Ion Stagkos Efstathiadis (CID: 01074696)

1. Task 1

For this task, the effect of regularization of classification models on the fairness of their classification is studied. Grid search of the regularization hyperparameters is performed for three types of classifiers, (logistic regression classifier, support vector classifier, random forest classifier) on three different datasets, (adult income, german, compas). For brevity, only results on the adult income dataset are reported. Some results on the compass dataset ca be found in the appendix. The effect of varying the regularization on accuracy and fairness metrics is reported. The hyperparameters giving the most accurate and most fair classification for the logistic regressor and the random forest on the adult income dataset are used to train a new model and test it against unseen test data.

Although different fairness metrics provide different advantages and disadvantages [1], for consistency across this report, average odds difference is used as the main indicator of fairness. The reason for this choice is that, unlike disparate impact and statistical parity difference for example, average odds difference considers the rate at which unprivileged samples are classified in each class in relation to the rate at which they actually occur in each class.

1.1. Cross Validation

Cross validation was used to study the effect of regularization on classifier accuracy and fairness for different models on different datasets. The results are shown only for the adult dataset for lack of space, but they were similar on all three datasets studied. Results on the compass dataset are found in the appendix. Note that the results were similar for all three classifier types.

The plotted figure shows that there is a range of regularization parameter values within which varying the regularization parameter influences both fairness and accuracy. In this report, this will be called the effective regularization spectrum. In the random forest classifier for example, (see Figure 3), this range is between 3 and 6. Within this range, increasing regularization causes accuracy to decrease by 2 to 5 percentage points but fairness to increase significantly, from unsatisfactory to satisfactory values.

To better understand the effect of regularization on fairness, Table 1 presents several regularization metrics for λ values at the two opposite sides of the effective regularization spectrum. All fairness metrics improve significantly as a result of increased regularisation. Interestingly, as portrayed by the confusion matrix of the unprivileged group, without regularization, the classifier classifies all members of the unprivileged groups as low-income, despite a minority of them actually having a high income. This is corrected by regularisation.

λ=	100	100000
Total Accuracy	0.80	0.79
Balanced Accuracy	0.66	0.67
Disparate Impact	0.00	0.34
Statistical Parity Difference	-0.22	-0.15
Equal Opportunity Difference	-0.46	-0.22
Average Odds Difference	-0.27	-0.14
Unprivileged Group **Confusion Matrix**	[12245 0] [1525 0]	[11561 684] [1139 386]

Table 1: Accuracy and Fairness metrics for regularization parameter "λ" values at the two opposite ends of the effective regularization range. This table is for the case of the logistic regression classifier on the adult income dataset.

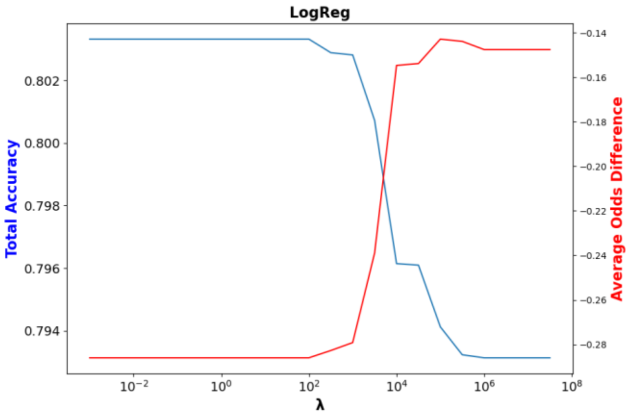


Figure 2: Accuracy and Fairness (represented as average odds difference) plotted against the regularization parameter "λ" for the case of a logistic regression classifier on the unweighted adult income dataset.

