation values corresponded to much more discriminative decision making. With the Adult dataset, the effect of regularisation strength on fairness was the inverse effect seen in accuracy, with EOP dropping to a more negative value the same bracket of 1E-05 to 1E-01 indicating increased unfairness. For the German data this opposing pattern was less apparent, with fairness instead being consistently poor across all regularisation strengths (although a disturbance still occurs in the same bracket). The standard deviations for this data are also much larger during the cross-validation process which is unsurprising given the fact that it is a much smaller dataset and we would therefore expect a greater variance in accordance with the law of large numbers in probability theory.

Fairness is poor overall for both datasets, with distinct bias towards privileged groups in both cases. This is in line with the literature which points towards there being a distinct trade-off between accuracy and fairness in classical machine learning approaches [3][4]. Figure 3a highlights the nature of this trade-off clearly as accuracy and fairness are negatively correlated. The decrease in regularisation strength can be seen to allow for a more complex model thereby improving generalisation, but essentially fitting to the bias present within the datasets.

It is worth noting that regularisation strength only influences the performance of the model within certain limits. As identified above, values of C between 1E-05 and 1E-01 is the key range and either side of this bracket an effect is not apparent. This is likely since for very high regularisation strengths, there is a limit to how simple the model can be (the extent to which you can penalise the weights). At the very least in this case, the lower limit of C appears to be in a local minimum. The same is true for the lower limits of regularisation in which the logistic regression cannot fit a more complex model to the data, as regularisation strength tends to zero.

3.2. How does this tradeoff change after the application of fairness-based reweighing of the dataset?

There have been several algorithmic fairness methods developed which attempt to mitigate the bias in the labelled dataset in order to reduce or remove any discrimination before the classifier is trained (these are known as pre-processing techniques). For the following experiment, we employ a technique known as reweighing in which the training dataset is made “discrimination free” by assigning each combination of the outcome (Y) and sensitive feature (A) - which for binary classification datasets with binary sensitive features means there are 4 subgroups - a weight related to the inverse of its relative frequency of occurrence in the dataset. The data with a higher weight is therefore sampled more frequently during the training process, thereby making the outcome independent of the sensitive attribute, A [6].

In our fairness-based (reweighed) model, we again see this jump occur at a C of around 1E-02 for both accuracy and fairness. This marks an increase in the overall accuracy of the models (with both datasets) as seen in section 3.1. However, unlike the first experiment, the fairness has been vastly improved for both datasets, with the Adult dataset retaining a near-zero (<0.05) fairness value at all regularisation strengths essentially removing the trade-off (see Fig. 2). With the German dataset, the C value of 1E-02 leads to an increase rather than decrease in fairness, therefore aligning the direction of desired regularisation strength for fairness (and this EOP is significantly reduced to values <0.011) with that of accuracy which also means there is no concept of a trade-off. This can be seen to agree with the discovery of an ideal distribution as discussed in [5]. This negation of the trade-off seen in the Adult dataset can clearly be seen in Fig.3 in which the negatively correlated metrics in Fig.3a

become aligned in Fig.3b (where there is no correlation apparent).

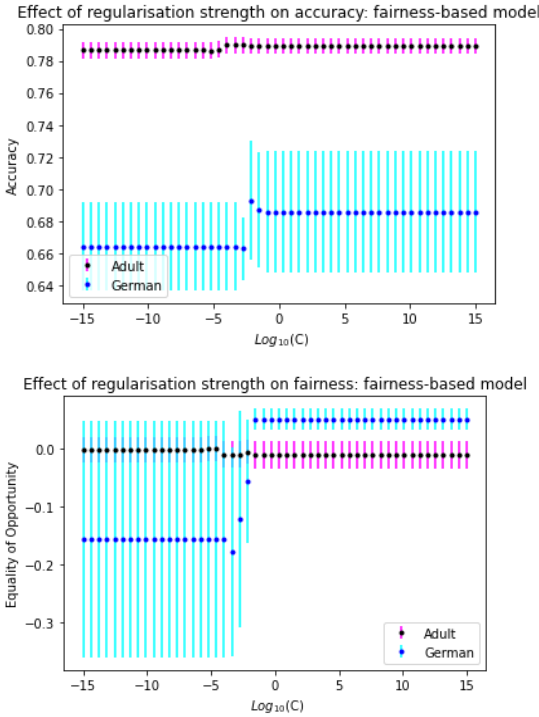

Figure 2. Fairness-based models. Average of 5-fold cross-validation for a) accuracy and b) EOP for inverse regularisation strength (C) for the two datasets. Error bars indicate the standard deviation from the mean.

Table 2 (see appendix) presents the results of the cross-validation in finding the regularisation strength which produced the most accurate and fair models after the reweighing process has been applied. It is worth noting that there is a small compromise made on the overall magnitude of the accuracy that can be achieved after the reweighing process, however this is a very small difference given the size of the dataset (1.36% and 2.34% for the Adult and German data respectively). As with any human-centric machine learning algorithm, the context of the application should be considered when deciding the significance of such compromises.

These results provide evidence that reweighing as an algorithmic pre-processing method for improving fairness can align the accuracy and fairness metrics such that better generalisation can indeed lead to fairer models.

