George Yiasemis (CID: 1108587)

Abstract

This assignment is part of the CO557 Ethics, Privacy, AI in Society module of Imperial College London. The task is the analysis of the effect of regularisation on accuracy-fairness trade-off obtained by a binary classifier on a dataset with sensitive/protected/private attributes.

1. Introduction

Nowadays, a lot of decisions in relation to our lives are being taken by artificial intelligent agents. Job recruitment, healthcare, university admissions, insurance coverage, loan provisions are only a few examples of decisions taken by AI. Such decisions can affect our lives, thus, it is of great importance that such agents can make not only accurate predictions but also be fair and/or unbiased. In machine learning, a given algorithm is said to be fair, or to have fairness if its outcome is independent of attributes considered to be private and more importantly, not related with it. Such attributes are: gender, age, ethnicity, sexual orientation, religion etc. This coursework requires the implementation of a regression model such as Logistic Regression and the analysis of the trade-offs between the accuracy of the model and some fairness metrics when a regularization factor is applied.

2. Data

In this report with the help of aif360 library[2] we are going to be using two widely-used datasets:
Adult Dataset: The target attribute of this dataset is an individual's binary ($\leq$ 50k or $>$ 50k) annual income results from various factors. Intuitively, it is influenced by the individual's education level, age, gender, occupation, and etc. The dataset contains 2 sensitive binary information (sex, race). It contains approximately 48,000 samples [3].

COMPAS Dataset: Correctional Offender Management Profiling for Alternative Sanctions is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending (recidivism). The target attribute of COMPAS is an offender's (binary) likelihood of reoffending in the future based various factors. The private attribute used for this dataset is the race (Caucasian-privileged-: and non-Caucasian). It contains approximately 5300 samples (2100 in the privileged class)[1].

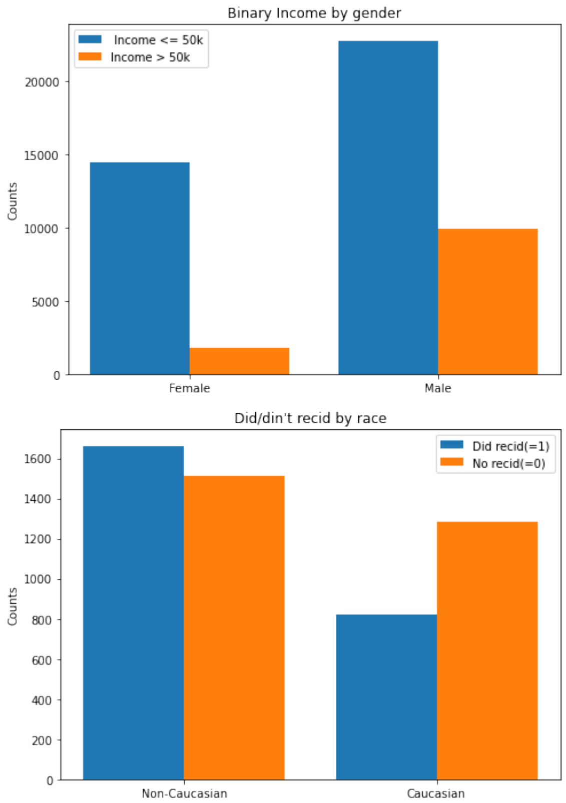

Figure 1. Bar plots of the counts in each class with respect to the target attribute for the two datastets; top: Adult Dataset, bottom: COMPAS Dataset

Figure 1 illustrates the unbalance of the two datasets between the privileged and non-privileged classes. For example, the first graph points out in the case of the Adult dataset, that there exists bias in favor of Male adults, as the percentage of the high-paid males (30%) is much greater than the percentage of the high-paid females (11%).

3. Logistic Regression

Given a training dataset

$$\mathcal{D} = \{(X, Y)\} = \{(\vec{x}_1, y_1), ..., (\vec{x}_N, y_N)\}$$

where $X \in \mathbb{R}^{N \times d}$ are the input features, $\mathcal{A} \in \{0,1\}^N$ the sensitive attribute and $Y \in \{-1,1\}^N$ the target binary label, in Logistic Regression(LR) our objective is to minimise the logistic loss:

$$\sum_{i=1}^{N} -\log\Big(\sigma(y_i \vec{w}^T \vec{x}_i)\Big) \quad (1)$$

where $\vec{w}$ corresponds to the learnable parameters of the regression model and $\sigma : \mathbb{R} \longrightarrow (0,1)$ to the logistic function, i.e. $\sigma(z) = \dfrac{1}{1 + exp(-z)}$.

