yy3219

1. Introduction

This report aims to study model selection that considers accuracy and fairness metrics when training standard and fairness-based machine learning models. In particular, I performed logistic regression on two datasets from the `AIF360` library. For each dataset, reweighing may be applied, and I selected the regularisation parameter C based on three metrics, resulting in six models in total. The model performances are evaluated using 5-fold cross-validation (CV). The train/test split is 0.7/0.3.

1.1. Data

I used the AdultIncome dataset [1] and the Compas dataset [2] in the `AIF360` library. For the AdultIncome dataset, I chose ”sex” as the sensitive feature, inspired by the OECD study of the gender wage gap observed in many countries [3]. The aim is to predict whether a person makes over $50K a year. For the Compas dataset, I chose ”race” as the sensitive feature, inspired by the article [2] discussing how US criminals' predictions are biased against blacks. The aim is to predict whether a criminal would re-offend in two-years time.

1.2. Metrics

The performance metrics for model selection are accuracy, the Equality of Opportunity Difference (EOOD) and a self-proposed metric that combines accuracy and EOOD, which is accuracy - |EOOD|.

2. Results & Analysis

For each dataset, Table 1 and 2 shows the accuracy and fairness metrics for the six models on the test set. The following sections summarise key results and observations from both datasets, mainly illustrated by the plots of CV results for the AdultIncome dataset. More plots for the Compas dataset are in the Appendix.

AdultIncome dataset				
	Best C	Accuracy	EOOD	Acc - |EOOD|
M1	*1e+2	0.7994	-0.1548	-
M2	*1e-5	0.7622	0.0	-
M3	*1e+2	0.7854	0.0370	-
M4	*1e-5	0.7622	0.0	-
M5	*1e-7	0.7637	0.0	0.7637
M6	1e-1	0.7905	0.0351	0.7554

Table 1. The best C values for logistic regression and three metrics on the test set for the AdultIncome dataset. The * sign means there are multiple C values giving the same CV results and I select one of them randomly.

Compas dataset				
	Best C	Accuracy	EOOD	Acc - |EOOD|
M1	*1e+2	0.6806	-0.2081	-
M2	*1e-5	0.5922	-0.0952	-
M3	*1e+2	0.6597	0.0372	-
M4	*1e+8	0.6597	0.0372	-
M5	1e-2	0.6755	-0.1807	0.4948
M6	*1e+5	0.6597	0.0372	0.6226

Table 2. The best C values for logistic regression and three metrics on the test set for the Compas dataset. The * sign means there are multiple C values giving the same CV results and I select one of them randomly.

2.1. Model 1 & 2

By varying the regularisation parameter C in the range of $(1e-15, 1e+15)$, we select Model 1 and 2 with the highest accuracy and highest EOOD, respectively. Figure 1 shows the changes in accuracy and EOOD with respect to C values for the AdultIncome dataset.

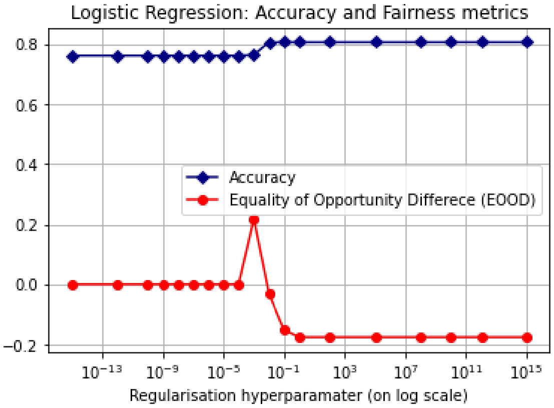

Figure 1. AdultIncome dataset: averaged CV results on the validation sets for different values of C (Model 1 & 2).

Varying the hyperparameter value C has an expected accuracy-fairness trade-off, as shown in Figure 1:

- Larger C values, $C \geqslant 1e-1$, have weak regularisation (as in Scikit-learn), which leads to higher accuracy but generally poorer fairness metrics (despite the spike at $C = 1e-2$).
- Compare the test metrics for Model 1 & 2 in Table 1 and 2, the absolute value of EOOD decreases for Model 2 at the expense of a drop in accuracy (although only a tiny drop for the AdultIncome dataset), showing the trade-off of accuracy and fairness. This suggests

that adopting the fairness metric for model selection could effectively correct bias without sacrificing too much of accuracy.

