Student Number:251041

1. Introduction

This report provides an information about algorithm fairness. There has been a question since the rise of AI and Machine Learning that there is a trade-off between fairness and accuracy [1]. Also, it is yet to decipher that accuracy should be less while increasing fairness. This report break-down into various tasks to achieve the aim of the report. It begins with inclusion of machine learning methods to study the accuracy and fairness for Adult and German dataset. Moreover, it provides a model selection mechanism by giving empirical evaluation.

2. Dataset

In this report, I have used two different datasets from aif360 [2]. I have done my empirical study on Adult and German dataset from aif360.

2.1 Adult dataset

Adult income dataset has almost more than 48,000 attributes comprises of: age, sex, education, salary level etc. The protected feature for adult dataset is sex and race. These attributes are binary ones.

2.2 German dataset

The German credit dataset has more than 10,000 samples. The main feature for German dataset is sex and age.

3 Machine learning models

In this report there are two types of machine learning models used. The two types of classification methods are Logistic regression from Sklearn and Logistic regression with PyTorch.

3.1 Logistic regression model

It is one of the powerful regression tools used in machine learning. This method is more suitable for classification and binary problems [3]. It is mostly used as a linear classification model rather than a regression one. The most used regression model has binary results like true and false. It has its own logistic function to model for a binary output variable [3]. It is different from the linear regression by the range bound in between 0 and 1. There are no linear dependencies between linear variables in logistic regression. Logistic function is defined below.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Loss function in logistic regression is binary, it is just opposite to the information gain. Sometimes it is known to as log loss. The loss function of logistic regression is:

$$Loss = -\sum_{j=1}^{Output\ size} x_j * \log x_j$$

Activation function is a mathematical function that changes the inside variable to find a new value. This process is done with the help of machine learning techniques like sigmoid function or ReLU. I have used logistic regression with PyTorch. PyTorch is a deep learning tool that works with Torch [4]. It is indeed a more comfortable way of building the model for classification tasks than the normal machine learning way. In PyTorch, first we will create a neural network with few parameters that will reevaluate after every iteration. So, each time it will pass through the given input dataset. Then the input will cross the network by forward propagation [4]. Here I will find the loss and will try to minimize it using the gradient descent.

3.2 Evaluation Metric for Accuracy and Fairness

The report aim is to find the best accuracy and fairness metric by using the given dataset. The metric for accuracy is used with the help of Sklearn standard accuracy score parameter. For fairness, the aif360 equal opportunity difference is chosen for the required task. Equality of opportunity odds provides the predictor to be supportive of sensitive attribute through the outcome permitting the best classifier with the highest accuracy score and better demographic parity [5].

3.3 Regularization method for logistic regression

It is the best way to negate the effect of overfitting when training the logistic regression model. It is used by taking a penalty and decreasing the value of loss function in training.

$$lambda\ ||w||^2 - \sum_{j=1}^{n} \log\left(1 + \exp\left(-x_j\, w * y_j\right)\right)^{-1}$$

Here the lambda is a positive parameter. The value of lambda effect on regularization. If the value is larger the effect will be more. In this report we test lambda for accuracy and fairness by changing its values to random one.

3.4 Fairness metric

Let B = B (Y, C) ∈ ⟨0, |1⟩ be a predictor. Here B fulfils the condition of equalized odd correspond to the protected feature C and target value Y.

$$P[B = 1 \mid C = 0, X = 1] = P[B = 1 | C = 1, X = 1] = TPR$$
$$P[B = 0 \mid C = 0, X = 0] = P[B = 0 | C = 1, X = 0] = TNR$$

Here the TPR is True positive rate and TNR is True negative rate. The condition of equalized off gets fulfillied when the logistic regression classifier provides the same value to each of them [5].

Demographic Parity

Let B = B (Y, C) ∈ ⟨0, |1⟩ be a predictor. Here B fulfills the condition of demographic parity for the sensitive attribute C.

$$P[B = 1 \mid C = 0] = P[B = 1 | C = 1] = \frac{TP}{TP+FP+TN+FN}$$

Predictive Parity

Let B = B (Y, C) ∈ ⟨0, |1⟩ be a predictor. Here B fulfills the condition of predictive parity corresponding to the protected feature C and target value Y.

$$P[B = 1 \mid C = 0, X = 1] = P[B = 1 | C = 1, X = 1]$$
$$P[B = 0 \mid C = 0, X = 0] = P[B = 0 | C = 1, X = 0]$$

3.4 Fairness methodology

The fairness methods aim to stop the bias in machine learning models. The methodology of algorithmic fairness is divided into the three steps [6]:

In preprocessing we try to mitigate the biasness before the classifier is introduced. In processing it is believed that the discrimination should be eliminate while training the classifier. Post processing is generally done after the model get trained and fairness metrics have been used on it. In task 2, we have reweighed the training data. For this task We have used the aif360 reweighing baseline. In task 3, for model selection, we have combined both the accuracy and fairness metrics.

