Student Number: 244851

1. Introduction

The goal of including fairness metric in machine learning model is an attempt to correct the algorithmic bias in automated decision processes where the model's decision impacts people's lives with respect to the variables which are considered sensitive, such as gender, sexual orientation, age, disabilities etc.[4]. This report consists of the study of model selection that considers accuracy and fairness metrics together so that the machine learning model can be used for fruitful predictions without being too much biased against any of the sensitive attributes.

The standard Logistic Regression model is used to train aif360 adult data set [1] and German data set [2] separately with 5-fold cross-validation along with tuning the hyper-parameter 'C'. The 'C' values corresponding to the highest accuracy and highest fairness metric (True Positive Rate Difference) are taken and used to train and test the corresponding data set (Task 1). Then the experiment is repeated by also noting the effect of reweighing training data in the change in accuracy and fairness values (Task 2). An extra analysis is also done using the Random Forest Classifier which is mentioned at the end of this report.

1.1 Related Works

As Philip Ball [3] has mentioned there can be different reasons for the bias like the Target variable bias, Data collection bias etc. Also there are different types of measures like Demographic parity, Equality of opportunity, Equalized odds, predictive parity etc. which can be used to measure the fairness of the model and if the measures are not satisfactory, we can tweak the model by using different methods like including the reweighing technique or by simply creating fairness through unawareness.

Kilbertus *et al* [5] has demonstrated that it is practical for certain outcome-based notions of the group fairness on real-world data sets to learn fairer models and still maintain the cryptographic privacy for all user's sensitive attributes.

2. Model

Logistic Regression is used as the model in Task 1 and Task 2, Where the Task 1 is to analyse the model's performance in prediction accuracy and the fairness metric without reweighing (Standard Machine Learning model) and Task 2 is to analyse the model's performance in prediction accuracy and fairness with reweighing (Fairness based Machine Learning model).

2.1 Logistic Regression

Logistic regression is a classification model which uses the logistic function to squeeze the output from a linear equation between 0 and 1 [6].

$$logistic(\eta) = \frac{1}{1 + \exp(-\eta)}$$

Usually, the threshold is set as 0.5 and the logistic regression calculates the probability of the output and if it is above 0.5 then it is mapped to 1 and if it is less than 0.5 then it is mapped to 0, where 1 and 0 indicates the binary class to which we have to make the classification.

3. Implementation

The pre-processed data of both adult data set and German data set was imported form the aif360 data set library. Both Task 1 and Task 2 is performed on both the data sets. For both the tasks, each data set is divided into train and test data with a train to test ratio of 0.7.

For both the data sets the privileged groups are identified from the aif360 data set documentation. For the adult data set the privileged groups are male and for the German data set, the privileged group is with age > 25. Unprivileged groups for the data sets are Females and age <= 25 respectively.

The Logistic Regression model is imported from the sklearn python library. For Task 1, the hyper-parameter 'C' is varied to find out the best hyper-parameter for maximum accuracy and maximum fairness by doing the 5-fold cross validation on the train data. The corresponding 'C' values are used to train the complete train data and then it is used to find the accuracy and fairness the model can deliver by testing it on the train data.

For Task 2, the whole procedure is done again, but with implementing the reweighing (imported from aif360.alogorithms.processing.reweighing) on the train data. The instance weights of the reweighed train data is passed as the sample weights in the model and hence the model becomes fairness based machine learning model since the reweighing is done considering the corresponding privileged and unprivileged groups of the train data.

4. Analysis and Results

The hyper-parameter 'C' is the regularization parameter of the Logistic Regression model. C = $1/\lambda$. Lambda controls the model's ability to get as complex (overfit) it can. Small values of C means higher regularisation strength and hence it creates simple model which will under fit the data and higher value of C means making overfitting model.

One point to note here is that, I have used standard scalar for scaling the data, although MinMax Scaler would have given better results for these data sets. This is because, the standard scaler is more tolerant to the outliers in the data if there is any. The fairness metric used in the analysis is Equality of opportunity (The acceptance rates of the qualified applicants from each of the groups must be equal).

The analysis and result discussions for each task on each model are given below.

