# Student Number: 237740

## 1. Introduction

The subject of algorithmic bias has received increased interest over the past decade. In the field of machine learning the data that is used to train models often can contain ingrained bias within its attributes given the nature of what type of information is gathered, ie. race, gender and age. Specifically, age and sex for our given problem.

The issue of models over fitting to training data is not a new topic and there are many strategies invented to combat this issue. In this paper, the first task takes a look at the Euclidean Norm and is explored within the Logistical Regression model. The regularization parameter *λ* being varied to exam the effects generalization has on the fairness and accuracy of the models predictions. Upon tuning the regularization parameter the technique of adversarial mitigation was introduced to directly combat the bias during the training of the models. The use of Beutel et al [1], technique of pre-training the classifier and adversarial models in an effort to improve the stability at their individual tasks. In conjunction with this method, the use of the input rules for the adversarial network provided by Zhang et al [2]. Here, it is described that in order to train the model to remove de-bias of chosen fairness metric, the adversarial network receives specific inputs. In regards to the task at hand, equality of opportunity, the adversarial network should receive both the classifiers output and the ground truth label, as explained in both papers, it is only examined on cases when the ground true label is 1. That is, the successful outcome/opportunity. Additionally, equalized odds is looked at because as Hardt described that equal opportunity is a weaker notion of non-discrimination of the highest forms of algorithmic fairness equalized odds [3] . The difference between the two fairness metrics is discussed within the Methodology section. Furthermore, the projection term Zhang et al recommends to include in the loss update of the classifier is included. Due to the prone errors in training adversarial it was chosen to implement in task the method that enforces equalised odds as it pertained to a more stable model. The method of restricting the adversary training examples to exam cases when the *$y = 1$* is explored in the extension section near the end of the paper. Ultimately it was discovered mitigating unwanted bias through adversarial networks is easy to implement and effective without sacrificing the models ability to generalise to unseen data.

## 2. Datasets

The datasets used were the Bank Marketing [4] and the Statlog (German Credit) [5]. The bank contains 41,188 examples and 20 inputs but when converted and scaled into numerical values contains 63 attributes with a class label representing whether the customer did or did not subscribed to the bank term deposit. As shown in Figure 1, the Bank dataset has a large imbalance in both the sensitive attribute age and the class label.

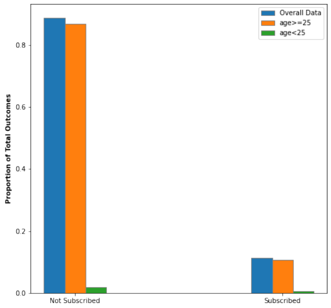

*Figure 1: Distribution of the class label in the Bank dataset.*

Given this imbalance within the bank dataset it's expected result would be a biased model in the absence of any deployed de-biasing techniques. With a large proportion of the data set class label being negative the model will have an issue with false negative predictions within both the sensitive groups.

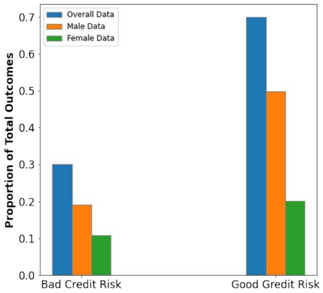

*Figure 2: Distribution of the class label of the German Credit dataset.*

The German credit dataset contains 1000 example and 11 attributes. Figure 2 shows the distribution of class label and sensitive attributes represented within. An imbalance in the direction of the positive class label with male having more representation within the good credit risk and less so in bad credit risk in comparison with the female representation. Again, without any de-biasing effort this will result in a model that is biased towards the male privileged group and prone to false positive from fitting to simply predicting positive outcomes. Both dataset were split into train and test data *(0.7, 0.3)* with 5 fold cross validation being performed on the train data.

## 3. Methodology

### 3.1 Performance Metric

The performance metric utilised was accuracy, that is the correct amount of predictions the model produced in regards to the ground truth label.

### 3.2 Fairness Metrics

The two metrics tracked in order to determine the fairness of the models are two commonly used measures and described by Zhang et al in regards to training an adversarial model.

