Candidate Number: 229540

1. Introduction

When training machine-learning models, accuracy is often considered the most important metric in evaluating performance. In recent years, there has been a greater the reliance on algorithms to make human-centric decisions [14] making it increasingly more important to consider a second metric when evaluating models – Fairness. Fairness is hugely important for models making human-centric decisions as they can profoundly affect people's lives (e.g. job offers or bank loans) [17]. It is widley accepted in the literature that machines are not free from bias and often learn and preserve historical human bias [17] [13]. This stems either from bias which has infiltrated the training data, or bias in the algorithms themselves. An example of bias in the training data would be highly imbalanced data, leading to underrepresentation of minority groups. An example of algorithmic bias would be the use of 'proxy' attributes by the model, meaning the model makes decisions based on protected attributes under the cover of other legitimate attributes [17]. A protected attribute is a characteristic that is illegal to discriminate against, for example, age, race and sex [21]. There are several ways to measure group fairness in machine learning models [11] however this report focuses solely on equalized opportunity difference and this metric will be referred to as 'fairness' from here on out.

In this report, several cross-validates models will be trained to make human-centric decisions, each will vary with respect to the strength of regularization. I predict that models with higher regularization will perform with better accuracy compared to models with lower regularisation, which will perform worse on accuracy, but better on fairness. This finding is well documented and commonly rereferred to as the bias-variance trade off [4]. Models which achieve the best accuracy and fairness will be selected for evaluation on the test data. Then I will use apply a de-biasing technique known as reweighting on the data which in short works by changing the 'importance' of data points in the dataset. The same selection and evaluation steps will be performed on reweighted models and these compared to the previous models. Based on previous findings, I expect reweighting to significantly improve fairness scores [18]. I also expect the same bias-variance trade-off relationship to hold true. The process described will be carried out on two datasets.

2. About the Data

Both datasets used were obtained from the UCI machine-learning repository [9]. The first is the Adult income dataset; a multivariate dataset with over 48,000 instances and 14 attributes, including several protected attributes. The classification task is to predict weather an individual makes more or less than 50K per year. The second is the Statlog German credit dataset, also a multivariate dataset with 20 attributes, including protected ones. The classification task is to predict weather the individual has 'good' or 'bad' credit. This dataset is much smaller than the adult one with just 1,000 instances. The data used had already been pre-processed by the IBM AIF360 toolkit, for details on variable codes and more see [11]. Before modelling, data was split randomly split into train (70%) and test (30%) sets. I ensured the privileged and unprivileged groups were defined; for the Adult dataset the protected attribute used was sex, with privileged being Males and unprivileged Females. For the German dataset the protected attribute used was age, with privileged being > 25 and unprivileged group being <=25. Lastly, I normalized the data using the Scikit-learn standard scaler [16], this is important in all machine learning pipelines to convert data into one common scale, ensuring variables don't exert more importance on models due to intrinsically larger values [15]. Looking at Figure 1, it is evident both datasets carry bias, with both having a very small amount of data pertaining to the unprivileged group for the 'positive' outcome.

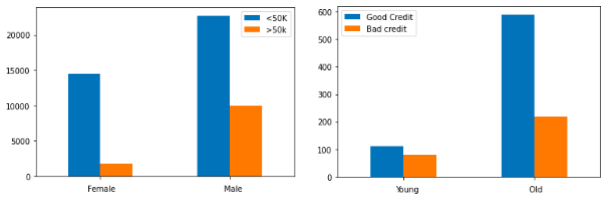

Figure 1. Bar graphs showing number of instances for each group and outcome for the Adult (left) and German (right) datasets

3. Model Type

Scikit-learn's Logistic regression models were fitted to the two datasets [16]. Logistic regression models are commonly used for binary classification tasks such as the ones in this report. The model learns a mapping from inputs to classes, using conditional probability. We obtain these probabilities using the logistic sigmoid function:

$$\sigma(x) = \frac{exp^x}{1 + exp^x} \qquad Eq.1$$

This function takes the linear transformation term $(W^T X + b)$ and transforms it to the [0,1] probability space:

$$\sigma(x) = \sigma(W^T X + w_0) \qquad Eq.2$$

The aim is to learn what function $f(x)$ maps our data X to the [0,1] output space best. This is achieved by learning which weights and bias terms create the best fitting function. The optimal weights are those which provide the maximum likelihood of getting our training labels y:

$$\arg\max_{w} \sum_{n=1}^{k} \log(\sigma(y_n W^T X_n)) \qquad Eq.3$$

The loss function being minimised in logistic regression is opposite to the equation above, max becomes min and log becomes -log. This is referred to as log loss and is minimised in Scikit-learn's liblinear logistic regression using co-ordinate decent. This iteratively updates just one weight at a time until there are no more updates to be made (aka. Convergence) [12].