4.1 Task 1 (Adult Data)

A custom 5-fold cross validation function is created and 15 C values are selected from range 1 to $1e^{-10}$ and best C values for maximum accuracy and maximum fairness is selected from the cross validation.

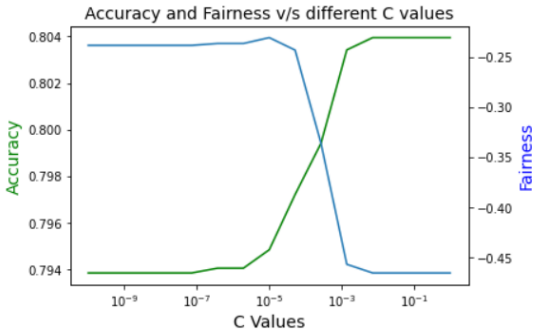

Figure 1. The accuracy and fairness values v/s different C values by Standard Logistic regression (Adult data set)

From the graph (Figure 1) it is evident that the as the C value increases the accuracy increases. On the contrary, the fairness metrics seems to get better as the C value decreases to $1e^{-5}$ and then very slightly get worse again.

The C values corresponding to the maximum fairness is $1e^{-5}$ and maximum accuracy is 1.0. One thing to keep in mind is that the adult data set has a comparatively large data size, hence the C values mentioned above held true while testing also.

4.2 Task 1 (German Data)

The same procedure as in adult data is repeated in the German data set and the following results are obtained.

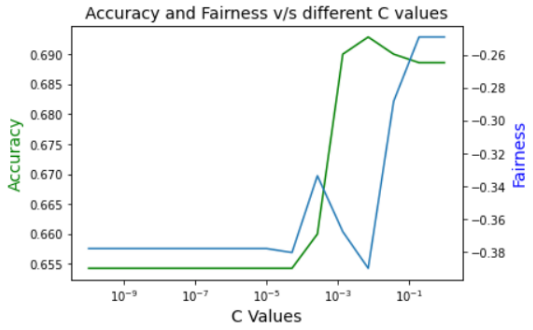

Figure 2. The accuracy and fairness values v/s different C values by Standard Logistic regression (German data set).

The graph above (Figure 2) is contradicting with the trend in fairness we observed in the adult data set. The C value corresponding to the maximum fairness while doing the cross validation is observed to be 1.0 and the C value corresponding to maximum accuracy is 0.007. Also, another discrepancy is that the while testing it on the test data set, both maximum accuracy and maximum fairness occurred for C = 1.0. These discrepancies are probably because the size of the total German data is very low (1000 rows). Hence on 0.7 split of train and test again the train data becomes almost only 700 rows and while doing the 5-fold cross validation, the data becomes sparse again.

So, the overall conclusion for the Task 1 is that better generalisation (lower C values) corresponds to fairer model and this might be possibly because better generalisation means strict regularisation and strict regularisation implies less emphasis on the weight of the training data. This will result in less emphasis of presence of large data supporting the privileged group and hence the model becomes fairer.

4.3 Task 2 (Adult Data)

In Task 2, we are studying the effect of reweighing the training data and passing the reweighed weights as the sample weight in the machine learning on the fairness and accuracy of the model. The same 5-fold cross validation used in Task 1 is used in Task 2 as well. Again, the C values corresponding to the maximum accuracy and maximum fairness will be found and will be passed to train the final model, which will be tested in the testing data.

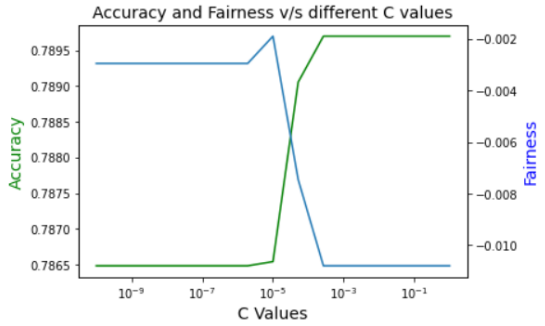

Figure 3. The accuracy and fairness values v/s different C values by Fairness based Logistic regression model (Adult data set)

From the graph above (Figure 3) it is evident that the fairness metric values are very close to zero (means the model is fair) compared to the fairness values we found on the same data set (adult data) in Task 1. This indicates that the model has become fairer in general for all the 'C' values and reweighing the data set is one of the effective strategies for making the model fairer. The trends in the variation of the accuracy and fairness remains the same as in Task 1 (adult data) with C value corresponding to maximum accuracy being 1 and C value corresponding to maximum fairness being $1e^{-5}$.