2.2. Model 3 & 4

This section applies a pre-processing fairness method, namely reweighing, to both datasets based on their chosen sensitive feature. We select Model 3 and 4 with the highest accuracy and EOOD, respectively, by varying the C values in the same range.

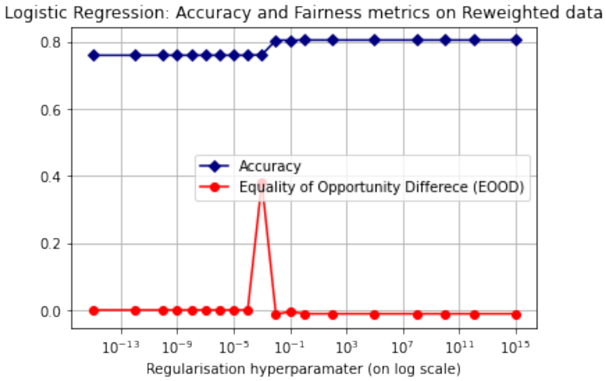

Figure 2. AdultIncome dataset: averaged CV results on the reweighed validation sets for different values of C (Model 3 & 4).

Varying the hyperparameter value does not display a significant accuracy-fairness trade-off, as shown in Figure 2:

- Larger C values, $C \geqslant 1e-1$, have weak regularisation, which leads to higher accuracy but almost no effect on the fairness metric (mostly around zero). This indicates that reweighing helps to correct bias in the model, as shown by the reduced magnitude of EOOD.

- Note that for the AdultIncome dataset, an outlier occurs at $C = 0.001$ where there is a big jump in bias towards the unprivileged group "female" (EOOD ≈ 0.4).

- From Table 1 and 2, we see that compared to Model 1 & 2, Model 3 & 4 have much smaller EOOD values (near zero) on the test set but similar accuracies, showing the effectiveness of reweighing in improving fairness for both datasets.

2.3. Model 5 & 6

To account for both accuracy and fairness, I proposed a new model selection metric: accuracy - |EOOD|, to account for the trade-off between the two metrics. We select models with C value(s) that maximises this criterion.

We select Model 5 and 6 that maximise this new metric on the original data and reweighed data, respectively, by varying the C values in the same range.

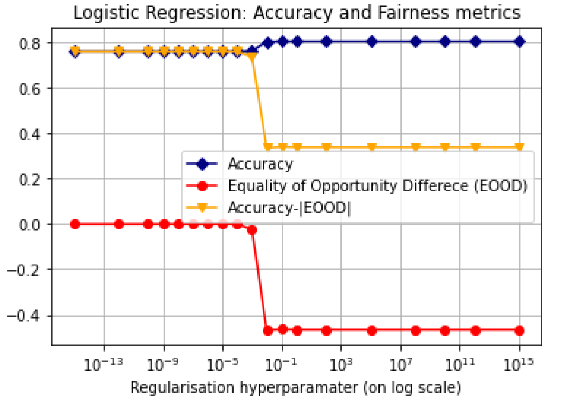

Figure 3. AdultIncome dataset: averaged CV results on the validation sets for different values of C (Model 5).

As shown in Figures 3 and 4, varying the hyperparameter value gives different trends in the new metric when applying to the original and reweighed data.

- In Figure 3, larger C values, $C \geqslant 1e-2$, have weak regularisation, which leads to higher accuracy but the poorer value in the new metric. As C increases, the proposed new metric follows a similar trend as the EOOD, with the same set of best C values.

- This shows that when there is bias in data and/or model with no effective correction (e.g. reweighing), the values of |EOOD| tend to be dominating in the proposed metric compared to accuracy.

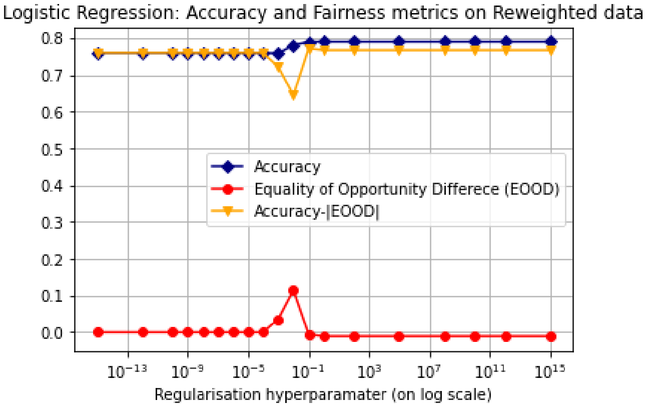

Figure 4. AdultIncome dataset: averaged CV results on the reweighed validation sets for different values of C (Model 6).

