01932146

1. Introduction

In this report, algorithmic fairness is explored using two data sets from the *aif360* Python library, the AdultDataset [3] and the CompasDataset [2]. In case of the AdultDataset, we work with sex and/or race 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 again sex and race, we aim to predict two-year recidivism in previously sentenced individuals. For both datasets, we have performed logistic regression and evaluated several fairness and accuracy metrics using 10-fold cross-validation (CV).
Note that throughout the report we focus mainly on accuracy and equal opportunity difference, but other metrics are provided as well for reference.

2. The AdultDataset

2.1. Task 1: Regularisation

The plot below shows performances when taking sex as our sensitive feature. 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. Some of the metrics do change for C in the range between 10^{-7} and 1, but for $C \geq 1$ they stay constant again.
The trade-off between accuracy and fairness is clearly visible in 2, although the accuracy only changed marginally while varying C, ranging between 79.6% and 80.5%. All the fairness metrics improve noticeably, however, when selecting $C < 10^{-3}$. We can see the largest improvement in the disparate impact metric, which jumps from 0 to 0.3.

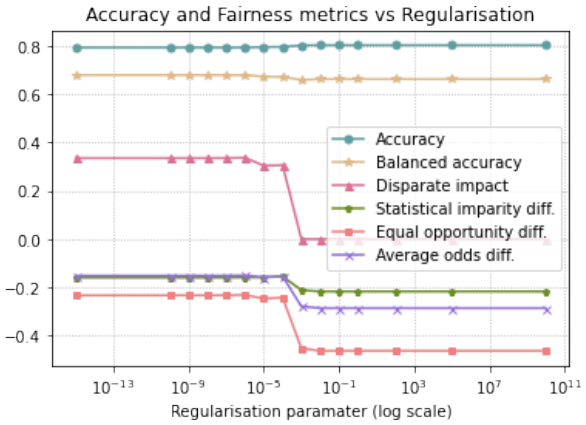

Figure 1. Averaged CV results on the validation set of the AdultDataset for varying levels of regularisation.

Thus, when selecting a model based on accuracy, we are choosing between $C = \{10^{-2}, 10^{-1}, 1\}$. The best option seems to be $C = 1$, since it has the best balanced accuracy score out of the three. When selecting based on the fairness metrics, the best result is achieved with $C = 10^{-6}$, regardless of which fairness metric we choose to look at.
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. However, there is still room for improvement, with the disparate impact metric still being far below the "4/5 rule" and the equal opportunity metric showing bias towards the privileged group.

C	Accuracy	Balanced accuracy	Disparate impact	Statistical parity 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	Disparate impact	Statistical parity difference	Equal opportunity difference	Average odds difference
(1)	e+00	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 fairness on the validation 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 *2.1*, we have seen that simply performing hyperparameter search while considering fairness metrics can lead to some improvement in model fairness. In this section, we

will explore whether this can be further improved using a fairness (pre-processing) method of reweighing.
Similarly to *2.1*, we have performed logistic regression and evaluated our models 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 results in *2.1*, though.

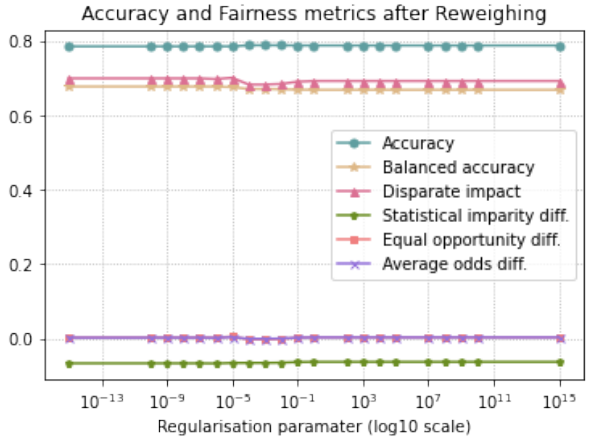

Figure 4. Averaged CV results on the validation set of the Adult-Dataset for varying levels of regularisation after reweighing.

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	Disparate impact	Statistical parity 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 than in the other direction. Selecting a model based on the fairness metrics is not as simple as in *2.1*. From the *Impossibility theorem* we know that the statistical parity and equal opportunity metrics are mutually exclusive, so it is better to choose one as a priority when selecting models. Disparate impact (which gives information similar to statistical parity) is generally considered to have more drawbacks than the equal opportunity difference metric, mainly because it does not consider the ratio of 'qualified' individuals from each group (privileged and unprivileged) in the dataset. Therefore, we will use the equal opportunity difference as our selection criterion (for more details see [1]). Coincidentally, this leads to selecting the same model as based on accuracy! We can see the performance on the held out test set for the model below.