3.3. Can we define and test a model selection strategy that accounts for both accuracy and fairness?

We ideally want to be able to define a metric, Z, which considers both fairness and accuracy, and finds a value which can balance the two. In order to do this, we wanted to define both accuracy and EOP as terms that could both be minimised in order to improve performance.

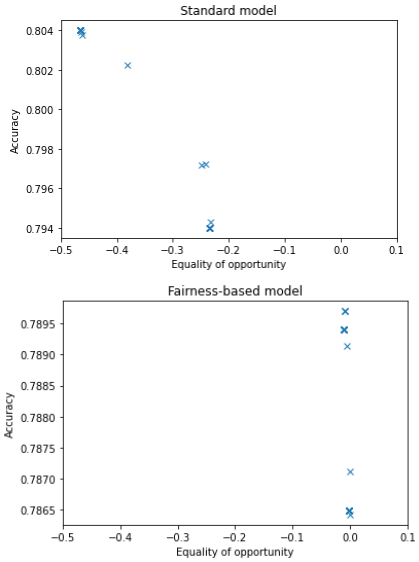

Figure 3. Accuracy against EOP for the Adult dataset for different values of C. a) The standard model shows a trade-off exists between accuracy and fairness while this disappears in the b) fairness-based model.

This required us to take 1 - accuracy, and likewise we want the absolute values of EOP in order to make sure the minimal value would always be the best (as this is closest to zero). We also wanted both accuracy and fairness to be of equal weighting when selecting our models, therefore we normalised both values to take a range between 0 and 1. The final metric is therefore defined as:

$$min\left(\left|E_{normalised}\right| + Q_{normalised}\right) \quad (3)$$

Where E is the EOP value and Q=1 - accuracy. Both terms have been normalised (scaled to take values between 0 and 1) by the following method:

$$X_{normalised} = \frac{\left(X - X_{min}\right)}{\left(X_{max} - X_{min}\right)} \quad (4)$$

Using this metric, we can show that we can achieve the optimal balance between accuracy and fairness in both of our datasets, and for both the standard and fairness-based models (Fig.4). The lowest Z value is therefore the most accurate and fair and can take values between 0 and 2 (with 2 being both the least accurate and the least fair of all values, and 0 being that it managed to achieve a

regularisation strength which yielded both the most accurate and the most fair out of all values). Close to zero therefore indicates a strong aligning of the two metrics (accuracy and fairness) for a given regularisation strength. Table 3 (in appendix A) shows that the value of C which yields the lowest Z value achieves a high accuracy and low fairness for both the standard and fairness-based models.

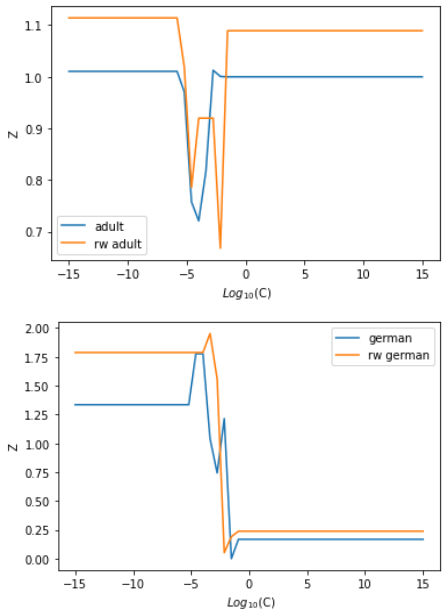

Figure 4. Z-metric as a function of C for a) Adult and b) German dataset

It is important to note that as a result of this process of normalising the data, this accuracy-fairness metric is not comparable between datasets and between standard and reweighed data. It only acts to give the optimal balance for that dataset instance and is therefore designed to be very general. For instance, with the unweighted adult dataset, the actual fairness values vary between -0.22 and -0.48, whereas in the reweighed instance, the fairness only varies between 0 and -0.01. However, given the normalisation for the values to take a range between 0 and 1, our fairness metric does not reflect how at lower regularisation strengths we can favour higher accuracies at the expense of fairness with the reweighed data given that a change of 0.005 can be considered negligible. This trade-off should not be made, however, when the dataset has not been reweighed, given that we see changes of over 0.2 in the fairness for lower regularisation of the weights.

In the case where we know that we will perform reweighing (or perhaps some other kind of algorithmic fairness method) which will provide us with two different models, we can adjust this metric in order to make the two comparable. This involves normalising the data based on the global minimum and maximum values of the accuracy and fairness metrics out of the two models. This is a much less general metric but works well when we know that we are comparing two models for which one has been reweighed (or some other form of algorithmic fairness method has been implemented) and we wish to compare between the two. Figure 5 this example of global normalisation, making the Z values for standard and fairness-based models comparable). This graph highlights how the reweighing method improves the overall Z score that can be achieved between the two models given that the magnitude of EOP is so drastically improved in the cases of both datasets.

Depending on the context of the application, we could conceive of a scenario in which we wish to favour one metric (out of accuracy and fairness) over another. In this case, an interesting avenue for future work in this area would be to add coefficient terms into the Z-metric equation which allows you to favour accuracy or favour fairness with a different weighting.