4 Results

The important task is based on the two datasets. Hence, there are two sets of results obtained from the two different models in separate two different parts. First let's discuss the result obtained from the Adult dataset and later for the German dataset.

4.1 Results on Adult dataset

The table 1 and 2 are the results of the adult dataset using logistic regression while using the PyTorch. The dataset has been split into 70 and 30 percent for training and testing purposes. From table 1 and 2, we can observe how the changing the value lambda is affecting the both accuracy and fairness metric. When we start the process with lambda equal to zero then there was no regularization effect on the train dataset. As soon as we start varying the value from the 1 to 10 the fairness metric(eq_opp_diff) gets better with decreasing accuracy. In addition, we can see from figure 1 how regularization is affecting the accuracy. Also, the other metrics like equal odds and predictive parity with TPR and TNR are not giving a promising hint to more fairness. The best model of task one got the best accuracy around 80%. And the best fairness was around -0.25.

lambda	accuracy	eq_opp_diff
1e-05	0.805022	-0.435016
1.0	0.802429	-0.363517
7.0	0.802361	-0.387759
10.0	0.798403	-0.247216

Table.1 Task one result on adult data set with PyTorch

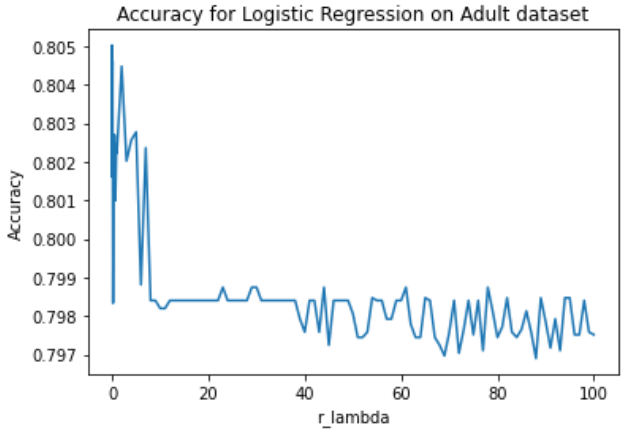


Figure.1. Accuracy vs r_lambda for adult dataset

On task two, we did reweigh and from table 2 we can see that the accuracy for the best model is around 79% and it has kept on decreasing with increasing the value of lambda. Meanwhile, the fairness metric has seen a positive trend and reweighing has definitely improved the fairness. Moreover, the equal odd and predictive parity condition has satisfied task 2.

lambda	accuracy	eq_opp_diff
1e-05	0.7905545	-0.198435
1.0	0.7905548	-0.198435
7.0	0.7866484	-0.198435
10.0	0.7839350	-0.198435

Table.2 Task two result on adult data set with PyTorch

On the other hand, with Sklearn, the model got accuracy around 80% with no biasness. And, the metrics have not changed much on reweighing.

C value	accuracy	eq_opp_diff
100	0.804089	0.00
2.33	0.795168	-0.441414

Table.3 Task one result on adult data set with Sklearn

C value	accuracy	eq_opp_diff
100	0.804089	0.00
2.33	0.795168	-0.441414

Table.4 Task two result on adult data set with Sklearn

4.2 Results on German dataset

The German dataset has been divided into train and test set with a ratio of 70 and 30. In comparison to Adult the German dataset has less attributes so has the less training data set. Which rises the problem of outlier so we have to take care of each data point for model selection.

lambda	accuracy	eq_opp_diff
1e-05	0.516666	-0.216957
1.0	0.503333	-0.108267
7.0	0.45	0.2492969
10.0	0.28	-0.3605174