	C	Accuracy	Balanced accuracy	Disparate impact	Statistical parity difference	Equal opportunity difference	Average odds difference
(3),(4)	e-03	0.8055	0.6774	0.6785	-0.0606	0.0587	0.0311

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

As a criterion to select a model that considers both fairness and accuracy, we will use the sum of error $(1 - accuracy)$ and the absolute value of the equal opportunity difference, which has been selected from the fairness metrics in *2.2*. Our best model will simply be the one with the lowest score.

	Task 1: Regularisation		
C	Accuracy score	Fairness Score	Final score
≤e-07*	0.2040	0.2208	0.4248
e-06	**0.2037**	**0.2125**	**0.4162**
e-05	0.2015	0.2379	0.4394
e-04	0.2011	0.2425	0.4436
e-03	0.1953	0.4506	0.6458
e-02	0.1951	0.4587	0.6538
e-01	0.1951	0.4610	0.6561
≥e+00*	0.1951	0.4620	0.6571

	Task 2: Reweighing		
C	Accuracy score	Fairness Score	Final score
≤e-06*	0.2126	0.0077	0.2203
e-05	0.2125	0.0087	0.2212
e-04	0.2097	0.0001	0.2097
e-03	**0.2097**	**0.0001**	**0.2097**
e-02	0.2104	0.0064	0.2168
≥e-01*	0.2105	0.0032	0.2137

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*.

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.

In 9 we can see the performance mean and standard deviation of the models trained in *2.1, 2.2* and *2.3* for five different random splits. For space reasons, only selected metrics are displayed, for full results please see the code submitted with this report.

	C	Fairness Score	Accuracy score	Final score
(5)	e-06	0.1892	0.2321	0.4213
(6)	e-03	0.1945	0.0587	0.2532

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

		Accuracy	Statistical parity difference	Equal opportunity difference
≤e-07*	avg	0.7973	-0.1512	-0.2155
	std	0.0085	0.0117	0.0421
e-06	**avg**	**0.7976**	**-0.1482**	**-0.2109**
	std	**0.0084**	**0.0152**	**0.0464**
e-05, e-04	avg	0.7990	-0.1495	-0.2333
	std	0.0085	0.0063	0.0419
e-03	**avg**	**0.8047**	**-0.2006**	**-0.4376**
	std	**0.0079**	**0.0039**	**0.0115**
e-02	avg	0.8041	-0.2037	-0.4412
	std	0.0070	0.0029	0.0113
≥e-01*	avg	0.8041	-0.2047	-0.4427
	std	0.0070	0.0029	0.0124

		Accuracy	Statistical parity difference	Equal opportunity difference
≤e-06*	**avg**	**0.7894**	**-0.0655**	**0.0243**
	std	**0.0087**	**0.0060**	**0.0249**
e-05	avg	0.7900	-0.0643	0.0281
	std	0.0088	0.0051	0.0241
e-04, e-03	**avg**	**0.7922**	**-0.0623**	**0.0258**
	std	**0.0087**	**0.0074**	**0.0219**
≥e-02*	avg	0.7912	-0.0597	0.0323
	std	0.0092	0.0092	0.0195

Figure 9. Averaged results (mean and standard deviation) for all the models trained in the previous sections (regularization on the left, reweighing on the right).
* The values of C used are described in the previous sections.

3. The CompasDataset

3.1. Task 1: Regularisation

As described in the introduction, we perform linear regression, this time including sex and race as sensitive features. In 10, we observe similar behaviour in terms of sensitivity to the value of C to *2.1*, but this time we see the opposite relationship, i.e. that we get better results when regularizing less. The accuracy is significantly lower than for our first data set, but it only changes within half a percentage point as we vary C. The best model based on accuracy is for $C = 10^{-1}$, and based on fairness $C = 10^{-2}$. Note that the parameter values are very close, unlike in *2.1*, and the performance differences are small, too.