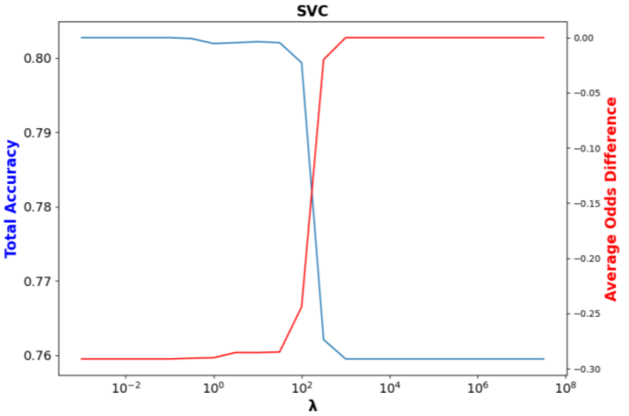


Figure 2: Accuracy and Fairness (represented as average odds difference) plotted against the regularization parameter "λ" for the case of a support vector classifier on the unweighted adult income dataset.

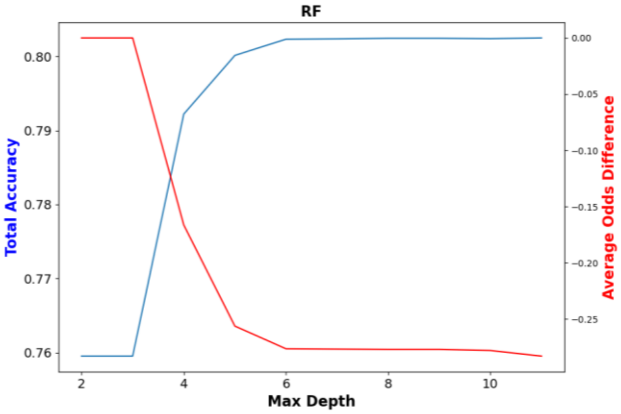


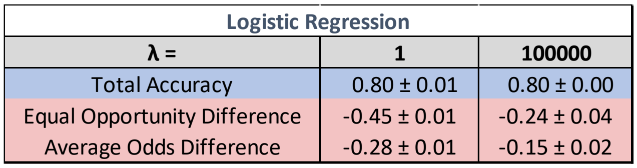
Figure 3: Accuracy and Fairness (represented as average odds difference) plotted against the regularization parameter "max depth" for the case of a random forest classifier on the unweighted adult income dataset. Note that here, unlike Figures 1 and 2, increasing the value of the regularization parameter reduces regularization.

1.2. Final Results

Cross validation allowed the identification of the regularization hyperparameter values yielding the fairest and most accurate classifiers. For each of these two values, and five different train-test splits, a model of each classifier type was trained and tested on each dataset. For brevity, only the tabulated results of the logistic regression and random forest classifiers on the adult income dataset are presented here.

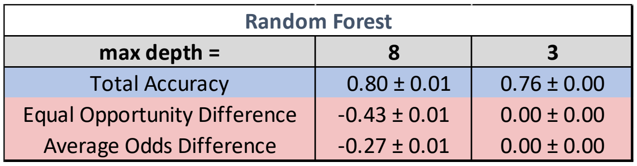
Tables 2 and 3 confirm the results found in the cross-validation section. They show that increased regularization causes a minor decrease in accuracy and a significant improvement in fairness. Note that low standard deviations in both tables testify towards the validity of the data.

Importantly, although both according to Figure 3 and Table 3, a random forest of maximum depth of 3 is a perfectly fair classifier, (average odds difference = 0 and equal opportunity difference = 0), inspecting the confusion matrix, (not included here), reveals that the classifier, simple as it is, assigns all samples, privileged and unprivileged, to the majority class. Although fair, this behaviour is far from desirable.



Logistic Regression		
λ =	1	100000
Total Accuracy	0.80 ± 0.01	0.80 ± 0.00
Equal Opportunity Difference	-0.45 ± 0.01	-0.24 ± 0.04
Average Odds Difference	-0.28 ± 0.01	-0.15 ± 0.02

Table 2: Accuracy and Fairness metrics of most accurate (λ=1) and fairest (λ=100000) logistic regression classifier on the adult income dataset. Metrics presented as: mean ± stdev.



