01932146

1. Introduction

In this report, algorithmic fairness is explored using two data sets from the *aif360* Python library, the AdultDataset and CompasDataset. In case of the AdultDataset, we work with sex as the protected feature, and we train a model to predict whether an individual has an income over $50k. With the CompasDataset, where the protected features are sex and race, we aim to predict two-year recidivism in previously sentenced individuals.

2. The AdultDataset

2.1. Task 1: Regularisation

We have performed logistic regression on the data and evaluated the accuracy and several fairness metrics using 10-fold cross-validation (CV). In the plot below, we can see that the model is not very sensitive to changes in the regularisation parameter (C). The same values were obtained for any $C \leq 10^{-7}$ tried. As we can see in the plot, some of the metrics do change for C in the range between 10^{-7} and 1, but for $C \geq 1$ they stop changing again.
The trade-off between accuracy and fairness is clearly visible in *Figure 2*. However, the accuracy only changed marginally while varying C, ranging between 79.6% and 80.5%. All the fairness metrics improve noticeably, though, when selecting $C < 10^{-3}$. We can see the largest improvement in the disparate impact metric, which jumps from 0 to 0.3.

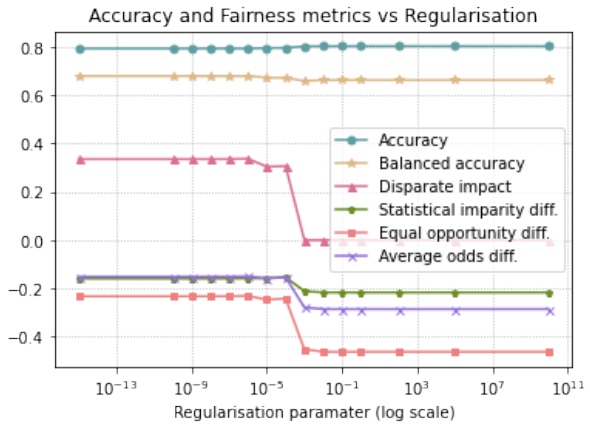

Figure 1. Averaged CV results on the validation set of the AdultDataset for varying levels of regularisation.

Thus, when selecting a model based on accuracy, the best option is to select $C = 10^{-2}$, since for larger C's the accuracy remains the same, but the other metrics perform slightly better. When selecting based on the fairness metrics, the best result is achieved with $C = 10^{-6}$, where the best scores for 4 out of the 5 fairness metrics are obtained. Thus, we can see that even an approach as simple as tuning the C parameter can help us create a more fair model with a loss in accuracy below one percentage point.

C	Accuracy	Balanced accuracy	Disparity impact	Statistical disparity difference	Equal opportunity difference	Average odds difference
≤e-07*	0.7960	0.6791	0.3346	-0.1562	-0.2208	-0.1448
e-06	**0.7963**	**0.6768**	**0.3425**	**-0.1507**	**-0.2125**	**-0.1385**
e-05	0.7985	0.6733	0.3053	-0.1537	-0.2379	-0.1517
e-04	0.7989	0.6733	0.2956	-0.1558	-0.2425	-0.1549
e-03	0.8047	0.6595	0.0000	-0.2082	-0.4506	-0.2768
e-02	**0.8049**	**0.6619**	**0.0000**	**-0.2128**	**-0.4587**	**-0.2824**
e-01	0.8049	0.6626	0.0000	-0.2143	-0.4610	-0.2841
≥e+00*	0.8049	0.6629	0.0000	-0.2148	-0.4620	-0.2848

Figure 2. Averaged CV results on the validation set. Each column is colour-coded (green = better, red = worse). The lines with the two selected models are highlighted.
* $C = \{10^{-15}, 10^{-10}, 10^{-9}, 10^{-8}\}$ have been tried and all yielded the result described in line 1. $C = \{10^{2}, 10^{5}, 10^{10}, 10^{15}\}$ have all yielded the result described in the last line.

As we can see below, on the held-out test set, model (2) outperforms model (1) in the fairness metrics, but their accuracy only differs marginally. Note that here the evaluation is done using the same train/test split as above. Varying the random seed can thus only affect the model initialization and does not introduce any variance in the results in this case.

	C	Accuracy	Balanced accuracy	Disparity impact	Statistical disparity difference	Equal opportunity difference	Average odds difference
(1)	e-02	0.8143	0.6627	0.0000	-0.1996	-0.4462	-0.2718
(2)	e-06	0.8108	0.6882	0.2842	-0.1619	-0.2321	-0.1516

Figure 3. Performance on the test set for models selected (1) based on accuracy & (2) based on the fairness metrics on the val. set.

