Student Number:243655

1. Introduction

It is often seen that ML techniques suffer from algorithmic biases which results in them making biased predictions favouring certain groups or individuals and putting others at a disadvantage.To avoid this problem, a set of tools are needed that can help in manataining a balance between accuracy and fairness.

In this report, the effect of regularization on accuracy and fairness is inspected using Logistic Regression as our classifier and the method of cross-validation is used for finding the best hyperparameter given the task at hand. An attempt is then made to come up with a score that accounts for both fairness and accuracy in the model based on which the final model will be selected. The problem of a non-binary sensitive attribute is dealt with using a GerryFair classifier, wherein an auditor would classify an individual as a part of a privileged or unprivileged group. Another approach is investigated for this problem which involves adding a constraint to the classification for minimizing the covariance between the concerned attribute and the classifier's prediction.

2. Datasets Used

A total of two datasets have been used for carrying out the analysis.

2.1. Adult Dataset

The Adult or the Census Income dataset has over 14 attributes about an individual along with a label attribute which indicates whether a person's income exceeds 50K or not.This dataset has 48842 data points.

2.2. Bank Dataset

The Bank dataset had data from a marketing campaign of a Portugese banking institution with target labels indicating whether a product was susbscribed or not.It had a total of 17 attributes with 45211 instances.

3. Models and Techniques Used

3.1. Logistic Regression

Logistic Regression is used as the model of choice for the task of classifying labels.It basically estimates $P(y = +1|x)$, the probability of a positive outcome, where $y \in \{+1, -1\}$.The genreral logistic function is used for estimating this probability such that $P(y = +1|w, x) = \frac{1}{1+e^{-(w_0+w^Tx)}}$ where w and w_0 are the parameters and x is the input matrix.The parameters are found by minimising the cost function $-\sum_{n=1}^{N} log(p(y_n|x_n, w))$.To prevent overfitting, we use add a regularization parameter λ to the cost function which then gives us the cost function $\sum_{n=1}^{N} -log(p(y_n|x_n, w)) + \frac{\lambda}{2}w^Tw$. This parameter is varied to check for its impact on accuracy and fairness of the model.

A major limitation of Logistic Regression is that it has a linear decision surface which means possible non linear relationships cannot be modelled using this technique.

3.2. Fairness Metrics

Equal Opportunity Difference was the metric of choice for assesing fairness. Equal Opportuniy Difference calculates the difference in true positive rates between the unprivilleged and the privilleged group. A value close to 0 for this metric indicates that both the groups are getting positive outcomes at the same rate, given that the people in it are qualified.

3.3. Reweighing

To enforce fairness upon the model, a technique called reweighing is used in which each instance or data point is assigned a weight according to its membership in a group.The weights are determined by the formula:

$$W(Y = 1, A = 1) = \frac{N(Y = 1)N(A = 1)}{N(Y = 1, A = 1)n}$$

, where $Y == 1$ represents a favourable outcome,n is the total number of instances and $A = 1$ represents membership to a privileged group, $N(x)$ denotes the number of instances of the condition x.This value is then multiplied by the cost function value for that data point.It is basically adjusts data point's weights such that the membereship in a group should be statistically independent to the classification outcome.

3.4. Overall Score Estimation(Fairness+ Accuracy)(Extra)

A scoring measure was devised which could rate models on the basis of both accuracy and fairness. To do this, a modified fairness measure,F_m of a model m is calculated such that $F_m = 1 - |eq_m|$, where eq_m is the equal opportunity difference score of the model m. Let A_m be the accuracy of the model. Then the Overall Score of the model O_m is calculated as

$$O_m = \frac{2}{\frac{1}{A_m} + \frac{1}{F_m}}$$

This is basically the harmonic mean of the modified fairness measure and accuracy.The model with the maximum value of O_m is chosen as the optimal model. Harmonic mean was chosen for this task as it penalizes lower values much more than the arithmetic mean and since this task would involve comparing models with a very small difference in accuracy and fairness, harmonic mean would be the best choice.

4. General Procedure

- In the results presented below, the C value represents the inverse of λ which is the regularization parameter in the Logistic Regression.

- The final scores calculated are the average of scores across all 5 folds.

- The model which obtained the best score on the cross validation set is retrained on the entire training data again before testing it on the testing set.