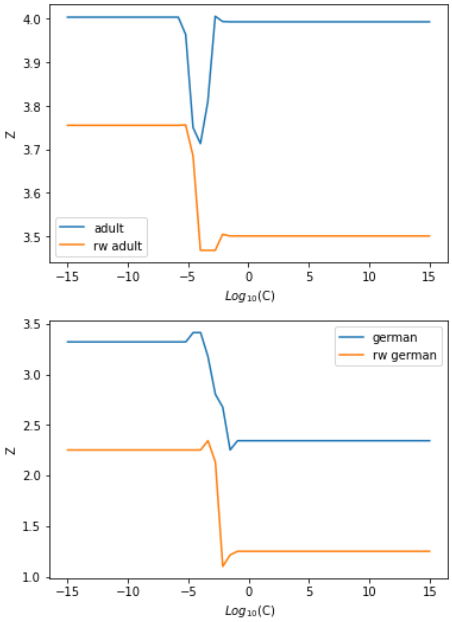

Figure 5. Globally normalised Z-metric (according to the minimum and maximum values seen between the standard and fairness-based models) against C for a) Adult and b) German datasets.

References

[1] D. Pessach and E. Shmueli, "A Review on Fairness in Machine Learning", *ACM Computing Surveys*, vol. 55, no. 3, pp. 1-44, 2022.

[2] Goodfellow, Y. Bengio and A. Courville, *Deep learning*. Cambridge, Massachusetts: The MIT Press, 2016, pp. 96-161.

[3] Chen, I. Y., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In Advances in Neural Information Processing Systems, pp. 3539–3550, 2018.

[4] Zhao, H. and Gordon, G. J. Inherent tradeoffs in learning fair representation. arXiv preprint arXiv:1906.08386, 2019.

[5] S. Dutta, S. Wei, P. Chen, S. Lui and K.R. Varshney, *"Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing*", PMLR, 119, 2020

[6] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination", *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1-33, 2011.

[7] M. Hardt, E. Price and N. Srebro, "*Equality of Opportunity in Supervised Learning*", arXiv:161206830, 2016.

Appendix A

Model No.	Design	Dataset	C	Accuracy (%)	EOP
1a	Most accurate	Adult	2.95E-02	80.42	*-0.463*
1b	Most accurate	German	2.95E-02	70.34	*-0.463*
2a	Most fair	Adult	6.25E-06	*79.66*	-0.242
2b	Most fair	German	2.95E-02	*69.0*	-0.463

Table 1. Standard model. C-values selected via a process of 5-fold cross-validation and then hyperparameter was used to train and test final model.

Model No.	Design	Dataset	C	Accuracy (%)	EOP
3a	Most accurate	Adult	1.05E-04	79.06	*-0.155*
3b	Most accurate	German	7.20E-03	68.0	*0.082*
4a	Most fair	Adult	2.56E-05	*79.84* *63.67*	0.036
4b	Most fair	German	2.95E-02	*69.67* *68.0*	0.0824

Table 2. Fairness-based model. C-values selected via a process of 5-fold cross-validation and then hyperparameter was used to train and test final model.

Model No.	Model	Dataset	C	Z	Ac (%)	EOP
5a	Standard	Adult	1.05E-04	0.721	79.84	-0.247
5b	Standard	German	2.95E-02	0.0	69.67	-0.437
6a	Fairness-based	Adult	7.20E-03	0.668	80.42	0.035
6b	Fairness-based	German	7.20E-03	0.052	69.67	0.066

Table 3. Accurate-and-fair model selection using Z-metric. C-values selected via a process of 5-fold cross-validation and then hyperparameter was used to train and test final model.

Appendix B

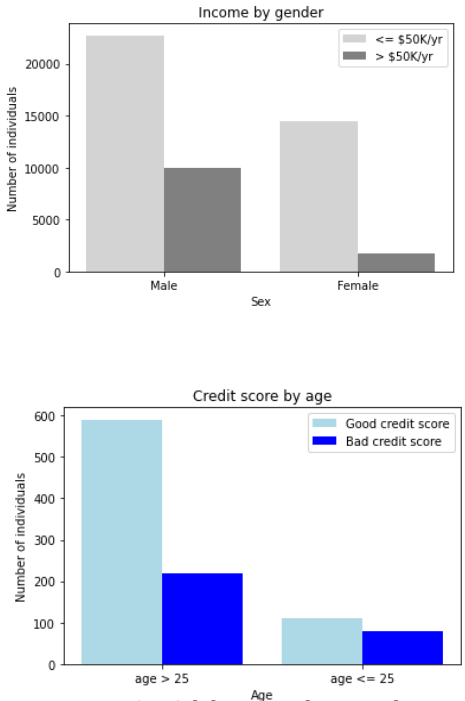

Figure 1. a) Adult dataset showing that a higher proportion of males earn over 50k per year which violates equality of opportunity between privileged and unprivileged groups. B) German dataset showing that a greater proportion of young people (unprivileged group) are given bad credit scores.