We have used Scikit-learn's logistic regression function, where smaller values of C denote stronger regularization. Thus, we can see that models with stronger regularization (up to an extent) did perform better on the fairness metrics.

2.2. Task 2: Fairness metrics

In *Task 1*, we have seen that simply performing hyperparameter search while considering fairness metrics can lead to some improvement in model fairness. However, there still is room for improvement. Thus, in this section, we will look at a fairness method of reweighing.

Similarly to the approach in *Task 1*, we have performed logistic regression on the data and evaluated the accuracy and several fairness metrics using 10-fold CV. However, this time we also performed reweighing of our samples and used these weights when fitting our models. In the plot below we can see that after reweighing, the regularisation parameter has almost no effect on model performance. The same values were obtained for any $C \leq 10^{-6}$ tried, and similarly for any $C \geq 10^{-1}$ tried. Even in the range between 10^{-6} and 10^{-1}, all the metrics' range is about one percentage point. The values obtained do differ significantly from the *Task 1* results, though.

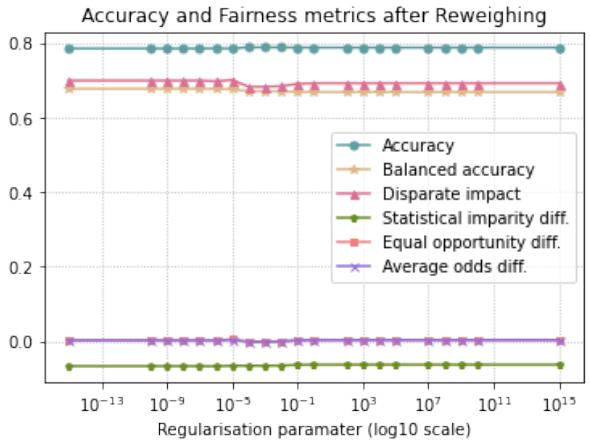

Figure 4. Averaged CV results on the validation set of the Adult-Dataset for varying levels of regularisation.

While almost all the fairness metrics improve, the most noticeable difference is in the disparate impact, where the best performance jumped from 0.34 to 0.71 compared to the *Task 1* results. Further, the (best) equal opportunity difference reduces from -0.21 to practically zero, i.e. the best possible outcome. In terms of accuracy, the best performance reduces by about 1 percentage point.

C	Accuracy	Balanced accuracy	Disparity impact	Statistical disparity difference	Equal opportunity difference	Average odds difference
≤e-06*	0.7874	0.6778	0.7049	-0.0645	0.0077	0.0063
e-05	**0.7875**	**0.6769**	**0.7071**	**-0.0636**	**0.0087**	**0.0071**
e-04	0.7903	0.6714	0.6873	-0.0637	0.0001	0.0022
e-03	**0.7903**	**0.6714**	**0.6873**	**-0.0637**	**0.0001**	**0.0022**
e-02	0.7896	0.6684	0.6981	-0.0602	0.0064	0.0062
≥e-01*	0.7895	0.6669	0.6848	-0.0624	-0.0032	0.0005

Figure 5. Averaged CV results on the validation set of the Adult-Dataset for varying levels of regularisation after reweighing. Each column is colour coded (green = better, red = worse).
* $C = \{10^{-15}, 10^{-10}, 10^{-9}, 10^{-8}, 10^{-7}\}$ have been tried and all yielded the result described in line 1. $C = \{1, 10^2, 10^3, 10^5, 10^7, 10^{10}, 10^{15}\}$ have all yielded the result described in the last line.

This time, when selecting a model based on accuracy, the best option is to select either $C = 10^{-3}$ or 10^{-4}. We have selected $C = 10^{-3}$, just because as the value of C increases, the reduction in accuracy is marginally smaller. Selecting a model based on the fairness metrics is not as simple as in *Task 1*. We have decided to consider the 'difference from the ideal state', e.g. take $1 - accuracy$ or the absolute value of the equal opportunity difference. We have summed up the values obtained for all the fairness metrics and selected the model with the lowest sum, which is the model with $C = 10^{-5}$. Note that applying this method in *Task 1* would not change the decision made.

We can see the performance on the held out test set for the two selected models below. Note that on the test set, the performance differs only marginally across all the metrics.

	C	Accuracy	Balanced accuracy	Disparity impact	Statistical disparity difference	Equal opportunity difference	Average odds difference
(3)	e-03	0.8055	0.6774	0.6785	-0.0606	0.0587	0.0311
(4)	e-05	0.8035	0.6847	0.6781	-0.0654	0.0610	0.0309

Figure 6. Model performance on the held-out test set for the models selected (3) based on accuracy and (4) based on the fairness metrics on the validation set.