3.1. Regularization

Regularization is a method used in Machine Learning used to avoid overfitting when training an algorithm. It is implemented by adding a "penalty" term when minimising the loss function during training. It can achieve a lower variance when the algorithm is used to make predictions on the test. We minimise instead of (1):

$$\lambda||\vec{w}||_2^2 - \sum_{i=1}^{N} \log\left(1 + \exp(-y_i \vec{w}^T \vec{x}_i)\right)^{-1} \tag{2}$$

where λ is a non-negative hyper-parameter. Note that the greater the value of λ is, the greater the effect of the regularization is. In this assignment we test its effect on the trade-off between accuracy and fairness of logistic regression on our datasets by tweaking its value.

4. Fairness

4.1. Fairness Metrics

An algorithm is said **not** to be fair when its decisions are taken based on the private attributes of the data that it has been trained on. There are various definitions/ criteria of fairness that have been proposed in the literature. Some of those are Statistical parity, Equalized Odds and Predictive Parity. However, the "Impossibility Theorem" states that any two of these criteria are mutually exclusive except in non-degenerate cases. [6].

4.1.1 Equalized Odds

Let $\mathcal{C} = \mathcal{C}(X, \mathcal{A}) \in \{0, 1\}$ be a predictor. $\mathcal{C}$ satisfies Equalized Odds (EO) with respect to the protected attribute A and target Y, if "$\mathcal{C}$ is independent of $\mathcal{A}$ conditioned on Y" [5], in other words the following equations hold:

$$\mathbb{P}[\mathcal{C} = 1|\mathcal{A} = 0, Y = 1] = \mathbb{P}[\mathcal{C} = 1|\mathcal{A} = 1, Y = 1] = \text{TPR}$$
$$\mathbb{P}[\mathcal{C} = 0|\mathcal{A} = 0, Y = 0] = \mathbb{P}[\mathcal{C} = 0|\mathcal{A} = 1, Y = 0] = \text{TNR}.$$

Note that TPR denotes the True Positive Rate and TNR the True Negative Rate of each protected class. Equalized Odds is satisfied when the classifier yields equal TPRs and TNRs across both (or all if more than two) protected classes.

For example, for a recruitment AI tool, it would be told to have EO if its acceptance and rejection rates of the qualified and unqualified respectively, from each of the sensitive groups would be equal.

4.1.2 Demographic Parity

Let $\mathcal{C} = \mathcal{C}(X, \mathcal{A}) \in \{0, 1\}$ be a predictor. $\mathcal{C}$ satisfies Demographic Parity (DP) with respect to the protected attribute A, if "$\mathcal{C}$ is independent of $\mathcal{A}$" [5], in other words the following equation holds:

$$\mathbb{P}[\mathcal{C} = 1|\mathcal{A} = 0] = \mathbb{P}[\mathcal{C} = 1|\mathcal{A} = 1] = \frac{\text{TP+FP}}{\text{TP+FP+TN+FN}} = \text{PR}.$$

For a recruitment AI tool, DP would be satisfied if its acceptance rates of each of the sensitive groups are equal.

4.1.3 Predictive Parity

Let $\mathcal{C} = \mathcal{C}(X, \mathcal{A}) \in \{0, 1\}$ be a predictor. $\mathcal{C}$ satisfies Predictive Parity (PP) with respect to the protected attribute A and target Y, if "Y is independent of $\mathcal{A}$ conditioned on $\mathcal{C}$", in other words the following equations hold:

$$\mathbb{P}[Y = 1|\mathcal{A} = 0, \mathcal{C} = 1] = \mathbb{P}[Y = 1|\mathcal{A} = 1, \mathcal{C} = 1]$$
$$\mathbb{P}[Y = 0|\mathcal{A} = 0, \mathcal{C} = 0] = \mathbb{P}[Y = 0|\mathcal{A} = 1, \mathcal{C} = 0].$$

4.2. Fairness Methods

Fairness methods are methods used in artificial learning algorithms that aim to combat biases in the data and achieve algorithmic fairness.
These methods can be split into three categories [7]:
(i) Pre-processing: Pre-processing techniques aim to remove the underlying discrimination before the classifier is learned.
(ii)In-processing: In-processing techniques aim to remove discrimination during the training of the classifier by constraining learning with fairness metrics.
(iii) Post-processing: Post-processing techniques aim to remove discrimination from a trained model using fairness metrics.