1. Equality of opportunity: A classifier satisfies equal opportunity with respect to the class $Y$, if $\hat{Y}$ and $Z$ are conditionally independent on $Y = y$, for a particular value of the positive outcome of the true label $Y$, $P(\hat{Y} = \hat{y})$ is identical for all the values of the protected attributes. $P(\hat{Y} = \hat{y} \mid Y = y) = P(\hat{Y} = \hat{y} \mid Z = z, Y = y)$. In regards to the adversarial input, it essentially means that it will see both the classifiers output $\hat{y}$ as well as the ground truth label $y$. Although, it will only be examined on the cases of successful outcome. Hence, when $y = 1$.

In regards to the German credit data $Z$ is sex. The male being privileged (1) and female unprivileged (0). Within the Bank dataset it is age. Here, 25 and above is considered the privileged (1) within the class and under 25 unprivileged (0).

2. Equality of Odds: A classifier satisfies equalized odds with respect to the class $Y$, if if $\hat{Y}$ and $Z$ are conditionally independent with respect to all possible values of $Y$. Equalized odds is trying to achieve equality for both the outcomes of the class label $Y$ regardless of the sensitive attributes and is regarded as one of the highest forms of algorithmic fairness. Given the input to the adversarial network, here it sees the classifiers output $\hat{y}$ and the ground truth label $y$, but the difference here is that both the cases of of the class label are examined, hence $y = 1$ and $y = 0$.

### 3.3. Models

### 3.4. Classifier

For task one the model used for both datasets is the standard Logistical Regression model which is mathematically represented as:

$$d = x_1 w_1 + x_2 w_2 + .. x_i w_i + \beta$$

$$\check{y} = sigmoid(d)$$

where:

$$sigmoid = \frac{1}{1 + \mathrm{e}^{-d}}$$

With the class labels being of a binary nature the use of binary cross entropy was implemented and can be written as such:

$$J(\check{y}) = \frac{-1}{m} \sum_{i=1}^{m} Y_i \log(\check{y}) + (1 - y_i)(\log(1 - \check{y}_i)) + \frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2$$

Here, m is the number of training examples and $\lambda$ is the regularization hyper-parameter that determines the trade off between fitting the training data well and maintaining the parameters small enough to avoid over fitting.
From the plethora of choices for optimizing the loss function the Adaptive Movement Estimation algorithm *(ADAM)* was chosen to take advantage of the benefits of slowly decreasing the learning rate has upon reaching the loss's minimum accurately and quickly.

### 3.5. Classifier + Adversary

The adversarial network took the same form as the classifier, logistical regression. Here, the adversary is seeing the ground truth label, which in both datasets is discrete: 1 for the privileged and 0 for unprivileged. Also, receives the classifier's prediction from the sigmoid layer, which is the connection between the two models.

## 4. Task One Generalization parameter tuning.

Tuning the weight decay *(λ)* hyper-parameter of the Euclidean norm added to the update of the weights within the optimization process took the values show in Figure 3:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *λ:* | 0 | 0.001 | 0.01 | 0.1 | 0.5 | 1 | 10 | 100 | 1000 | 10,000 |

*Figure 3: Different values for the weight decay hyper-parameter.*

Covering the spectrum from no regularization *(0)*, to high level of regularization to analyse if generalising to the training data improves the models ability to accurately and fairly predict unseen data. The closer the equality of difference and equalised odds difference is to zero the fairer the model.

### 4.1. Results and analysis of tuning the hyper-parameter λ: Bank data.

Both datasets verified what was expected to be witnessed, containing significant bias with the absence of a de-biasing method. First, the bank dataset shows a significant bias in favour of the unprivileged groups. Figure 4 showcases how approaching the actual value of the  Euclidean norm the the model becomes even more bias in favour of the unprivileged group (*age < 25)*, in respect to the true positive rate difference between the protected groups. The model never substantially improves and steadily decreases in accuracy. Resulting in, values for λ above 0.5 producing poorer models.

| Metrics | λ: | 0 | 0.001 | 0.01 | 0.1 | 0.5 | 1 | 10 | 100 | 1000 | 10,000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Val Accuracy | | 90.9 | 90.89 | 90.84 | 90.45 | 89.63 | 88.82 | 81.86 | 77.48 | 66.58 | 58.77 |
| Eq odds diff | | 0.119 | 0.122 | 0.112 | 0.158 | 0.226 | 0.264 | 0.199 | 0.156 | 0.141 | 0.165 |
| Eq opp diff | | 0.094 | 0.095 | 0.093 | 0.117 | 0.181 | 0.233 | 0.222 | 0.174 | 0.111 | 0.135 |