4.4 Task 2 (German Data)

The process for Task 2 (adult data) is repeated on the German Data and following observations are made.

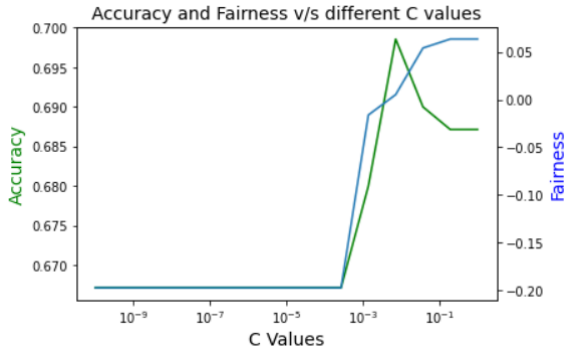

Figure 4. The accuracy and fairness values v/s different C values by Fairness based Logistic regression model (German Data)

Again, the general trend of the accuracy and fairness metric remains the same as in Task 1 (German Data), however the model has become fairer as the fairness metric values are now near zero, which is observable from the Figure 4. It is interesting to note that when the C value is near 1.0, the fairness metric value is slightly positive which means a very slight bias against the labelled privileged group. As observed in Task 1, the German data set is not behaving in the same way as the Adult data set. This might be again because of the lower data size of the German data compared to Adult data. The C value corresponding to maximum accuracy is 0.037 and C value corresponding to maximum fairness is 0.00719. However, the accuracy and fairness corresponding to both the C values remains the same.

The discrepancy we observed in the C value corresponding to the maximum accuracy in cross validation v/s testing holds here also. In the testing data, the C value corresponding to maximum accuracy remains 1.0.

So, the overall conclusion for Task 2 is that, the pre-processing technique of reweighing the data can make the model fairer.

4.5 Result Table

Data	Task	C value		Accuracy %	Fairness
Adult	1	Max Accuracy	1	80.42	-0.44
		Max Fairness	0.00005	79.75	-0.21
	2	Max Accuracy	1	79.05	0.035
		Max Fairness	0.0005	78.76	0.033
German	1	Max Accuracy	0.007	67	-0.38
		Max Fairness	1	71	-0.21
	2	Max Accuracy	0.037	68	0.082
		Max Fairness	0.00719	68	0.082

Table 1. The overall results for all both the tasks on both the dataset.

In the above table (Table 1) the accuracy and fairness metric represent the values we got by testing the corresponding models in the test data. In general, the accuracy of the model trained on German data is low compared to models trained on the Adult data. Also, it can be observed that the fairness in the Task 2 is better for both the data set since the reweighing is done on the training data.

5. Extra Content.

I have done extra analysis by using Random Forest model to perform both the Tasks (Task 1 and Task2) in both the data sets (Adult and German Data).

Random Forest Classifier is imported from the sklearn.ensemble library.

5.1 Random Forest Classifier/Regressor

The random forest algorithm consists of an ensemble of decision trees, each of which is made up of a data sample selected from a training set with replacement, known as the bootstrap sample [7]. One third of that training data is set as test data, known as out-of-bag sample. Feature bagging is then used to inject another instance of randomness into the dataset, increasing its diversity and decreasing the correlation between decision trees. The method for determining the prediction will differ depending on the type of problem. In a regression task, the individual decision trees will be averaged, and in a classification job, the predicted class will be determined by a majority vote—that is, the most common categorical variable. Finally, the out of bag sample is used for cross-validation, which seals the deal on the prediction [7].

Here we have to use the Random Forest model for classification purpose, so we use it as a Classifier.