4.2.1 Pre-processing

In this assignment we are using the pre-processing technique to examine the trade-offs between accuracy and fairness when applying regularization. The method of pre-processing used here is **Reweighing** the training data, i.e. those with higher weight are used more often and those with lower weight are used less frequently.
The weights $W_{(Y=y,\mathcal{A}=a)}$ for $y \in \{-1, 1\}$ and $a \in \{0, 1\}$ are calculated using:

$$W_{(Y=y,\mathcal{A}=a)} = \frac{\mathbb{P}[Y = y]\,\mathbb{P}[A = a]}{\mathbb{P}[Y = y, A = a]}. \tag{3}$$

Then we train the model by minimising:

$$\lambda||\vec{w}||_2^2 - \sum_{i=1}^{N} W_{(Y=y_i,\mathcal{A}=a_i)} \log\left(1 + \exp(-y_i \vec{w}^T \vec{x}_i)\right)^{-1} \tag{4}$$

5. Results

In this section we present our results on both, Adult and COMPAS datasets. We investigate the effects of regularization on accuracy-fairness trade-off by training a LR model in each case while varying the regularization term λ.

5.1. Adult Dataset

Figures 3 and 4 and Table 2 show how varying the regularization parameter affects the accuracy of Logistic Regression on the hold-out test set as well as the fairness metrics without and with performing pre-processing, respectively. Note that for the results presented in Figures 3 and 4 and Table 2 we split the data in 70% training and 30% test data.

5.1.1 Without Pre-processing

Observing Figure 3 and Table 2 we can see that varying the regularization parameter λ slightly affects the accuracy as we have a slight decrease from 80.18% ($\lambda = 0$, no regularization) to 79.52% ($\lambda = 100$). Moreover, we can observe from Figure 3 and 2 that none of the fairness metrics (Equalized Odds, Predictive Parity, Demographic Parity) are satisfied, i.e. the algorithm with no pre-processing is not fair. Additionally, it seems that regularization does not affect fairness either as the fairness metrics get a slight change.

5.1.2 With Reweighing

From Figure 3 and Table 2 we can see that varying the regularization parameter λ in the case of Reweighing affects the accuracy which decreases from 79.55% ($\lambda = 0$, no regularization) to 68.18% ($\lambda = 100$). However, we can observe that regularization affects fairness either as the fairness metrics improve. Figure 3 shows that Equalized Odds is satisfied as the TPRs and TNRs of the two groups are approximately equal for λ. Nevertheless, as expected since Equalized Odds is satisfied, the other two fairness criteria are not. Therefore, we can conclude that regularization improves fairness but worsens accuracy.

5.2. COMPAS Dataset

Figures 5 and 4 show how varying the regularization parameter affects the accuracy of Logistic Regression on the hold-out test set as well as the fairness metrics without and with performing pre-processing, respectively. Note that for the results presented we split the COMPAS dataset in 80% training and 20% test data.

5.2.1 Without Pre-processing

Observing Figure 5 we can see that varying the regularization does not really affect the accuracy nor the fairness metrics as they remain mostly unchanged for all the values of λ. Moreover, the only metric that it seems to be satisfied partially is Predictive Parity as PPP is equal in the two groups.

5.2.2 With Reweighing

Observing Figure 6 we see that varying the regularization parameter λ in the case of Reweighing slightly affects the accuracy and fairness metrics as they fluncuate around a single value. However, Figure 6 shows that Equalized Odds is satisfied as the TPRs and TNRs of the two groups are approximately equal for λ. Nevertheless, as expected since Equalized Odds is satisfied, the other two fairness criteria are not.

6. Model Selection

Our aim is to select the best model possible in terms of maximising simultaneously both accuracy and fairness. We have seen how varying the regularization parameter when training a model on reweighed data can improve fairness, but this comes with a price; a reduced accuracy.
As we seen in section 5, applying reweighing yields a fair classifier with respect to the Equalized Odds criterion. Our approach here is to choose the model (and therefore the value of the regularization parameter) with the higher accuracy but also balances TPR and TNR, i.e. the probability of predicting as capable those who are truly capable and the probability of predicting as not capable those who are not, are equal. Our experiments have shown in the case of the Adult dataset that the model with $lambda = 50$ yields an accuracy of approximately 71% while $TPR = TNR \approx 0.7$ for both sensitive attributes (see Figure 4).
In Table 1 are presented the results of that model (mean and standard deviation) over **ten** 70/30% random train/test splits.