Random Forest		
max depth =	8	3
Total Accuracy	0.80 ± 0.01	0.76 ± 0.00
Equal Opportunity Difference	-0.43 ± 0.01	0.00 ± 0.00
Average Odds Difference	-0.27 ± 0.01	0.00 ± 0.00

Table 3: Accuracy and Fairness metrics of most accurate (max depth = 8) and fairest (max depth = 3) random forest classifier on the adult income dataset. Metrics presented as: mean ± stdev.

2. Task 2

2.1. Cross Validation

In this task, similarly to task 1, the effect of regularization on the accuracy and fairness of different classification models is studied. This is done for three datasets, (adult income, german and compas) but for brevity, only the results on the adult income dataset are reported (see appendix for logistic regression results on compass dataset). This task differs from task 1 in that the data have been preprocessed to have the bias removed from them. This is achieved through the popular technique of reweighing.

As a result of reweighing, there is no inherent imbalance in the datasets that train the classifiers. This is reflected by the fact that for cases of SVC and RF classifiers, fairness seems to be nearly perfect irrespective of regularization (see Figures 5, 6). For a reason not yet discovered during the writing of this report, regularization within the effective regularization spectrum of the logistic regression classifier seems to significantly promote unfairness in favor of the initially unprivileged group. This is revealed by the positive values that the average odds difference takes in Figure 4.

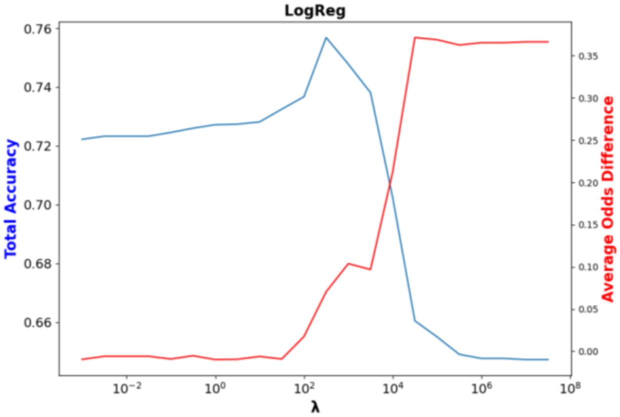


Figure 4: Accuracy and Fairness (represented as average odds difference) plotted against the regularization parameter "λ" for the case of a logistic regression classifier on the reweighted adult income dataset.

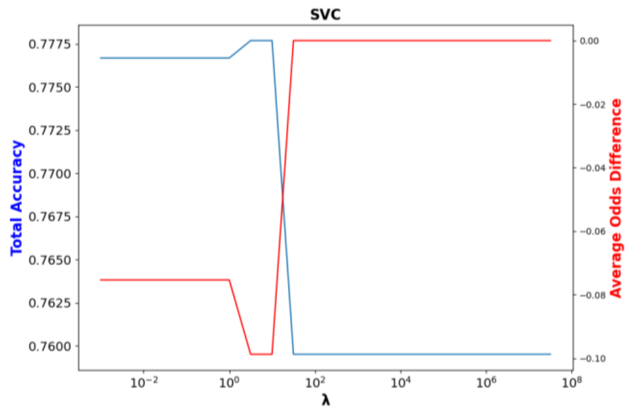


Figure 5: Accuracy and Fairness (represented as average odds difference) plotted against the regularization parameter "λ" for the case of a support vector classifier on the reweighted adult income dataset.

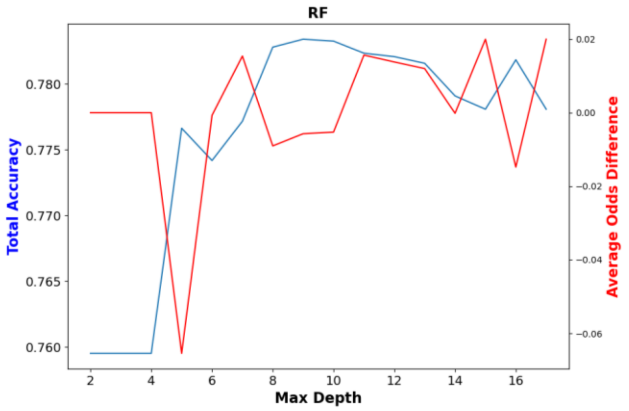


Figure 6: Accuracy and Fairness (represented as average odds difference) plotted against the regularization parameter "max depth" for the case of a random forest classifier on the reweighted adult income dataset. Note that here, unlike figures 4 and 5, increasing the value of the regularization parameter decreases regularization.