For each dataset 8 models were trained both before and after reweighting, each varying with respect to the C hyperparameter, which controls the amount of regularization in a logistic regression. Regularisation constrains weights therefore keeping complexity of models lower to avoid overfitting to the training data [1]. This regularisation term is added to the end of the log loss equation:

$$\text{Minimise}_{w} \sum_{n=1}^{N} -\log(P(y_n \mid \boldsymbol{x}_n, w)) + \frac{\lambda}{2} w^T w \qquad Eq.4$$

The C hyperparameter relates to Equation 4 because C = 1/λ. Smaller lambda values mean that weights are allowed to grow larger and regularization is weaker. The same is therefore true of higher C values, as it is the inverse of lambda. I chose logistic regression over alternative models due to its simplicity. Simpler models should generalize better on small datasets [14], like the German dataset. For the adult dataset which is large, using a simpler model will facilitate faster computation of results.

4. 5-fold cross validation

5-fold cross validation was performed using Scikit-learn's cross validation tools [16]. Stratified K-fold was cross validation was used to ensure folds are representative of the whole dataset class distribution [8], important when datasets are imbalanced. 5-fold cross validation splits the training data into 5 random sections. Essentially, 5 models are fitted each using K – 1 folds, with the additional fold used for validation. The validation fold differs for each of the 5 models, thereby so does the combination of training folds. Accuracy is obtained by averaging accuracy scores across the 5 models. By varying which folds are used for training and validation, cross validation reduces overfitting to the training data, allowing for better generalization at test [20].

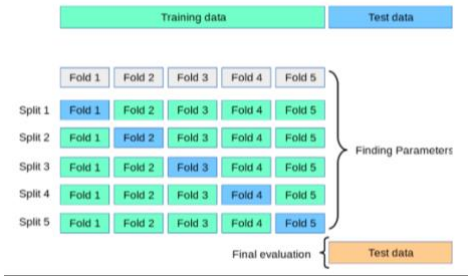

Figure 1. Schematic of 5-fold cross validation [16]

5. Model selection and Evaluation

The metrics used for model selection were accuracy and fairness. Fairness was calculated using Equal opportunity difference from the AIF360 toolbox [11], with a value of 0 indicating perfect fairness. Equality of opportunity is satisfied when all qualified individuals have an equal chance of receiving the 'positive' label regardless of their protected / sensitive attributes:

$$P[C=1 \mid A=0, Y=1] = P[C=1 \mid A=1, Y=1] \qquad Eq.5$$

Metrics were computed for the 8 adult dataset models, with the fairest and most accurate models selected using a selection algorithm. The most accurate model (Model 1a) had a C value of 0.01, accuracy of 80.4% and fairness of -0.449. The fairest model (Model 2a) had a C value of 1×10^{-8}, accuracy of 79.6% and fairness of -0.224. Full results can be found in Appendix 1. The same was done for the German dataset; The most accurate model (Model 1b) had a C value of 0.01, accuracy of 72.0% and fairness of -0.313. The fairest model (Model 2b) had a C value of 0.0001, accuracy of 66.0% and fairness of -0.239. Full results can be found in Appendix 2. The models selected at this step were then evaluated on the test data, see Table 1 for the results. Models 1b and 2b had much worse fairness scores at test, suggesting an inherent difference between training and test sets here.

Model number	Accuracy %	Fairness score
1a	80.3	-0.451
2a	79.6	-0.231
3a	79.1	-0.003
4a	78.9	0.015
5a	79.6	-0.231
6a	78.9	0.015
1b	67.0	-0.503
2b	63.3	-0.408
3b	68.0	-0.034
4b	68.0	-0.034
5b	67.0	-0.503
6b	68.0	-0.034

Table 1. Accuracy and Fairness scores when selected models for the Adult dataset (1-6a) and German dataset (1-6b) were evaluated on the test set, numbered as per the brief

6. Reweighting

Despite selecting for the fairest model, fairness for both datasets is still very poor, particularly for the German dataset. If these models were used to make human centric decisions, it would be considered discrimination. Fortunately, several methods are available for improving fairness in machine learning models. One such example is a pre-processing technique known as reweighing and is similar to up/down sampling your data. In the case of this report we have 2 outcomes x 2 possibilities for the protected characteristic, meaning we have four groups. In reweighting, each of these four groups is assigned a new weight and the value of this weight depends on whether the group should bare more or less importance on the model. For example, in the Adult dataset, instances of Males who earn more than 50K per year are likely to receive a weight <1 (down sampling) whereas Females who earn less than 50k per year will receive a weight >1 (up sampling). New weights are calculated as follows:

$$\frac{\#(Y=1)\,\#(A=0)}{\#(Y=1,\ A=0)N} \qquad \textit{Eq. 6}$$

The values of Y and A change depending on which group's weight is being calculated. This was implemented using the reweighting function from the AIF360 toolbox [11]. As before, 8 models were trained on the reweighted data, all with varying C values and were selected for most accurate and fair. For the Adult dataset, the most accurate model (Model 3a) had a C value of 0.0001, accuracy of 78.9% and fairness score of 0.005. Full results can be found in Appendix 3. The fairest model (Model 4a) had a C value of 100000, accuracy of 78.9% and fairness score of 0.001. For the German dataset, the models selected as most accurate (Model 3b) and fair (Model 4b) were the same. The Model had a C value of 100000, accuracy of 70.7% and fairness of 0.007. Again, selected models were then evaluated on the test data, refer back to Table 1 for the results of this. It is evident that reweighting has vastly improved fairness scores across all models. Interestingly, the bias-variance trade off that was evidenced before is gone and now the most accurate and fair models are more closely aligned. In the German dataset accuracy even improves after reweighting, showing that fairness interventions need not always be in competition with accuracy [2].