Table 1. Random train/test splits results for model selection

Accuracy(%)		Male/ Female				Absolute Difference			
		TPR(%)		**TNR**(%)		**TPR**(%)		**TNR**(%)	
Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
71.0	0.4	70.2/ 70.5	0.9/ 1.5	71.8/ 70.1	0.9/ 0.9	1.0	0.9	1.7	0.8

7. Extrapolation of what was taught

7.1. Analysis of algorithmic fairness methods beyond binary sensitive features

7.1.1 Categorical sensitive-feature

Given a training dataset $\mathcal{D} = \{(\vec{x}_1, y_1), ..., (\vec{x}_N, y_N)\}$ where $X \in \mathbb{R}^{N \times d}$ are the input features, $\mathcal{A} \in \{0, 1, .., M-1\}^N$ the sensitive attribute and $Y \in \{-1, 1\}^N$ the target binary label, where M denotes the number of classes of the sensitive feature. For example, if we assume that the sensitive feature is "race" then $\mathcal{A}$ could lie in $\{0, 1, 2, 3, 4\}$ where

each integer corresponds to ”White”, ”Black”, ”Latino”, ”Asian”, ”Malayan”, etc. In such a case we treat the classification in the same way that we would as in the binary case with the difference that fairness metrics should be satisfied for all sensitive classes.
For instance, the Equalized Odds metric would be satisfied if: ”$\mathcal{C}$ is independent of $\mathcal{A}$ conditioned on Y”, in other words the following equations hold:
$\mathbb{P}[\mathcal{C}=1|\mathcal{A}=i, Y=1] = \mathbb{P}[\mathcal{C}=1|\mathcal{A}=j, Y=1] = \text{TPR}$
$\mathbb{P}[\mathcal{C}=0|\mathcal{A}=i, Y=0] = \mathbb{P}[\mathcal{C}=0|\mathcal{A}=j, Y=0] = \text{TNR}$
for all $i, j \in \{0, 1, .., M-1\}$.

7.1.2 Continuous sensitive-feature

Given a training dataset $\mathcal{D} = \{(\vec{x}_1, y_1), ..., (\vec{x}_N, y_N)\}$ where $X \in \mathbb{R}^{N \times d}$ are the input features, $\mathcal{A} \in (a,b)^N$ the sensitive attribute and $Y \in \{-1, 1\}^N$ the target binary label, where the interval $(a,b) \subset \mathbb{R}$ denotes a range of real numbers. For example, if we assume that the sensitive feature is continuous such as ”age” then we could split the interval into M sub-intervals and treat it as the previous case.

7.1.3 Reweighing

As a fairness method Reweighing can be used similarly to the binary case. The weights for $y \in \{-1, 1\}$ and $a \in \{0, .., M-1\}$ are calculated using:

$$W_{(Y=y, \mathcal{A}=a)} = \frac{\mathbb{P}[Y=y]\,\mathbb{P}[A=a]}{\mathbb{P}[Y=y, A=a]}.$$

7.1.4 Results and Analysis on Adult dataset

To perform an analysis in the case of non-binary sensitive-features we use again the Adult dataset but now we consider both race and sex, as sensitive attributes, therefore we have four sensitive groups; White Males, Non-white males, White Females and Non-white females. We consider as the privileged group the White Males. Figure 2 illustrates the counts in each sensitive group with respect to their Binary Income ($\leq 50k$, $> 50k$). We perform a 70%-30% train/test split. The method of reweighing is then applied on the train set and then a LR model is trained. We check the accuracy and Equalized Odds metric on the test set. Figure 7 and Table 3 show the results for different values of the regularization parameter. We can observe, that regularization enforces the trade-off between accuracy and fairness as accuracy drops while Equalized Odds is satisfied. Note that these results are consistent with those in the binary case.

8. Git Repository

You can find the implementation’s code here: https://gitlab.doc.ic.ac.uk/gy615/fairness-coursework.