- In Figure 4, larger C values, $C \geqslant 1e-1$ leads to higher accuracy and also slightly higher values in the new metric. As C increases, the proposed new metric follows a similar trend as the accuracy (despite the big spike at $C = 0.001$ due to a large EOOD value). The EOOD values mostly stay around zero despite the change in C values.

- This suggests that after correcting the bias using reweighing, EOOD contributes very little to the new metric and accuracy becomes the main contribution.

While the above results apply to both datasets, when looking at test metrics for Model 5 and 6 in Table 1 and 2, we observe different results for each dataset:

- **AdultIncome dataset**: With reweighing (Model 6), the new metric on the test set is slightly lower than Model 5, with an increase in EOOD and a slight gain in accuracy, i.e. the best standard ML model outperforms the best fairness ML model. This shows that to combat biases, choosing a model based on the new metric (that accounts for both accuracy and fairness) could be a valuable alternative to fairness methods such as reweighing.

- **Compas dataset**: With reweighing (Model 6), the new metric on the test set is higher by 0.13 than Model 5, with a decrease in EOOD at a relatively small cost of accuracy. In this case, the best fairness ML model outperforms the best standard ML model. This shows that despite the effectiveness of the new metric, using fairness methods such as reweighing may help improve fairness further on top of using the new metric.

3. Critique and Improvements

- In this study, I used the same test set for all folds of CV. One possible improvement is splitting to train-val-test multiple times after data shuffling and averaging the metrics over multiple test sets.

- For both datasets, there are often multiple best hyperparameter values that give equal metric values on the validation set. I randomly chose one best C in such cases, but the test set's metric values could be slightly different from the other best Cs. One possible improvement is evaluating all the best values on the test set and averaging to produce more reliable test metrics.

- For each dataset, I only chose one (binary) sensitive feature out of 'sex' and 'race'. Further work could apply the same analysis to the other feature and compare whether the dataset is more biased towards 'sex' or 'race'.

- To generalise the results, other classification models such as (boosted) decision trees and k-nearest neighbours, potentially with reweighing, could be applied and compared with logistic regression.

4. Extrapolation

4.1. Exclude Sensitive Feature

This section explores the changes in performances for Model 1-6 when we exclude the sensitive feature ”sex” from the AdultIncome dataset and the sensitive feature ”race” from the Combat dataset.

AdultIncome dataset (removed ”sex”)				
	Best C	Accuracy	EOOD	Acc - |EOOD|
M1	*1e+3	0.7900	0.0113	-
M2	*1e-8	0.7637	0.0	-
M3	*1e+5	0.7905	0.0351	-
M4	*1e-9	0.7637	0.0	-
M5	*1e+3	0.7648	-0.0093	0.7555
M6	*1e+7	0.7596	0.0144	0.7452

Table 3. The best C values for logistic regression and three metrics on the test set for the AdultIncome dataset, after removin the feature ”sex”. The * sign means there are multiple C values giving the same CV results and I select one of them randomly.

Combat dataset (removed ”race”)				
	Best C	Accuracy	EOOD	Acc - |EOOD|
M1	*1e+3	0.6723	-0.1425	-
M2	*1e-8	0.6080	-0.0452	-
M3	*1e+5	0.6723	-0.1425	-
M4	*1e-9	0.6168	-0.0683	-
M5	*1e-9	0.6250	-0.0878	0.5372
M6	*1e-10	0.6250	-0.0878	0.5372

Table 4. The best C values for logistic regression and three metrics on the test set for the Compas dataset, after removin the feature ”race”. The * sign means there are multiple C values giving the same CV results and I select one of them randomly.

Observations:

- Compare Table 3 and 4 with Table 1, I found that changes to test metrics for Model 1-6 are marginal, with a maximum change of around 0.05 in accuracy and 0.10 in the magnitude of EOOD. The direction of changes also seems to be random. This suggests that removing the sensitive feature itself may not necessarily help improve models' fairness or accuracy.