2.3. Task 3: Final selection

	Task 1: Regularisation				Task 2: Reweighing		
C	Fairness Score	Accuracy score	Final score	C	Fairness Score	Accuracy score	Final score
≤e-07*	0.3016	0.2040	0.2528	≤e-06*	0.1392	0.2126	0.1759
e-06	**0.2965**	**0.2037**	**0.2501**	e-05	0.1391	0.2125	0.1758
e-05	0.3130	0.2015	0.2572	e-04	0.1414	0.2097	0.1756
e-04	0.3168	0.2011	0.2590	**e-03**	**0.1414**	**0.2097**	**0.1756**
e-03	0.4552	0.1953	0.3252	e-02	0.1413	0.2104	0.1758
e-02	0.4584	0.1951	0.3267	≥e-01*	0.1429	0.2105	0.1767
e-01	0.4594	0.1951	0.3273				
≥e+00*	0.4597	0.1951	0.3274				

Figure 7. Averaged CV results for both *Task 1 & 2* on the validation set using the selected criterion.
* Values of C used are described in sections 2.1 and 2.2.

To select the final model, we adopt an approach similar to selecting a model based on fairness metrics in *Task 2*. We look at the 'difference from the ideal state' score for both the accuracy and the fairness metrics. To account for the fact that we compare several fairness metrics to a single value representing accuracy, we take the average of the scores obtained for each fairness metric. We take the average of the scores obtained for accuracy and the fairness metrics and select the model with the lowest resulting value. For simplicity, we take the arithmetic mean, but we could also opt for weighted average (either when calculating the mean fairness score or the overall score) to account for the

importance of each component for a specific use case. For more detail on how the score is obtained see the appendix.
Note that the best-performing models based on this criterion correspond to models (2) and (3) selected in the previous two sections.
We can also see from the table above that reweighing generally results in much better final scores for the AdultDataset, as we achieve a significant improvement on the fairness metrics at a small cost in terms of accuracy.

			Accuracy	Balanced accuracy	Disparate impact	Statistical parity difference	Equal opportunity difference	Average odds difference
(5)	e-06	avg	0.7976	0.6754	0.3331	-0.1482	-0.2109	-0.1354
		std	0.0084	0.0092	0.0446	0.0152	0.0464	0.0242
(6)	e-03	avg	0.7922	0.6704	0.6838	-0.0623	0.0258	0.0147
		std	0.0087	0.0083	0.0338	0.0074	0.0219	0.0106

Figure 8. Averaged results (mean and standard deviation) for 5 different random seeds on the held-out test set.

3. The CensusDataset

3.1. Task 1: Regularisation

3.2. Task 2: Fairness metrics

3.3. Task 3: Final selection

4. Conclusion

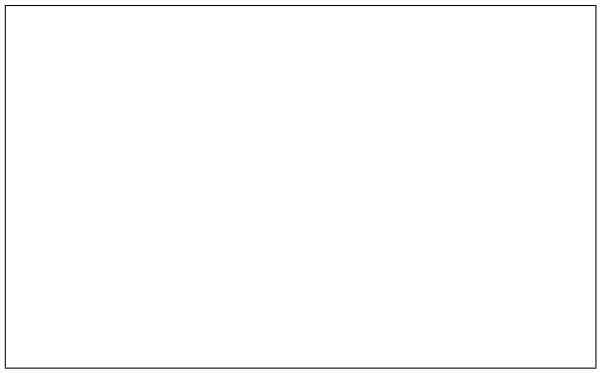

Figure 9. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

Formatting requirements etc.

Length: no longer than 4 pages in length including tables and figures. List of references can be in the fifth page. All text must be in a two-column format. The total allowable width of the text area is 17.5 cm wide by 22.54 cm high. Columns are to be 8.25 cm wide, with a 0.8 cm space between them. The student number (on the first page) should begin 2.54 cm from the top edge of the page. The bottom margin should be 2.86 cm from the bottom edge of the page for 8.5×11-inch paper (US letter); for A4 paper, approximately 4.13 cm from the bottom edge of the page.

Please number all of your sections.

Wherever Times is specified, Times Roman may also be used. Main text should be in 10-point Times, single-spaced. Section headings should be in 10 or 12 point Times. All paragraphs should be indented 1 pica (approx. 0.422 cm). Figure and table captions should be 9-point Roman type as in Figure 9.

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your report. When referenced in the text, enclose the citation number in square brackets, for example [1]. Where appropriate, include the name(s) of editors of referenced books.

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the report. We will choose to print your report in order to read it. You cannot insist that we do otherwise, and therefore must not assume that we can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

References

[1] Authors. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as additional material `fg324.pdf`. 3