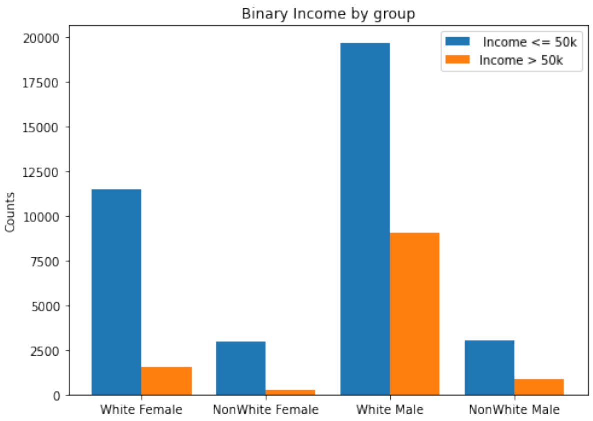

Figure 2. Bar plots of the counts for the four sensitive groups in the Adult Dataset

9. Conclusion

It is important that AI algorithms are fair as they are increasingly being used and have an impact in our lives. In this report, by using a pre-processing technique, Reweighing, we have shown how fairness can be achieved through some metrics. Moreover, we have explored the effect of regularization in Logistic Regression on accuracy-fairness trade-offs. Trade-offs between accuracy and fairness usually exist. Regularization can improve fairness in a biased dataset that favors a specific group and disfavors another. However, we have seen that the accuracy of the model may worsen. Therefore, in applications of AI in which either the data are skewed or unbalanced and there is a danger of unfair predictions, a threshold must be decided between fairness and accuracy must be set depending on the application.

10. Discussion/ Future Work and Possible Extensions

In this paper we have only investigated some aspects of the fairness-accuracy trade-offs when using regularization. There are several possible ideas for future work:
A fairness metric that could be used is the Absolute Between-ROC Area (ABROCA) statistic as proposed in [4]. ABROCA is defined as the absolute difference between the are under the ROC curves (AUC) of each group ($|AUC_{\text{priv}} - AUC_{\text{unpriv}}|$). Note that the lower is the value of this metric the fairer would be predictor.

An other possible approach of training a fair and accurate model would be to ”constraint” our objective function by placing a fairness constraint when minimising our loss function in (4). An approach of this is presented in [8]. As an extension of [8] we minimise the loss function with respect to both the parameters and the regularization function.

References

[1] M. Barenstein. Propublica's compas data revisited, 2019. 1

[2] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, Oct. 2018. 1

[3] D. Dua and C. Graff. UCI machine learning repository, 2017. 1

[4] J. Gardner, C. Brooks, and R. Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics Knowledge*, LAK19, page 225–234, New York, NY, USA, 2019. Association for Computing Machinery. 4

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

[6] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores, 2016. 2

[7] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning, 2019. 2

[8] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118, New York, NY, USA, 23–24 Feb 2018. PMLR. 4