- After removing the sensitive feature, one change I noticed is that, for both datasets, when we select models based on the new metric (Model 5 & 6), reweighing no longer help correct bias. This is illustrated by Figures 5 and 6 below (AdultIncome data), where the trend for EOOD does not change after reweighing (Model 6), thus giving the same trends for the new metric for Model 5 and 6. This suggests that removing the sensitive feature and applying the accuracy + fairness metric together could be an effective alternative to preprocessing methods such as reweighing.

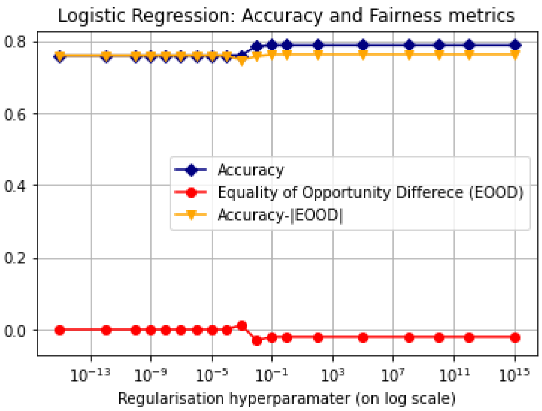

Figure 5. AdultIncome dataset (removed ”sex”): averaged CV results on the validation sets for different values of C (Model 5).

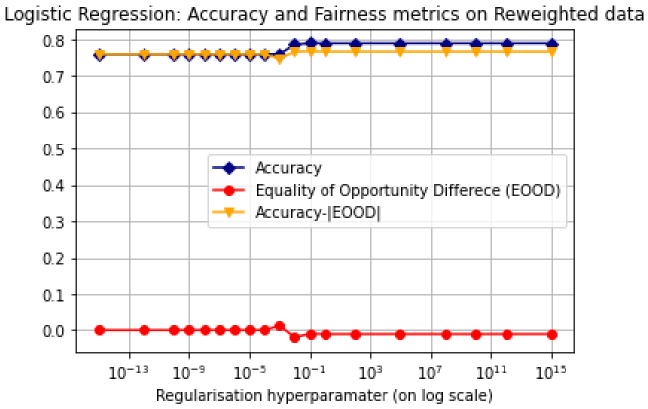

Figure 6. AdultIncome dataset (removed ”sex”): averaged CV results on the reweighed validation sets for different values of C (Model 6).

5. Appendix

5.1. Plots for Combat dataset (Task 1-3)

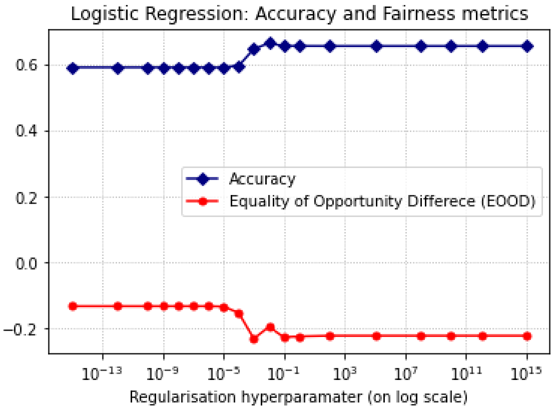

Figure 7. Combat dataset: averaged CV results on the validation sets for different values of C (Model 1 & 2).

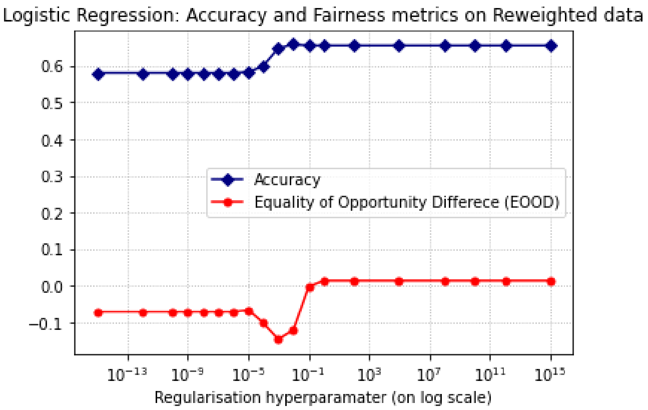

Figure 8. Combat dataset: averaged CV results on the reweighed validation sets for different values of C (Model 3 & 4).