*Figure 4: The effects the weight decay has upon the models on the given metrics, accuracy and fairness metrics round to 2 and 3 decimal places. The red indicating best fairness scores and green the highest validation accuracy.*

### 4.2. Final results on Test data, Bank:

| Model | Test Accuracy | Eq Opportunity | Equalised odds |
|---|---|---|---|
| HA: Biased | 91.1 | 0.137 | 0.101 |
| MF: Biased | 91.2 | 0.155 | 0.113 |

*Figure 5: Final results conducted on the Bank Test data where highest accuracy (HA) took the λ value 0 and most fair model (MF), took the value 0.01.*

Interestingly, *λ=0* was the highest average validation accuracy, but switched to being the most fair model once trained on the whole dataset and predictions conducted on the test set. This demonstrates the small scale of difference that a small L2 penalty can achieve. Additionally, the accuracy of the model has increased by a fraction and has generalized well to new data but at the cost of an increase in bias.

### 4.3. Results of tuning the hyper-parameter λ: German data

| Metric | λ: | 0 | 0.001 | 0.01 | 0.1 | 0.5 | 1 | 10 | 100 | 1000 | 10,000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Val Accuracy | | 68.33 | 68.33 | 68.32 | 68.41 | 67.18 | 66.54 | 63.88 | 63.2 | 61.6 | 60.77 |
| Eq odds diff | | -0.057 | -0.057 | -0.057 | -0.057 | -0.187 | -0.294 | -0.231 | -0.148 | -0.121 | -0.108 |
| Eq opp diff | | -0.144 | -0.144 | -0.144 | -0.144 | -0.275 | -0.363 | -0.311 | -0.171 | -0.089 | -0.112 |

*Figure 6: The effects the weight decay has upon the models on the given metrics. Accuracy and fairness metrics round to 2 and 3 decimal places. The red indicating best fairness and green the highest validation accuracy.*

### 4.4 Final results on Test Data, German:

| Model | Test Accuracy | Eq Opportunity | Equalised odds |
|---|---|---|---|
| HA: Biased | 69.67 | -0.157 | -0.275 |
| MF: Biased | 69.67 | -0.157 | -0.275 |

*Figure 7: Final results for task one conducted on the German test.*

Here, the λ was *0.1* for the highest accuracy and *0* for the fairest model. Both resulting in exactly the same metrics in the final predictions on the test data. Although, significantly higher than the average fairness metrics retrieved in the 5 fold cross validation.

## 5. Task Two, Adversarial De-biasing

### 5.1 Pre-training

The method for training the german data set was adopted from Beutel et al [1]. Both the adversary and the classifier were pre-trained to stabilise the losses for their individual tasks. During their respective pre-training, the model not being directly optimised had its  weights frozen. The adversary was trained after the classifier in order to have a stabilised input into the adversary. Both models for each data set used 3 epochs for pre-training.

### 5.2. Adversarial stage

The process of training the classifier and adversary involves a battle between the two models over a given number of epochs. Within the epoch the adversary uses the output  of the classifier to predict the sensitive attribute *Z.* A fair classifier would predict Y given *X* completely unrelated to the sensitive variable Z. In other words, having knowledge of *Y* would have no benefit in predicting *Z*, and should be completely up to chance.

The original mitigation through an adversarial network can be mathematically represented as:

Adversary Objective:

$$min_{\Theta adv}\left[Loss_z\left(\Theta_{clf},\Theta_{adv}\right)\right]$$

Classifiers Objective:

$$min_{\Theta clf}\left[Loss\left(\Theta_{clf}\right)-\alpha Loss\left(\Theta_{clf},\Theta_{adv}\right)\right]$$

The difference in our method is the update to the classifier will include a projection term shown in the gradient calculation equation below. The Classifier’s loss is defined as $L_p(y,\check{y})$ and the adversary loss is set as $L_A$ .

$$\nabla_W L_p - proj_{\nabla_W L_A}\nabla_W L_p - \alpha\nabla_W L_A$$

Zhang et al mention that it is possible for the gradient of the classifier to move in the direction that will help the adversary, the projection term resolves this issue by projecting the gradient onto the adversary's. The final term , $-\alpha \nabla_W L_A$ ensures that the predictor tries to hurt the adversary. This works because the sign is set to the opposite of the classifier's loss leading to the minimizing of the classifier loss and the maximizing of the adversary's loss obtaining the same sign and goal. To try prevent any information being leaked to the adversary, in both datasets the sensitive features were removed from the input to the classifier.