11. Appendix

11.1. Figures and Tables

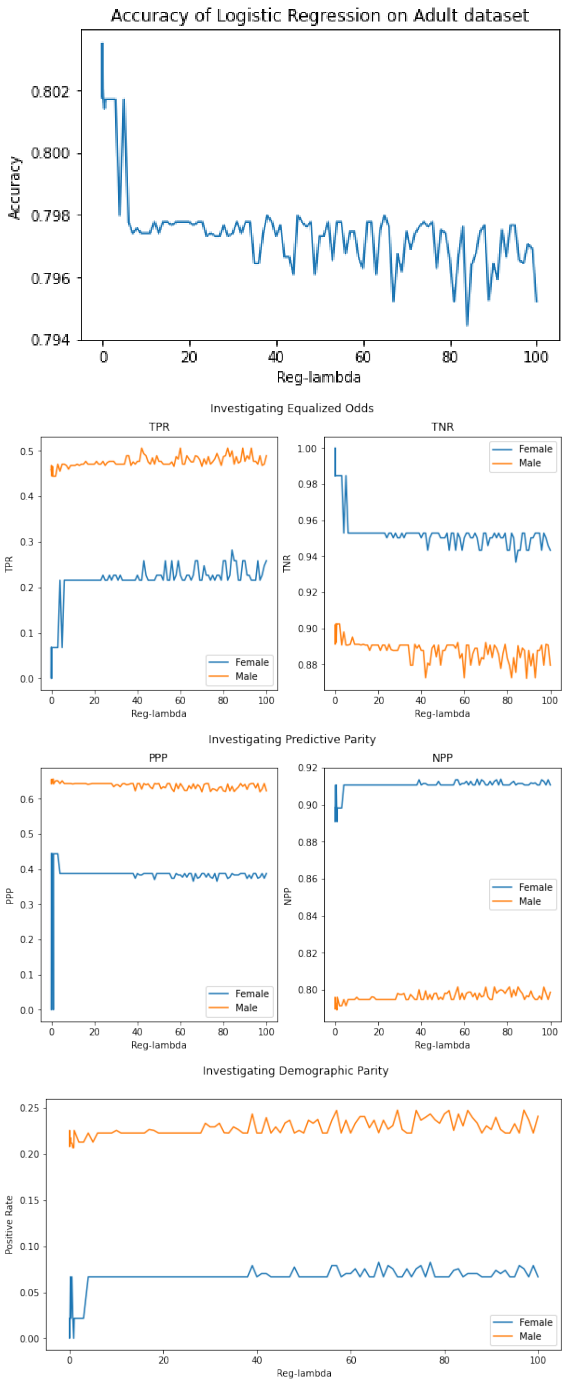

Figure 3. Accuracy and fairness metrics when varying the regularization parameter for the Adult dataset with no pre-processing done. From top to bottom: Accuracy, TP and TN rates (Equalized Odds), Positive and Negative Predictive Parity, Positive Rate (Demographic Parity)

Table 2. Accuracy and fairness of Logistic Regression for the Adult dataset for varying λ

λ	W/o Reweighing							With Reweighing						
	ACC (%)	**Fairness**						**ACC (%)**	**Fairness**					
		Male			Female				Male			Female		
		TPR(%)	TNR(%)	PR(%)	TPR(%)	TNR(%)	PR(%)		TPR(%)	TNR(%)	PR(%)	TPR(%)	TNR(%)	PR(%)
0.0	80.18	45.5	89.8	21.3	6.77	98.5	2.17	79.55	44.4	90.2	20.4	28.2	93.7	13.8
0.5	80.14	46.6	89.2	21.2	6.77	98.5	6.67	78.75	41.1	91.0	20.4	41.1	90.5	13.8
10	79.74	47.0	89.1	22.5	21.6	95.2	6.67	78.17	51.6	87.0	26.5	50.2	86.3	19.3
30	79.74	47.0	89.0	22.9	22.6	95.0	6.67	74.61	62.3	77.9	34.3	62.3	75.3	29.2
50	79.73	47.7	88.8	22.5	22.7	95.0	6.67	71.22	68.0	73.5	40.3	70.0	70.1	34.4
80	79.66	48.7	88.3	24.3	25.9	94.3	6.67	70.73	69.5	72.3	41.6	69.7	69.3	35.7
100	79.52	48.9	87.9	24.1	26.0	94.3	6.67	68.18	75.6	66.5	40.5	76.5	65.0	35.1