- In case that there is a tie in the scores, the model with the lowest C value is selected.

- ‘Sex‘ is the sensitive attribute for adult dataset where ‘Male‘ is the privilleges group while ‘Age‘ is the sensitive attribute for the bank dataset and the privilleged group is people with age greater than 25.

5. Results and Analysis for Adult Income Dataset

5.1. Results without Reweighing

Table 1: Results for Unweighted Adult Income Data

C	Average CV Accuracy	Equal Opportunity Difference
0.000001	0.795226	-0.218182
0.000010	0.796601	-0.222567
0.000100	0.796981	-0.252441
0.001000	0.803329	-0.442790
1.000000	0.803212	-0.448435
100.000000	0.803212	-0.448435
10000.000000	0.803212	-0.448435
100000.000000	0.803212	-0.448435

From the table we see that although the accuracies remain relatively constant, it decreases for very low values of C which is expected since low values of C imply more regularization effect.The fairness score clearly decreases with more regularization implying that increased regularization improves fairness.

The model with best accuracy(C=.001) and best fairness(C=1e-06) was chosen for the testing set.

On the testing set, the most accurate cross-validation model set obtains an accuracy of 0.805 and an equal opportunity score of -0.44 while the most fair model obtains an accuracy of 0.7964 and an equal opportunity score of -0.198.

5.2. Results with Reweighing

From the table below, it is seen that the fairness values are much lower and close to zero when compared to the unweighted case.In this case however, we see that the fairness scores are higher for very low values of C implying that increased regularization is actually increasing the bias of the model with respect to the unprivilleged group.The reason for this could be that increasing regularization is preventing the model to learn the additional complexities that have been introduced in the data by weighing the instances differently.The loss function has become more complex with reweighing and increasing regularization inhibits the model's abilities to learn these complexities which are related to enforcing fairness. Accuracies have decreased in general for similar values of C compared to the previous case(unweighted) which is in line with the expectation of a tradeoff between fairness and accuracy as mentioned in [3].

Table 2: Results for Reweighted Adult Income Data

C	Average CV Accuracy	Equal Opportunity Difference
0.000001	0.787329	0.010420
0.000010	0.787651	0.010816
0.000100	0.790108	0.001680
0.001000	0.790108	0.001680
1.000000	0.789932	0.003824
100.000000	0.789932	0.003824
10000.000000	0.789932	0.003824
100000.000000	0.789932	0.003824

The same model(C=.0001) had the best accuracy and the best fairness and this was chosen for testing on the test set.

This model obtained an accuracy of 0.7895 and equal opportunity difference score of 0.003.

5.3. Combined Scores with Reweighing(Extra)

Table 3: Results for Reweighted Adult Income Data

C	Average CV Accuracy	Equal Opportunity Difference	score
0.000001	0.787329	0.010420	0.876944
0.000010	0.787651	0.010816	0.876988
0.000100	0.790108	0.001680	0.882094
0.001000	0.790108	0.001680	0.882094
1.000000	0.789932	0.003824	0.881146
100.000000	0.789932	0.003824	0.881146
10000.000000	0.789932	0.003824	0.881146
100000.000000	0.789932	0.003824	0.881146

The combined scores results are presented in the table above. The model(C=0.0001) has performed the best and it achieves a test accuracy of 0.78959 with an equal opportunity difference score of 0.003.

5.4. Combined Score without Reweighing(Extra)

The results without reweighing are presented above and we see that the model with C=1e-6 has the best score in this

Table 4: Results for Unweighted Adult Income Data

C	Average CV Accuracy	Equal Opportunity Difference	score
0.000001	0.795226	-0.218182	0.788465
0.000010	0.796601	-0.222567	0.786900
0.000100	0.796981	-0.252441	0.771479
0.001000	0.803329	-0.442790	0.658008
1.000000	0.803212	-0.448435	0.654017
100.000000	0.803212	-0.448435	0.654017
10000.000000	0.803212	-0.448435	0.654017
100000.000000	0.803212	-0.448435	0.654017

case. Its testing accuracy is 0.7964 and the equal opportunity difference score is -0.1981.

5.5. Conclusions Regarding Model Selction with Combined Score(Extra)

It should be noted that the best model for unweighted case according to the combined score had a lower C value than the best model selected for the reweighted case. This tells us that higher regularization achieves a better accuracy and fairness tradeoff in the case when the model is trained on raw unweighted data compared to when the model is trained on reweighted data.