5.2 Advantages and Disadvantages of RF

The main advantage of using random forest classifier is that the model will not overfit the training data as it uses multiple uncorrelated decision trees for making the decision. Since boot strapping and cross validation already happens in the random forest model, we do not have to do additional cross validation as we have done in logistic regression model. Another advantage of the random forest is that it can be used for both regression and classification problems.

However, there are some disadvantages also for the Random forest model. Since random forest has to create multiple uncorrelated decision trees, and compute each of them, the process is time consuming. It also requires more computing and dynamic data storage resources.

5.3 Results and Discussion

Data	Task	Accuracy %	Fairness
Adult	1	80.09	-0.46
	2	79.36	-0.04
German	1	70.33	-0.15
	2	69.33	-0.11

Table 2. The accuracy and Fairness of the Random Forest Classifier for both the Tasks on both the data set.

From the above table (Table 2), it can be observed that Tasks performed on German data gives low accuracy compared to Tasks performed on Adult data. Also, it can be observed that the reweighing training data makes the model fairer. It should be also noted that while performing Task 1, the adult data has more bias (unfair) compared to the German data set. This might be because of the presence of large pool of data belonging to the privileged group in adult data set compared to the low amount data belonging to the privileged group in the German data set.

6. Future Work

As we have observed, when we do the reweighing, the accuracy of the model decreases compared to standard machine learning output. It might be possible to come up with a strategy or model which might not affect the accuracy of the model but increases the fairness of the model (i.e., model which is accurate + fair)

Also, the performed analysis can be done using different fairness parameters other than Equality of opportunity and can be compared to each other.

Thee is also the possibility of exploring the counterfactual fairness where we make the algorithm-led decisions fair by making sure the outcomes are same in the actual world and a ‘Counterfactual world’ where an individual belongs to a different demographic.

7. Conclusion

The given Tasks (Task 1 and Task2) are performed using the Adult and German data sets by using the standard Logistic Regression model.

By performing Task 1 it can be conclude that better generalisation leads to fairer machine learning models. From completing Task 2 it can be learned that pre-processing Fairness technique like reweighing can decrease the bias of the machine learning model and hence help in creating fairer models. Also, during the analysis we could see the effect of training data size in hyper parameter selecting and in the fairness metric values.

Finally, as extra analysis, both the Tasks (Task 1 & Task2) is performed using Adult and German data set using Random Forest classifier. The advantages and disadvantages of the Random Forest Classifier compared to other machine learning model is also learned.

References

[1] aif360.readthedocs.io. (n.d.). *aif360.datasets.AdultDataset — aif360 0.4.0 documentation*. [online] Available at: https://aif360.readthedocs.io/en/stable/modules/generated/aif360.datasets.AdultDataset.html#aif360.datasets.AdultDataset [Accessed 1 May 2022].

[2] aif360.readthedocs.io. (n.d.). *aif360.datasets.GermanDataset — aif360 0.4.0 documentation*. [online] Available at: https://aif360.readthedocs.io/en/stable/modules/generated/aif360.datasets.GermanDataset.html#aif360.datasets.GermanDataset [Accessed 1 May 2022].

[3] Ball, P. (2018). *Fairness in Machine Learning with Causal Reasoning*. [online] Available at: https://www.mlmi.eng.cam.ac.uk/files/ball_thesis.pdf [Accessed 6 May 2022].

[4] Wikipedia. (2022). *Fairness (machine learning)*. [online] Available at: https://en.wikipedia.org/wiki/Fairness_(machine_learning) [Accessed 7 May 2022].

[5] Kilbertus, N., Gascón, A., Kusner, M., Veale, M., Gummadi, K. and Weller, A. (n.d.). *Blind Justice: Fairness with Encrypted Sensitive Attributes*. [online] Available at: https://www.fatml.org/media/documents/blind_justice_fairness_with_encypted_sensitive_attributes.pdf [Accessed 8 May 2022].

[6] Molnar, C. (n.d.). *4.2 Logistic Regression | Interpretable Machine Learning*. [online] *christophm.github.io*. Available at: https://christophm.github.io/interpretable-ml-book/logistic.html.

[7] www.ibm.com. (n.d.). *What is Random Forest?* [online] Available at: https://www.ibm.com/cloud/learn/random-forest.