### 5.3 Adversarial loss weight

The hyper-parameter for controlling how much the classifier hurts the adversary is referred to as the adversarial loss weight. K-fold was performed on each value taken from 1 to 10. This was performed on both the datasets. The weight decay hyper-parameter for each model was derived from the analysis in task 1. The german model with *λ = 0.1* (L2 Norm) and *λ = 0* for the bank model.

### 5.4 Results and Analysis of Parameter Tuning, German model:

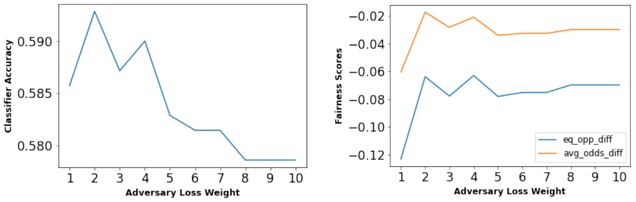

*Figure 8, on the left, is the impact of increasing adversarial loss weight upon the accuracy of the model and Figure 9 is the fairness metric scores. The y axis represents the averages calculated using 5 cross validation folds.*

The results show a clear indication of the de-biasing measure producing a more fairer model, but at the cost of a lower validation accuracy, this result is not surprising. The amount the model has improved in both the fairness metrics is substantial. Finally, the fairest model is when α = *4*, with an equality of opportunity difference score of *-0.00629*. The highest average validation accuracy was achieved when α = *2*, scoring *59.29*.

### 5.4.1 Final Results of Task 2, German Test Data

| Model | Test Accuracy | Validation Acc | Eq Opp Diff | Eq Odds Diff |
|---|---|---|---|---|
| HA: Biased | 69.67 | 68.33 | -0.1567 | -0.275 |
| MF: Biased | 69.67 | 68.41 | -0.1567 | -0.275 |
| HA: Debiased | 61.67 | 59.29 | -0.02 | -0.0592 |
| MF: Debiased | 62.66 | 59 | -0.0033 | -0.05 |

*Figure 9, final metrics of the debiased model on german test data.*

The debiased model chosen with the highest accuracy, performed near satisfying both fairness metrics. The issue is that the test accuracy fell by *8* and still contains bias towards the male class. As seen in Figure 9 the equality of opportunity difference dropped to -0.02, but the fairest model score was -0.0033.

### 5.5. Results and Analysis of Parameter Tuning, Bank model:

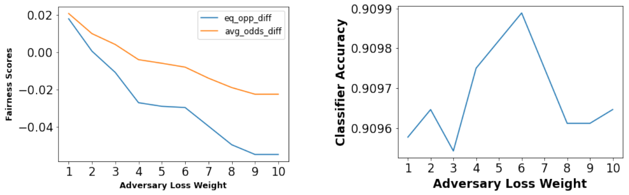

*Figure 10, the slow increase in bias towards the privileged class label in age (25 and above) as α is increased. The validation accuracy in Figure 11 showing an inconsistent trajectory.*

The performance across the folds with de-biasing enforced resulted in the true positive rate being in favour of the unprivileged group when α = 1. Contrary to the biased model in task one, it moved towards bias in the privileged group as α increases. Interestingly, the validation accuracy does increase and the highest average of the 5 folds achieved is *90.99*, when α = 6. For the fair model, α = 2 was chosen given it scored *0.0005* for the equality of opportunity difference and *0.0099 for equalised odds.* Bringing it close to satisfying both fairness measures.

### 5.5.1 Final Results of Task 2, Bank Test Data

| Model | Test Accuracy | Validation Acc | Eq Opp Diff | Eq Odds Diff |
|---|---|---|---|---|
| HA: Biased | 91.11 | 90.91 | 0.138 | 0.1008 |
| MF: Biased | 91.2 | 90.84 | 0.113 | 0.113 |
| HA: Debiased | 90.96 | 90.99 | 0.0645 | 0.0509 |
| MF: Debiased | 90.97 | 90.96 | 0.0853 | 0.0617 |

*Figure 11, final metrics of the debiased model on bank test data.*

Again, Figure 11 the highest accuracy model generalises better to the unseen test data than the selected most fair model. The method specifically had an improvement of over 50% on the biased model in regards to equality of opportunity even though the adversarial inputs and loss calculations was designed around equalised odds. Ultimately, both of the final models have mitigated to sum degree the bias in regards to their sensitive groups. Along with a minor decrease on the accuracy on unseen data.