Table.5 Task one result on German data set

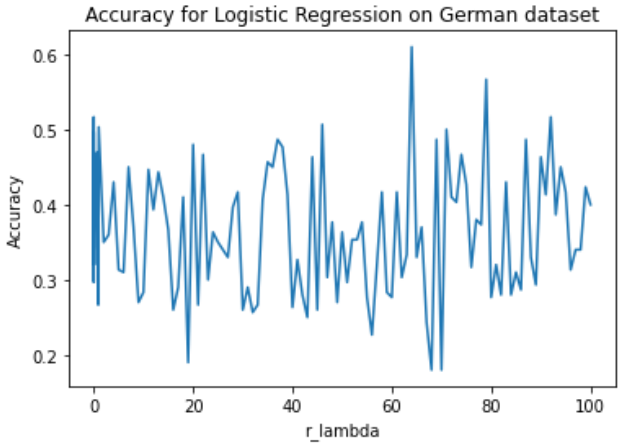


Figure.2. Accuracy vs r_lambda plot for German dataset

From table 5 and figure 2 we can clearly see that the accuracy is keep fluctuating with change in lambda. The best model has got an accuracy of 66%. The fairness metric has also got the same trend in task one. The close to zero value was around -0.10 for the best model.

On task 2, we reweigh the system for better fairness. Unlike the Adult dataset, the German dataset has performed very well on task two. The task has better accuracy than task one. Moreover, we discovered the zero difference in equal opportunity metric. Which means, there are no more biasness in the system. Both the equal odd and predictive parity metrics have also satisfied task two.

lambda	accuracy	eq_opp_diff
1e-05	0.443333	0.0
1.0	0.323333	0.0
7.0	0.526666	0.0
10.0	0.413333	0.0

Table.6 Task two result on German data set

5 Model Selection

Final task that is task 3, we have to pick the best model which justify for both accuracy and fairness. This model should be a combination of task one and task two. Earlier, we have already witnessed how the reweighing data and by changing the parameter affect the accuracy and fairness metric. So, now we want that the classifier has to produce the best accuracy also to balance the value of TPR and TNR. It meant the probability of expecting competent are those who are really a true competent one and other way round. For lambda equal to 24 we checked the accuracy which is almost around 76% but the TPR and TNR are around 1%. This suggests this model has almost the better accuracy with better fairness metric. The equal opportunity difference was close to zero for this model. Hence, this model seems to be the most accurate and fairer to the adult data set.

When we did the same procedure for German data the accuracy was around 43% with almost the same TPR and TNR values. This concludes that this model has better accuracy score and also stands better for fairness metric.

Data	Accuracy	(TPR/TNR)
adult	0.76	70/70.8
German	0.43	41/40

Table.6 Task three on German and adult dataset

5 Some other extrapolations

In this report we have to explore more on other task, such as finding the algorithmic fairness beyond the binary sensitive feature like including race or age with another binary feature. Also, perform an analysis with or without the main feature. Moreover, we have to compare the empirical evidence of both tasks.

5.1 Analysis of algorithmic fairness method beyond binary sensitive features

For further analysis we have added one more feature to our training dataset. This time, we have added one more feature race, which is also a categorical feature. When we have performed the data analysis with an adult dataset. It was found that white male is dominating the race feature hence we pick the white men as a privileged group and non-white like Black, Asian, Latino are the non-privileged ones. Henceforth, the classification process for this case is a replica of the previous one. The only major deviation is that the fairness measurement should be the same for both the race and age group. We have varied the hyperparameter lambda from zero to 0.1 and include the learning rate of $1*10^{-4.}$

The figure 3 illustrates accuracy for the adult data was escalating from 79% to 78% for a given hyperparameter.

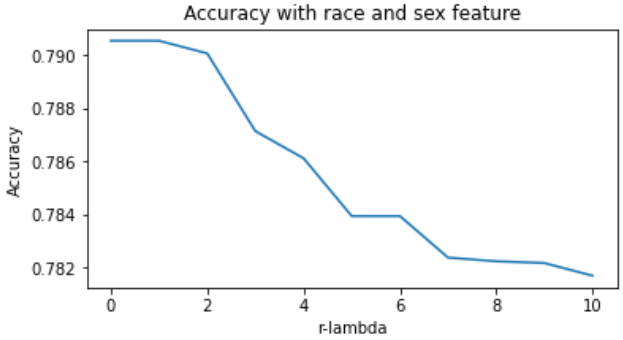


Figure.3 Accuracy and r_lambda graph for race and sex feature

5.2 Excluding the sensitive feature from the main input

We have investigated another situation where we have excluded the gender feature from the main input feature. Firstly, we transformed the adult data set into a data frame and dropped the required sensitive feature from it. It is too challenging and cumbersome to drop directly from the aif360 data. For further analysis, I have followed the same pipeline. It was observed from the empirical data both accuracy and fairness has increased from excluding the sensitive data. All the models have improved their accuracy and fairness. One can assume that it is better to drop the sensitive feature for better fairness but on the contrary, we have less data samples on our train data set, which will give us less information [7]. Hence it is not worth dropping sensitive attributes from the dataset. It plays a pivotal role for deciding the behavior of a machine learning model.