6. Results and Analysis for Bank Dataset

6.1. Results without Reweighing

Table 5: Results for Unweighted Bank Data

C	Average CV Accuracy	Equal Opportunity Difference
0.000001	0.802680	0.195242
0.000010	0.825688	0.228621
0.000100	0.883979	0.217437
0.001000	0.893820	0.128111
1.000000	0.897990	0.118641
100.000000	0.898084	0.121994
10000.000000	0.898084	0.121994
100000.000000	0.898084	0.121994

The results are roughly the same as that for the adult dataset with accuracy decreasing for lower C values implying regularization effect. The effect on fairness is very sporadic in this case.

The best model according to accuracy with C = 100 gets an accuracy of 0.9040 on the testing set and an equal opportunity difference score of 0.03519 on the test set while the most fair model(C=1) gets an accuracy of 0.904231 and an equal opportunity difference score of 0.01793.

6.2. Results with Reweighing

From the results given below, it is seen that that very low values of C and hence increased regularization has the effect of worsening fairness in the case of reweighted data as the overly regularized model fails to learn the more complex loss function created by reweighing the data.The accuracy behaves as expected and in the same way as it did in the unweighted case.

Table 6: Results for Reweighted Bank Data

C	Average CV Accuracy	Equal Opportunity Difference
0.000001	0.804414	0.168203
0.000010	0.829296	0.138253
0.000100	0.884401	0.077454
0.001000	0.893820	0.004166
1.000000	0.897990	-0.069797
100.000000	0.897990	-0.085912
10000.000000	0.897990	-0.085912
100000.000000	0.897990	-0.085912

The most accurate model(C=1.0) achieves an accuracy of 0.9041 and an equal opportunity difference score of -0.154 on the test set while the most fair model(C=.001) achieves an accuracy of 0.9008 an equal opportunity difference score of -0.0653 on the test set.

6.3. Combined Scores without Reweighing(Extra)

Table 7: Results for Unweighted Bank Data

C	Average CV Accuracy	Equal Opportunity Difference	score
0.000001	0.802680	0.195242	0.803718
0.000010	0.825688	0.228621	0.797610
0.000100	0.883979	0.217437	0.830185
0.001000	0.893820	0.128111	0.882718
1.000000	0.897990	0.118641	0.889597
100.000000	0.898084	0.121994	0.887932
10000.000000	0.898084	0.121994	0.887932
100000.000000	0.898084	0.121994	0.887932

The above results show combined scores of models trained on unweighted data. The best model(C=1.0) according to this score obtained an accuracy of 0.90423 and an equal opportunity difference score of 0.01793.

6.4. Combined Scores with Reweighing(Extra)

Table 8: Results for Weighted Bank Data

C	Average CV Accuracy	Equal Opportunity Difference	score
0.000001	0.804414	0.168203	0.817877
0.000010	0.829296	0.138253	0.845210
0.000100	0.884401	0.077454	0.903071
0.001000	0.893820	0.004166	0.942073
1.000000	0.897990	-0.069797	0.913813
100.000000	0.897990	-0.085912	0.905967
10000.000000	0.897990	-0.085912	0.905967
100000.000000	0.897990	-0.085912	0.905967

The above results show combined score for models trained on unweighted data. The best model(C=0.001) obtains an accuracy of 0.9008 and an equal opportunity difference score of -0.0653 on the test set.

6.5. Conclusions Regarding Model Selction with Combined Score(Extra)

Both the models selected according to the combined score in th eunweighted and reweighted case have similar C values implying that the effect of regularization on balancing the fairness-accuracy tradeoff is not clear. This could be due to the data having problems with respect to the process oflabelling sensitive groups.

7. Results on Droping Sensitive Features(Extra)

Given below are results obtained from dropping the sensitive features from the data on the Adult dataset.