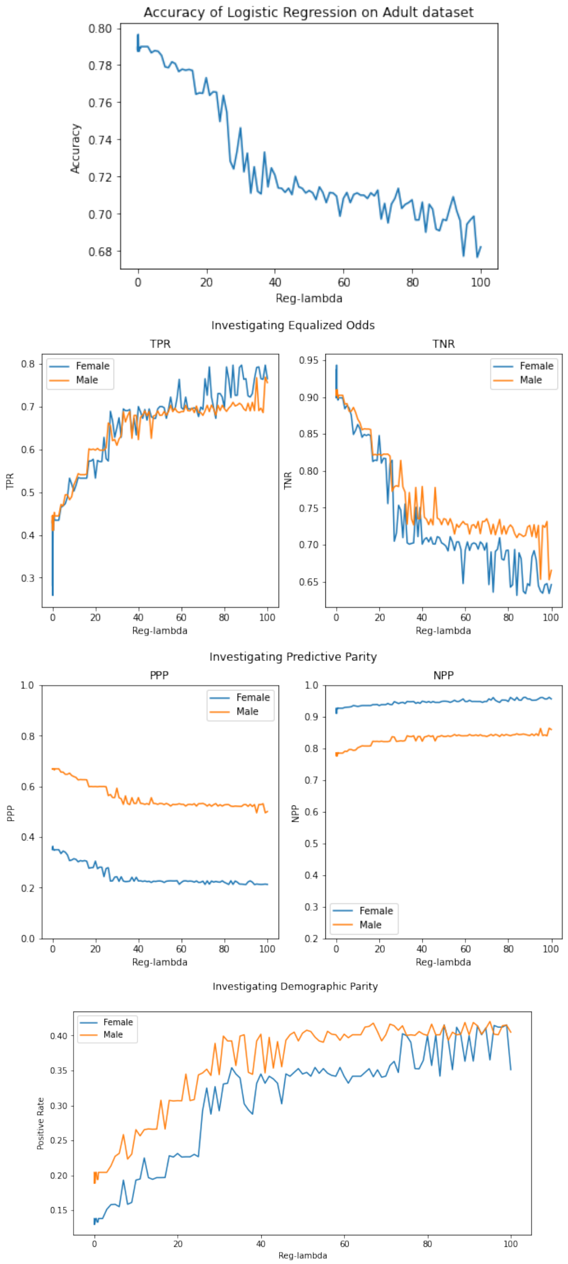

Figure 4. Accuracy and fairness metrics when varying the regularization parameter for the Adult dataset with Reweighing. From top to bottom: Accuracy, TP and TN rates (Equalized Odds), Positive and Negative Predictive Parity, Positive Rate (Demographic Parity)

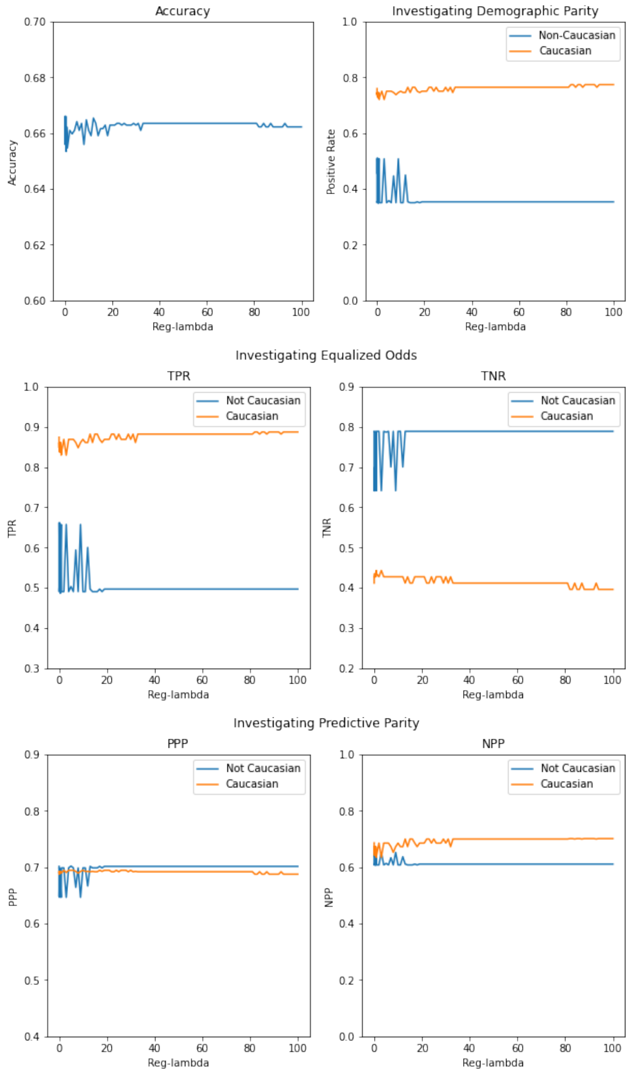

Figure 5. Accuracy and fairness metrics when varying the regularization parameter for the COMPAS dataset. Top: Accuracy (left), Positive Rate /Demographic Parity (right) — Middle: TP and TN rates (Equalized Odds) — Bottom: PPP and NPP