5.3 Data Analysis on Adult dataset

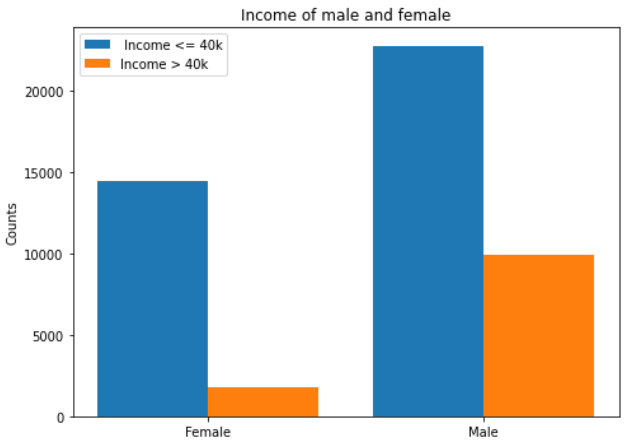
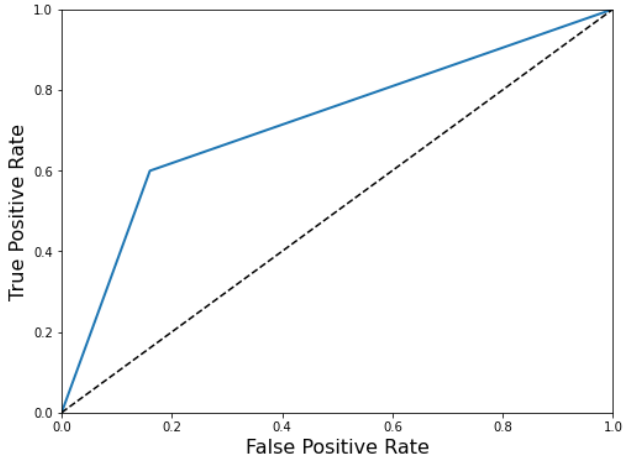


Figure.4 Bar plot of each class from adult dataset

The above figure depicts the major difference between the income between male and female in a range of more or less than 40k. The change in figure 4 is influence by various factor like education, job, sex etc.

ROC Curve

When we are dealing with a classification problem in machine learning, it is important to visualize the receiver operating characteristic curve (ROC). The ROC curve is plotted with respect to TPR and FPR.



6 Conclusion

Nowadays a lot of decisions are commanded by AI and machine learning. In the world of artificial intelligence and machine learning, it is crucial that the algorithms should be fair. We wanted algorithms to be fair and work better than mankind. Effect of regularization can somehow control the biasness. If a dataset is more biased and favoring a particular class to cope in this situation, we can apply a machine learning model to balance the accuracy and fairness metric. In this report, we have used both the adult and German dataset for measuring the accuracy and fairness metric. We have seen that in most cases that both accuracy and fairness have different trends. We have applied methods such as reweighing the system for getting more fairness or excluding the sensitive feature from main input to gain more accuracy and fairness for a few models.

For further work, there are many methods yet to be decoded for measuring the fairness metric. Fairness can be measured using reject option classification (ROC).

References

[1] Dutta, Sanghamitra, et al. "Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing." International Conference on Machine Learning. PMLR, 2020.1
[2]aif360readthedocs. https://aif360.readthedocs.io/en/latest/modules/datasets.html.1
[3] Logistic regression. https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on.1
[4] Logistic regression with pytroch https://towardsdatascience.com/logistic-regression-with-pytorch-3c8bbea594be.2
[5] Moritz Hardt, Eric price. Equality of opportunity in supervised learning. https://arxiv.org/abs/1610.02413.2
[6] N.Mehrabi, F.Morstatter. A survery on bias and fairness in machine learning. 2019. https://arxiv.org/abs/1908.09635 2
[7] Tianxiang Zhao, Enyan Dai, Kai Shu, Suhang Wang. Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features. https://arxiv.org/pdf/2104.14537.4

7 Appendix