## 6. Extension

### 6.1 Equality of Opportunity Adversarial training method.

The difference in the training for the adversary compared to the previous debiasing method is that the input to the adversary is restricted to cases where the ground truth label is a positive outcome. In other words, when $y = 1$. It still sees the classifiers prediction. This technique allows for the adversary to tackle the bias in regards to equality of opportunity of the sensitive groups directly. Similarly to task two, both the adversary and the classifier were pre-trained on 5 epochs. Further adaption of Bonks training technique was applied during the adversarial stage when the models are attempting to trick one another. Instead of even training, the adversary is given the advantage by being trained on all the values in the training data set. The loss for the adversary is only examined on the examples when the ground truth label is positive to satisfy equality of opportunity [1]. Within the same iteration a batch of 64 is drawn at random from the dataset and the ‘fight’ is undertaken on these examples. Another key difference is the absence of the projection term, this could potentially result in the classifier aiding the adversary at its task. The weight decay value was set to *0*, identical to task two model.

#### 6.2.1 Results

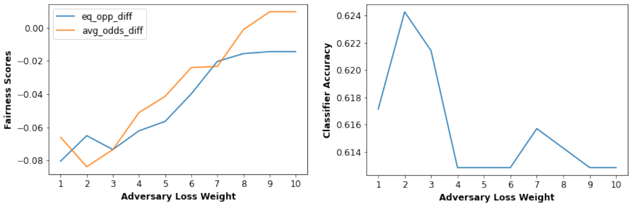

*Figure 12, left, shows the effects on the metrics over the increasing α on the fairness scores. Figure 13, right, the validation accuracy of the models.*

The model has a trade off, when α is low the accuracy is higher at the expense of weaker fairness scores. The validation score is similar when *α* takes values 4 to 6 and 9 to 10. From 4 to 6 the fairness metrics improving rapidly.

| Model | Test Accuracy | Validation Acc | Eq Opp Diff | Eq Odds Diff |
|---|---|---|---|---|
| HA: Biased | 69.67 | 68.33 | -0.1567 | -0.275 |
| MF: Biased | 69.67 | 68.41 | -0.1567 | -0.275 |
| HA: Debiased | 61.67 | 59.29 | -0.02 | -0.0592 |
| MF: Debiased | 62.66 | 59 | -0.0033 | -0.05 |
| HA: Debiased Ext | 61.67 | 62.43 | -0.077 | -0.254 |
| MF: Debiased Ext | 62.67 | 61.23 | -0.187 | -0.22 |

*Figure 13, final predictions on the test dataset for the extension task compared with the previous models.*

The proposed method did implement a level of algorithmic fairness and was successful at improving its objective of removing the bias in regards to the unprivileged female group. Decreasing the equality of opportunity difference from *-0.1567* to *-0.0077*. Unlike the equalised odds technique deployed in task 2, the final test accuracy decreased from the validation score. It can be concluded that this method is good for when the focus is on improving the true positive rates across the sensitive attributes.

### 7. Conclusion.

The initial analyses on the introduction of the L2 regularisation term into the loss of the logistical regression model in task one proved little in improvements in terms of accuracy and algorithmic fairness of the models trained. Both models over fitted to the training data even with regularization due to the class label imbalance ingrained within both datasets. The implementation of adversarial network to the training of the classifier moved the equality of opportunity difference and equalised odds difference in the direction of fairness without sacrificing accuracy and the models ability to generalise to unseen data.

### 8. References

[1] Beutel, A.; Chen, J; Zhao, Z; and Chi, E.H. 2017. *Data decisions and theoretical implications when adversarially fair representation.*

[2] Zhang, B.H.; Lemoine, B.; Mitchell, M. 2018. *Mitigating unwanted bias with adversarial learning.*

[3] Hardt, M.; Price, E.; Srebro, N.;et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315-3323.

[4] Moro, S.; Cortez, P.; Rita, P. 2014. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31

[5] Professor Dr. Hans Hofman Institute for Statistics, University of Hamburg, 2000