Table 9: Results obtained after dropping sensitive attribute from adult data

C	Average CV Accuracy	Equal Opportunity Difference	score
0.000001	0.787680	0.011695	0.876661
0.000010	0.788792	0.005075	0.879948
0.000100	0.790108	0.001680	0.882094
0.001000	0.789991	-0.002316	0.881772
1.000000	0.789201	-0.017926	0.875136
100.000000	0.789201	-0.017926	0.875136
10000.000000	0.789201	-0.017926	0.875136
100000.000000	0.789201	-0.017926	0.875136

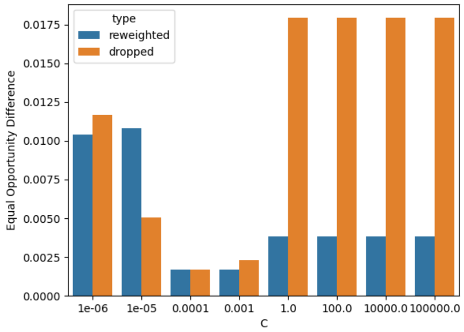

Figure 1: Comparision of the absolute value of Equal Opportunity Difference score of case where the data was reweighted according to the sensitive attribute and the case where the sensitive attribute was dropped.

It can be seen from the figure below that reweighing according to the sensitive attribute has clearly produced better results in terms of fairness when compared to just dropping the sensitive feature.The reason for this is that there would be some correlation in the other features with the sensitive attribute due to which just dropping the feature does not make as much of an impact as reweighing. Again, the negative effect of regularization on fairness is clearly visible as models trained with lower values of C perform similarly or worse than models trained on the dropped dataset.

8. Dealing with Non-Binary Sensitive Attribute(Extra)

8.1. GerryFair Classifier

8.1.1 General Description

A GerryFair Classifier has two kinds of classifiers.There is a learner which works on accurately classifying the data instance according to the labels and there is an auditor which works on identifying a group which has been unfairly treated compared to other individuals. The auditor is generally a linear thresholding function over protected attributes.This makes it applicable to the non-binary sensitive attribute case as the input features to the auditor function can be either binary or non-binary.The whole process is run over several iterations where the learner tries to minimize an objective function which includes both classification error and unfairness on the groups/subgroups found by an auditor from all the previous iterations, while the auditor tries to find group/subgroup on which the uniform distribution of all fairness classifier functions from the previous iterations violates fairness constraint the most. More details regarding the algorithm is given in appendix A.1.

There are mainly two parameters for this classifier:γ, which represents the upper bound for fairness score and C, which is related to the strength attributed to the fairness contraint in the cost function of the learner. γ is the parameter which is varied and experimented with while C is set to 10 for consistency.Larger values of this parameter was not found to have much impact on the results according to [5].The learner as well as the auditor is taken as a linear classifier in this case.

8.1.2 Fairness Metric

Given a classifier D, distribution P and a function $g(x)$ for classifying groups we have,

$$\alpha_{FP}(g, P) = Pr[g(x) = 1, y = 0],$$

$$\beta_{FP}(g, D, P) = |FP(D) - FP(D, g)|$$

where $FP(D) = Pr_{D,P}[D(X) = 1|y = 0]$ and $FP(D, g) = Pr_{D,P}[D(x) = 1|y = 0, g(x) = 1]$ In this case $FP(D)$ is the false positive rate of the classifier on the whole data while $FP(D, g)$ is the false poitive rate over the sensitive group.The fairness measure is then $\alpha_{FP}(g, P).\beta_{FP}(g, D, P)$.

The fairness contraint is given as $\alpha_{FP}(g, P).\beta_{FP}(g, D, P) < \gamma$, where γ is a hyperparameter. The learner's objective function will include this constraint.

8.1.3 Results and Analysis

The Bank dataset from AIF360 library is taken for analysis where age is taken as the non-binary sensitive attribute. The score is calculated in the same way was as it was in 3.4.

The table below gives results for implementing the method on the training set.The classifiers are trained for a maximum of 150 iterations for each gamma value. It is clear from the table that there is no improvement in error or fairness on increasing the gamma value above 0.001.Therefore, this value of gamma is considered for all further testing.