3.2. Task 2: Fairness metrics

In this section, we perform reweighing analogically to *2.2*. While there is improvement in the equality of opportunity difference (as well as the other fairness metrics), it is much smaller than in *2.2*. The overall performance trends remain the same with and without reweighing, so only one plot was included (10) for space reasons. We can only observe moderate improvement in the values of the fairness metrics after reweighing, indicating that this approach

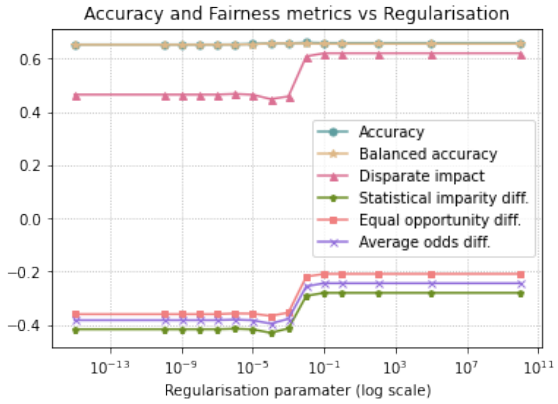

Figure 10. Averaged CV results on the validation set of the CompasDataset for varying levels of regularisation.

C	Accuracy	Balanced accuracy	Disparate impact	Statistical parity difference	Equal opportunity difference	Average odds difference
≤e-07*	0.6533	0.6523	0.3304	-0.6092	-0.5227	-0.5726
e-06	0.6537	0.6526	0.3304	-0.6092	-0.5227	-0.5726
e-05	0.6552	0.6540	0.3238	-0.6151	-0.5288	-0.5786
e-04	0.6579	0.6571	0.3065	-0.6308	-0.5405	-0.5939
e-03	0.6577	0.6574	0.3191	-0.5990	-0.5173	-0.5571
e-02	**0.6600**	**0.6566**	**0.4785**	**-0.4450**	**-0.3406**	**-0.4037**
≥e-01*	**0.6587**	**0.6554**	**0.4847**	**-0.4383**	**-0.3368**	**-0.3961**

Figure 11. Averaged CV results on the validation set of the CompasDataset for varying levels of regularisation. Each column is colour coded (green = better, red = worse).
* The values of C used are the same as in *Section 2*.

	C	Accuracy	Statistical parity difference	Equal opportunity difference
(1)	e-02	0.6496	-0.4095	-0.2998
(2)	e-01	0.6515	-0.3529	-0.2712

Figure 12. Performance on the test set for models selected (1) based on accuracy & (2) based on fairness on the validation set.

might not be the most suitable one (at least on its own) for this use case.

In 11 as well as 13, we can see, as already indicated above, that for the Compas Dataset the accuracy-fairness trade-off is much less present than for the Adult Dataset, with larger values of C yielding better accuracy as well as fairness.

3.3. Task 3: Final selection

Using the same methodology as in *Section 2*, we can see our model performances on our selected criterion below in 15. The best models corresponds to models (2) and (4) from *3.1* and *3.2*, i.e. the ones chosen based on the fairness metric. This is expected, as we saw the accuracy vary significantly less than the fairness metrics.

In 17, we see that the results on 5 random splits only differ minimally from the results obtained above, and we

C	Accuracy	Balanced accuracy	Disparate impact	Statistical parity difference	Equal opportunity difference	Average odds difference
≤e-06*	0.6524	0.6491	0.4860	-0.4214	-0.3407	-0.3759
e-05	0.6528	0.6494	0.4860	-0.4214	-0.3407	-0.3759
e-04	0.6583	0.6521	0.6090	-0.3214	-0.2187	-0.2744
e-03	**0.6611**	**0.6552**	**0.6257**	**-0.2992**	**-0.1927**	**-0.2512**
e-02	0.6608	0.6557	0.6361	-0.2846	-0.1826	-0.2338
e-01	0.6583	0.6526	0.6681	-0.2609	-0.1600	-0.2104
≥e+00*	**0.6562**	**0.6503**	**0.6758**	**-0.2553**	**-0.1570**	**-0.2051**

Figure 13. Averaged CV results on the validation set of the CompasDataset for varying levels of regularisation after reweighing. Each column is colour coded (green = better, red = worse).
* The values of C used are the same as in *Section 2*.

	C	Accuracy	Statistical parity difference	Equal opportunity difference
(3)	e-03	0.6591	-0.2796	-0.1696
(4)	e-00	0.6648	-0.2608	-0.1696

Figure 14. Performance on the test set for models selected (3) based on accuracy & (4) based on fairness on the validation set.

	Task 1: Regularisation				Task 2: Reweighing		
C	Accuracy score	Fairness Score	Final score	C	Accuracy score	Fairness Score	Final score
≤e-07*	0.3467	0.5227	0.8694	≤e-06*	0.3476	0.3407	0.6883
e-06	0.3463	0.5227	0.8690	e-05	0.3472	0.3407	0.6879
e-05	0.3448	0.5288	0.8737	e-04	0.3417	0.2187	0.5604
e-04	0.3421	0.5405	0.8826	e-03	0.3389	0.1927	0.5316
e-03	0.3423	0.5173	0.8596	e-02	0.3392	0.1826	0.5217
e-02	0.3400	0.3406	0.6806	e-01	0.3417	0.1600	0.5017
≥e-01*	**0.3413**	**0.3368**	**0.6780**	**≥e+00***	**0.3438**	**0.1570**	**0.5008**

Figure 15. Averaged CV results for both *Task 1* & *2* on the validation set using the selected criterion.
* Values of C used are described in *Section 2*.