2.2. Final Results

Figure 6 shows no clear effect of regularization on the fairness of the random forest classifier. The minor fluctuations of fairness shown in the graph can well be attributed to noise. Nevertheless, random forest classifiers were trained and tested for regularization hyperparameter values maximizing fairness and accuracy according to Figure 6, in a procedure like the one described for task 1.

The results presented in Table 5 show equal mean accuracies and fairness metrics for the different max depths. This verifies the hypothesis that the fluctuations in accuracy and fairness displayed in Figure 6 are results of random noise in the data, rather than a long-term trend. It also hints towards the fact that when a dataset is perfectly balanced, (as it was here in effect due to reweighing), regularization does not clearly affect fairness.

Unlike for other classifiers, Figure 4 shows that increasing regularization in the case of a logistic regression classifier trained by reweighted data, promotes unfairness. Here, similarly to task 1, λ values giving the most accurate and most fair models were used in the final testing whose results are reported in Table 4. Table 4 shows identical results for the slightly different λ values for which it is produced. This is deemed a testament to the unclear effect of regularization on fairness and it insignificant effect on accuracy of models trained by reweighted data. Notably, the two different λ values used for Table 3 are close to each other, (as cross validation reported the most accurate and most fair classifiers occurring for similar λs).

Logistic Regression		
λ =	Most Accurate **316.22**	Most Fair **31.62**
Total Accuracy	0.79 ± 0.01	0.79 ± 0.00
Equal Opportunity Difference	-0.03 ± 0.02	-0.03 ± 0.02
Average Odds Difference	-0.01 ± 0.01	-0.01 ± 0.02

Table 4: Accuracy and Fairness metrics of most accurate (λ=316.22) and fairest (λ=31.62) logistic regression classifier on the reweighted adult income dataset. Metrics presented as: mean ± stdev.

Random Forest		
max depth =	Most Accurate **8**	Most Fair **4**
Total Accuracy	0.79 ± 0.01	0.78 ± 0.01
Equal Opportunity Difference	-0.10 ± 0.06	-0.09 ± 0.07
Average Odds Difference	-0.06 ± 0.04	-0.06 ± 0.04

Table 5: Accuracy and Fairness metrics of most accurate (max depth = 8) and fairest (max depth = 4) random forest classifier on the reweighted adult income dataset. Metrics presented as: mean ± stdev.

3. Task 3

For this task, results of 4 models are reported: **i)** a logistic regression without reweighing with such regularization as to reconcile accuracy and fairness, **ii)** same as in i, but for a case with reweighing, **iii)** same as in i, but for a random forest, **iv)** same as in ii, but for a random forest.

Table 6 shows that for the case of the logistic regression classifier without reweighing, taking the regularization value λ = $10^{3.5}$ which according to Figure 1 constitutes a middle solution between focusing solely on accuracy and solely on fairness, indeed yields fairness metrics in between the figures for most accurate and most fair reported in Table 2. On the other hand, with reweighing on, the accuracy is ever so slightly reduced, while fairness metrics are very close to 0 (approximating total fairness).

Table 7 shows that for a random forest classifier without reweighing, using a "max depth" = 4 to reconcile accuracy with fairness indeed works as expected (compare with Table 3). On the other hand, with reweighing on, a moderate "max depth" = 7 yields a negligibly reduced accuracy with near perfect fairness.

Logistic Regression		
λ =	Reweigh Off 316.22	Reweigh On 10
Total Accuracy	0.80 ± 0.00	0.79 ± 0.01
Equal Opportunity Difference	-0.38 ± 0.04	'0.03 ± 0.02
Average Odds Difference	-0.23 ± 0.03	'0.01 ± 0.02

Table 6: Accuracy and Fairness metrics of best-of-both-worlds logistic regression classifiers on the unweighted and reweighted adult income dataset. Metrics presented as: mean ± stdev.