Table 10: Error and Fairness Disparity for various gamma values

	gamma	errors	fp_violations	score
0	0.0005	0.291588	0.000477	0.829158
1	0.0010	0.106087	0.000802	0.943628
2	0.0020	0.106087	0.000802	0.943628
3	0.0050	0.106087	0.000802	0.943628
4	0.0100	0.106087	0.000802	0.943628
5	0.0200	0.106087	0.000802	0.943628
6	0.0500	0.106087	0.000802	0.943628
7	0.1000	0.106087	0.000802	0.943628
8	0.1500	0.106087	0.000802	0.943628
9	0.2000	0.106087	0.000802	0.943628
10	0.2500	0.106087	0.000802	0.943628

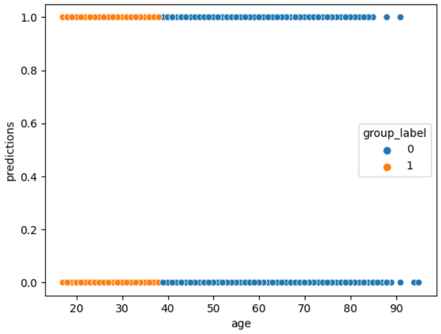

Figure 2: The privilleged(1) and unprivilleged groups(0) classified by the auditor after initial training for 2 iterations

The figure below shows results for classification of groups by the auditor after training for 2 iterations. Interestingly, the auditor has labelled the younger age group as privileged which contradicts the default labelling done by the AIF360 library on this dataset. Similar results were found in [2], which showed that using an age threshold of above 25 for the bank dataset results in the unprivilleged group being more likely to recieve a favourable outcome.

To test the efficacy of this method, a simple Logistic Regression classifier is fitted on the training set and the auditor function is run over it's soft predictions. This classifier gives a disparity score of 0.0004 with similar accuracy to the learner's classifier. On examining the training process, it is found that the GerryFair clasifier converges quickly(5 iterations) with the gamma value of 0.001. An attempt is made to retrain the classifier with lower gamma values in which case convergence in not achieved even with 1400 iterations. Hence, the conclusion is made that this method cannot achieve satisfactory compromises between accuracy and fairness and further testing is dropped.

8.2. Cost Function with Covariance Constraint(Extra)

8.2.1 Correlation as a Fairness Metric

A simple way of measuring fairness disparity is computing the correlation between the sensitive feature and the labels predicted by the classifier.Moreover, a constraint can be added to the cost function such that,

$$\frac{1}{N}|\sum_{i=1}^{N}(z_i - \bar{z})d_\theta(x_i)| < c$$

where z_i is the i^{th} value of the sensitive attribute, $\bar{z}$ is the mean of the attribute and $d_\theta(x_i) = \theta^T x_i$, where θ is the weight attributed to each feature.This formulation ensures that the cost function is convex.

8.2.2 Result and Analysis

The analysis is performed on the Bank dataset and age is used as the non-binary sensitive attribute. Logistic Regression is the classifier that is used for prediction and regularization parameter value is fixed as 1($\lambda = 1$). Cross validation process is used for correlation and accuracy score for various covariance thresholds.From this process, the model with the covariance threshold of 0.01 is chosen to be tested on the testing set as it had the best score.

Table 11: Cross Validation Results for different covariance thresholds

	Accuracy	correlation	score
Covariance Threshold			
0.0100	0.892352	0.031204	0.929004
0.2575	0.893523	0.044762	0.923351
0.5050	0.893757	0.049705	0.921159
0.7525	0.894226	0.060606	0.916253
1.0000	0.894811	0.070686	0.911736

The table above shows results for 3 types of classifiers.The 'Not Dropped' classifier is the one which is trained

Table 12: Test set Results for different classifiers

	Not Dropped	Dropped	Constrained Classifier
Fairness Score	0.063535	0.059033	0.002162
Accuracy	0.902700	0.903138	0.903356

using all the features in the data. The 'Dropped' classifier is trained on data that excludes the sensitive attribute while the 'Constrained Classifier' is the classifier which is trained using the constraint specified above.

The method suffers from the drawback that correlation only quantifies linear relationship between the sensitive attribute and the predictions and hence this can be an incomplete fairness measure.

9. Conclusion

- The use of Logistic Regression as a classifier limits our modelling capabilities as it has a linear decision boundary. To address this problem, neural network models could be investigated.

- Although the GerryFair classifier fails at the task at hand, the auditor method proves to be useful in identifying the correct privileged group from the non binary sensitive attributes.The constrained classifier is successful in solving this problem.

- The sporadic results obtained on performing our analysis for model selection on the bank dataset could be a result of incorrectly chosen thresholds by the library for labelling sensitive groups as demosnstrated by contradictory results obtained from the auditor.Use of intervals(for example label age group 30-50 as unprivileged) instead of thresholds could be explored for identification of sensitive groups.