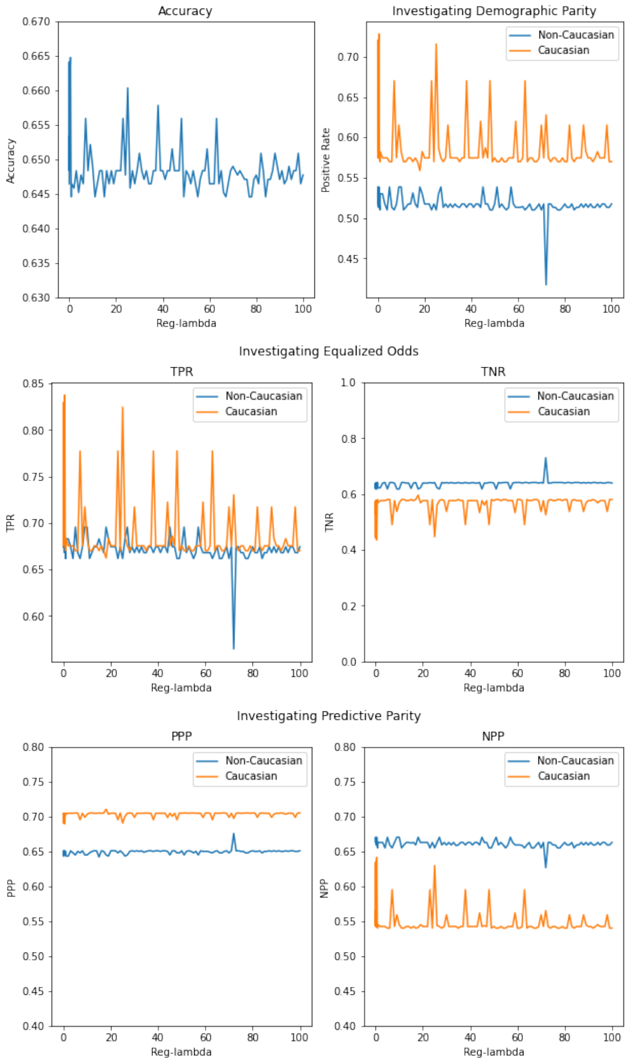

Figure 6. Accuracy and fairness metrics when varying the regularization parameter for the COMPAS dataset with Reweighing. Top: Accuracy (left), Positive Rate /Demographic Parity (right) — Middle: TP and TN rates (Equalized Odds) — Bottom: PPP and NPP

Table 3. Effect of regularization on a non-binary sensitive feature

λ	ACC (%)	**Fairness (%)**							
		White Males		Nonwhite Males		White Females		Nonwhite Females	
		TPR	**TNR**	**TPR**	**TNR**	**TPR**	**TNR**	**TPR**	**TNR**
0.0	79.0	46.6	89.7	30.0	95.1	43.6	88.2	42.5	89.8
10	78.1	50.2	88.0	56.5	85.0	47.6	85.9	53.8	88.0
30	75.8	61.8	81.4	66.1	81.0	58.8	78.5	61.3	78.9
65	70.0	71.0	71.7	71.7	71.4	70.0	66.3	68.8	67.6
100	69.8	71.0	71.7	71.7	71.4	72.5	65.0	68.8	67.6

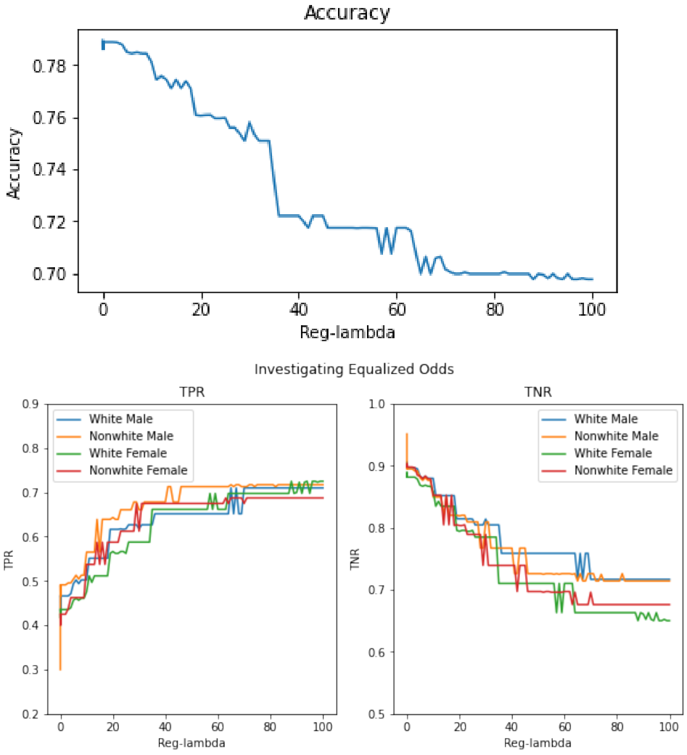

Figure 7. Accuracy and fairness metrics when varying the regularization parameter in the case of non-binary sensitive feature (Adult Dataset). Top: Accuracy Bottom: Equalized Odds