	C	Accuracy score	Fairness Score	Final score
(5)	e-01	0.3485	0.2712	0.6197
(6)	e+00	0.3352	0.1696	0.5049

Figure 16. Performance on the test set for models selected (5) based on accuracy & (6) based on fairness on the validation set.

would mostly end up selecting the same parameter C for our models.

4. Additional Analysis

Finally, let's explore how the performance of the previously trained models changes when we exclude the sensitive features from the dataset.

4.1. AdultDataset

When only varying the regularization parameter, the performance changes marginally when we decide to simply exclude the 'sex' variable from the data when training and making predictions. We can observe a slight decrease in accuracy (about 0.2% in most cases) and a similarly small shift in the fairness metrics in both positive and negative directions. When applying re-weighing, we use the sensitive features to obtain new weights and then proceed to fit the model with data without the sensitive features. This approach results in a decrease (though a slight one, up to 0.5 percentage point) across all the metrics and values of C, showing that this naïve approach can also cause harm, as ignoring sensitive features does not mean one gets rid of bias in the model, since it is often present implicitly and across many features.

		Accuracy	Statistical parity difference	Equal opportunity difference
≤e-07*	avg	0.6439	-0.5364	-0.4758
	std	0.0234	0.0562	0.0730
e-06, e-05	avg	0.6451	-0.5433	-0.4808
	std	0.0246	0.0561	0.0751
e-04	avg	0.6515	-0.5744	-0.5020
	std	0.0225	0.0550	0.0754
e-03	avg	0.6530	-0.5332	-0.4758
	std	0.0210	0.0432	0.0724
e-02	**avg**	**0.6591**	**-0.4221**	**-0.3445**
	std	**0.0120**	**0.0560**	**0.0381**
e-01	avg	0.6568	-0.3928	-0.3212
	std	0.0145	0.0561	0.0579
≥e-00*	**avg**	**0.6564**	**-0.3890**	**-0.3140**
	std	**0.0138**	**0.0558**	**0.0583**

		Accuracy	Statistical parity difference	Equal opportunity difference
≤e-05*	avg	0.6439	-0.4038	-0.3558
	std	0.0195	0.0268	0.0844
e-04	avg	0.6432	-0.3256	-0.2688
	std	0.0202	0.0527	0.1217
e-03	avg	0.6515	-0.2775	-0.2039
	std	0.0144	0.0487	0.0270
e-02	**avg**	**0.6553**	**-0.2832**	**-0.2177**
	std	**0.0139**	**0.0450**	**0.0360**
≥e-01*	**avg**	**0.6519**	**-0.2371**	**-0.1711**
	std	**0.0152**	**0.0341**	**0.0288**

Figure 17. Averaged results (mean and standard deviation) for all the models trained in the previous sections (regularization on the left, reweighing on the right).
* The values of C used are described in the previous sections.

On the other hand, when we look at the second sensitive feature in this data set, which is race, we find that simply excluding it (no reweighing yet) helps improve the fairness metrics significantly and at *practically zero cost* in terms of accuracy. The difference is striking, with disparate impact values jumping to over 60% for all values of C (note that the best we achieved in *2.1* was $\sim$ 35%). Similarly, the highest equal opportunity difference drops by an order of magnitude. This shows that in this data set is biased much more in terms of race than sex. Interestingly, reweighing results in slight increase (about half a percentage point) in accuracy across all values of C compared to *2.2*.

4.2. CompasDataset

Excluding the 'race' feature from the Compas Dataset without reweighing yielded a $\sim$ one percentage point increase in accuracy as well as a noticeable improvement (around 10%) in the fairness metrics, though not as dramatic as in case of the Adult Dataset.

The above shows that while simply ignoring sensitive features is definitely not a universal solution to algorithmic bias, it can be a fast and simple step resulting in notable improvements, sometimes not even in terms of fairness, but accuracy as well. For more details and numeric results please refer to the code submitted with this report.

References

[1] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. 2016. 2

[2] ProPublica. The compas dataset, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. 1

[3] U. M. L. Repository. The adult data set, 1996. https://archive.ics.uci.edu/ml/datasets/adult. 1