References

[1] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

[2] M. Hort and F. Sarro. Privileged and unprivileged groups: An empirical study on the impact of the age attribute on fairness. Association for Computing Machinery (ACM), 2022. 5

[3] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012. 2

[4] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

[5] M. Kearns, S. Neel, A. Roth, and Z. S. Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 100–109, 2019. 4

[6] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2015.

A. GerryFair Classifier

A.1. Prediction Method for Auditor and Learner

Given c_0 as the cost of predicting the label 0 and c_1 is the cost for predicting the label 1, both auditor and learner use the following process for prediction:

- Train two linear regression models r_0 and r_1, to predict c_0 and c_1 respectively.

- For a point x, use the trained models to predict the cost of classifying x as 0 or 1.

- Output the label which has lower predicted cost.

A.2. Scoring method

Let Error of the classifier be err and disparity score be $viol$, then The final score is, given by the formula,

$$\frac{2}{\frac{1}{1-err} + \frac{1}{1-|viol|}}$$

B. Constrained Classifier

B.1. Scoring Method

Let accuracy of the classifier be acc and correlation be $corr$, then the final score is given by the formula,

$$\frac{2}{\frac{1}{acc} + \frac{1}{1-|corr|}}$$

C. Graph of Accuracy against Absolute Value of Fairness Score

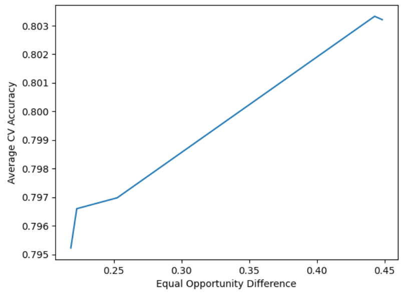

Figure 3: Unweighted Adult Dataset

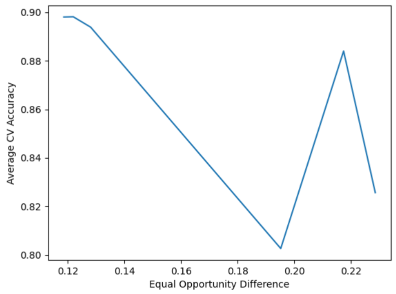

Figure 4: Unweighted Bank Dataset

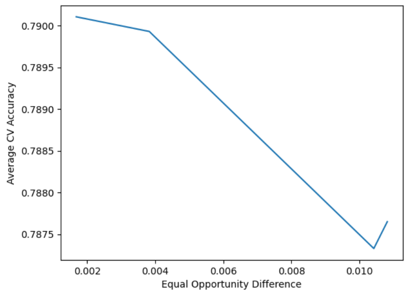

Figure 5: Reweighted Adult Dataset

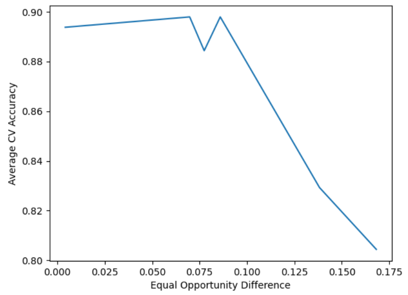

Figure 6: Reweighted Bank Dataset

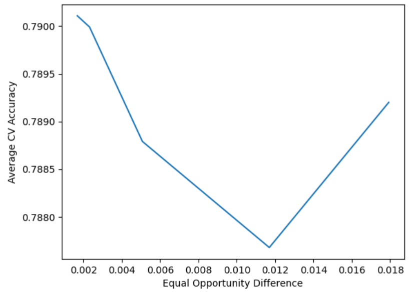

Figure 7: Reweighted Adult Dataset with sensitive features dropped

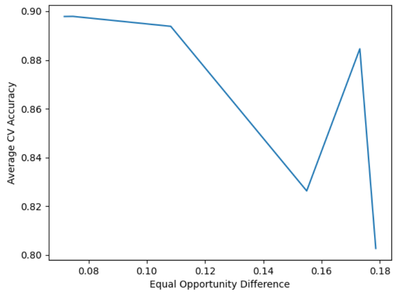

Figure 8: Reweighted Bank Dataset with sensitive features dropped