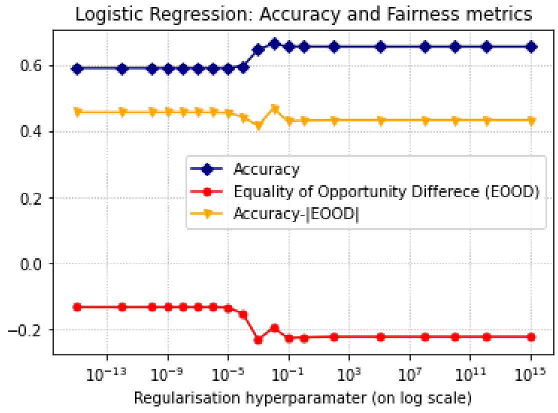

Figure 9. Combat dataset: averaged CV results on the validation sets for different values of C (Model 5).

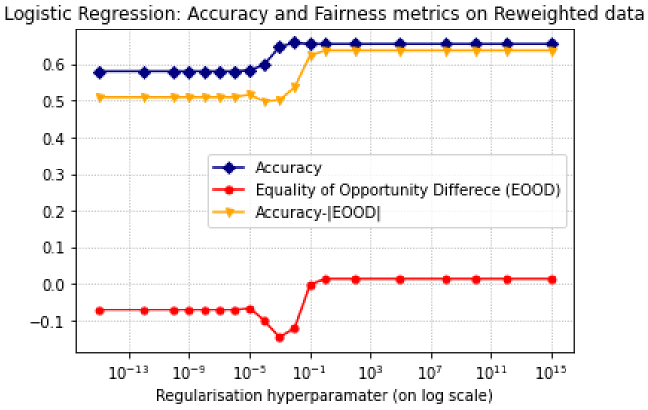

Figure 10. Combat dataset: averaged CV results on the reweighed validation sets for different values of C (Model 6).

5.2. Plots for Combat dataset (Removed "race")

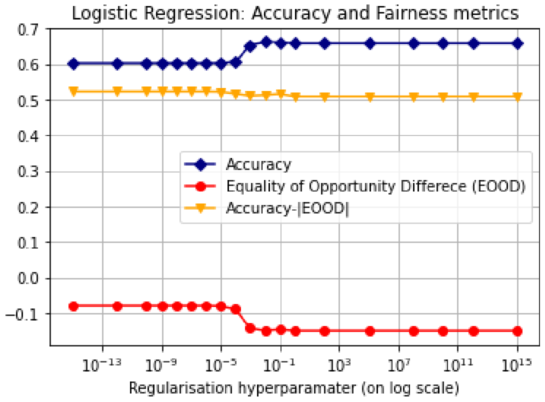

Figure 11. Compas dataset (removed ”race”): averaged CV results on the validation sets for different values of C (Model 5).

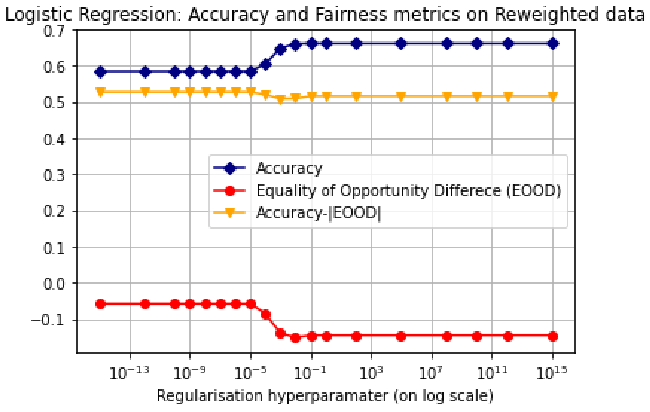

Figure 12. Compas dataset (removed ”race”): averaged CV results on the reweighed validation sets for different values of C (Model 6).

6. References

[1] Ronny Kohavi and Barry Becker, ”Adult Data Set”, 1996, predict whether income exceeds $50K/yr based on census data, https://archive.ics.uci.edu/ml/datasets/adult

[2] ProPublica, ”The Compas Dataset”, 2016, Machine Bias, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] OECD Data, ”Earnings and wages - Gender Wage Gap”, 2021, https://data.oecd.org/earnwage/gender-wage-gap.htm