Random Forest		
max depth =	Reweigh Off 4	Reweigh On 7
Total Accuracy	0.79 ± 0.00	0.78 ± 0.01
Equal Opportunity Difference	-0.27 ± 0.01	-0.04 ± 0.04
Average Odds Difference	-0.16 ± 0.01	-0.03 ± 0.02

Table 7: Accuracy and Fairness metrics of best-of-both-worlds random forest classifiers on the unweighted and reweighted adult income dataset. Metrics presented as: mean ± stdev.

4. Conclusion

The main take-home message from task 1 is that increasing regularization within the effective regularization spectrum drastically improves fairness while mildly reducing accuracy. This seems to hold across all classifier types and datasets.

The results of task 2 are slightly trickier to be interpreted. Similarly to task 1, increasing regularization seems to have a weak negative effect on accuracy across all classifiers. When it comes to fairness though, regularization does not seem to significantly affect it. This is to be expected as reweighing essentially renders the training dataset perfectly fair. This observation does not hold for the case of the logistic regression, where for a yet unknown reason, regularization seems to clearly hurt fairness.

In task 3, regularization values expected to offer a compromise between accuracy and fairness were chosen, and the results confirmed the expectations.

5. Future Work

The analyses performed in this report could in the future be expanded to include:

- A wider variety of fairness metrics such as predictive parity and calibration [1].
- A wider variety of bias mitigation algorithms such as adversarial debiasing and disparate impact remover [2] and their combinations.

6. Appendix

See below, cross validation graphs for tasks 1 and 2 of the logistic regression classifier on the compas dataset.

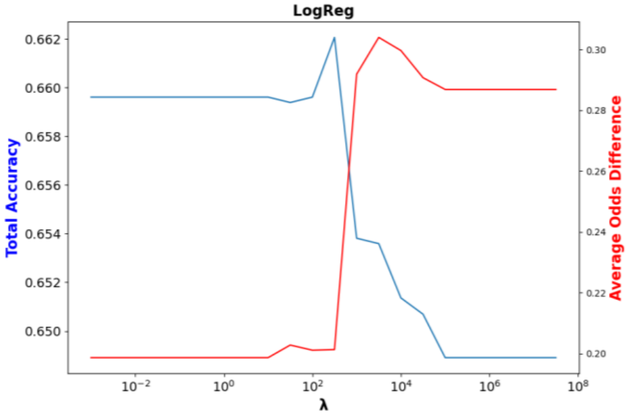


Figure 7: Accuracy and Fairness (represented as average odds difference) plotted against the regularization parameter "λ" for the case of a logistic regression classifier on the unweighted compas dataset.

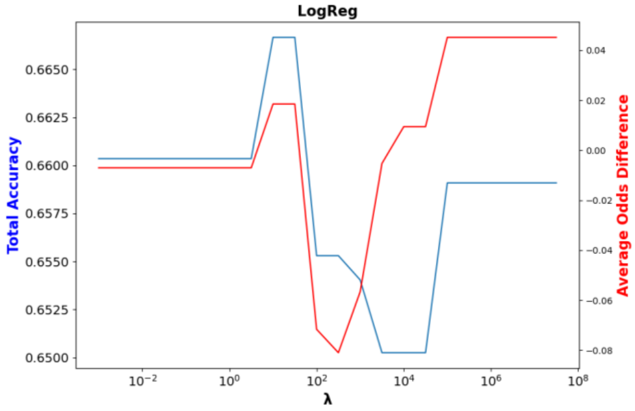


Figure 4: Accuracy and Fairness (represented as average odds difference) plotted against the regularization parameter "λ" for the case of a logistic regression classifier on the weighted compas dataset.

References:

[1] P. G. a. J. V. a. V. Foggo, "Fairness Metrics: A Comparative Analysis," 2020. [Online]. Available: https://arxiv.org/abs/2001.07864.

[2] IBM, "AI Fairness 360," IBM, 2021. [Online]. Available: https://aif360.mybluemix.net/. [Accessed 2021].