7. Final model selection

So far, accuracy and fairness have been considered as separate selection strategies. However, in a real-life scenario you would aim for a suitable balance between the two.

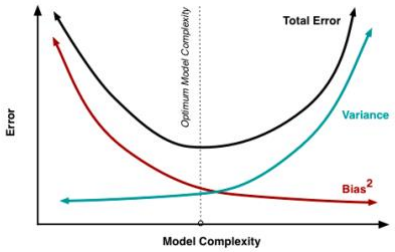

Figure 2. Optimal model complexity to balance accuracy and fairness [6]

Based on what I have observed, for the Adult dataset my selection strategy is as follows: The selected model will produce the fairness score closest to 0, so long as the accuracy does not drop by over 1%. An algorithm was created which would satisfy these conditions. The Model selected for the Adult dataset before reweighting (Model 5a) had a C value of 1 x 10^{-8}. After reweighting (Model 6a) the model selected had a C value of 100000. The models selected here were the same as the 'most fair' models from before. Accuracy and fairness metrics for these Models when evaluated on the test data can be found in Table 1. For the German dataset, I employed a similar strategy, assuming a slightly less stringent accuracy criterion of a drop no larger than 3%. This is because the dataset is smaller so small differences in misclassifications can change the overall accuracy by quite a bit. Running the selection algorithm before reweighting (Model 5b) chose the Model with a C value of 100000 which interestingly is the same as the 'most accurate' model from before. After reweighting (Model 6b), the same model was chosen as most fair and accurate has aligned in this case. To see the results of these models

evaluated on the test data see Table 1. I compared my selections to a very commonly used model selection method, Bayes information criterion. I found that for 3 out of 4 models this method chose models that were equal to the ones selected by my own strategy.

8. Fairness through unawareness

An alternative method of producing fair models is by removing protected attributes from model training, this is known as 'fairness through unawareness'[22]. This relies on the premise that models cannot be biased towards a characteristic that doesn't exist in the feature space. For the adult dataset, I removed the sex attribute, to ensure comparability to the findings above. When evaluated on the test set, the most accurate model achieved an accuracy of 79.1% and the fairest a fairness score of -0.003 meaning this technique has achieved comparable results to reweighting (Table 1). For the German dataset, the most accurate and fair models were the same with an accuracy of 68.0% and fairness of -0.034, again making this method equally as effective as reweighting. The reason that removing the protected attribute from the feature space does not lead to a fairness of 0 is because correlated features can 'leak' information into the model [5].

9. More than one attribute

Throughout this report, focus has remained on just one protected attribute per dataset despite both datasets having a second. For the adult dataset the second attribute is race and for German it is sex. To obtain more accurate fairness metrics, the same analysis was run taking both attributes into account. I expected to find that fairness scores would worsen with this change due to there being more conditions to satisfy. For the most part, this was true, expect for models 1b and 2b in which fairness was much worse with just one attribute. Despite fairness on the whole being worse with two attributes, low fairness scores were still achieved whilst presenting a more comprehensive picture of the models' potential for discrimination. The see the metrics for models when evaluated on the test data, refer to Appendix 5.

10. Discussion and concluding remarks

This paper empirically investigated the well documented bias-variance trade off [4] by training and comparing several models varying in their strength of regularisation. As hypothesised, I found that stronger regularisation led to more accurate models and weaker regularisation to less accurate but fairer models. Unexpectedly, this relationship did not uphold after the reweighing procedure. Again, as hypothesised, reweighting did lead to significant decreases in fairness scores uniformly. In addition, I found the method 'fairness through unawareness' had comparable effects on fairness scores to that of reweighting. This finding is of particular interest as it could serve as a useful tool for companies where storing sensitive data is in conflict with industry practices and legislation [23]. A second interesting finding was that the bias-variance relationship did not hold true after reweighting, suggesting something to do with the reweighting procedure has caused this behaviour. One last interesting finding was the volatility of fairness metrics on small datasets. For the German dataset, specifically in model 1 and 2b, a huge 15-20% difference in fairness was observed at test compared with train. This finding shows that caution must be taken when using very small datasets to create fair models. Interventions to improve model performance in these scenarios could include outlier removal or creation of synthetic samples [19].

The work of this report could be furthered by employing strict fairness through unawareness by removing protected attributes in addition to any which are highly correlated. This model would stand a good chance of achieving complete fairness but most likely at a cost as removal of too many features has a negative impact on model performance [3] Several papers have adopted strategies to bypass this issue, one found that by minimising correlations between related features and model predictions a fair and accurate classifier could be learnt [23]. Exploring this could be an interesting next step. Another direction for further study is to complete the same analysis with a more complex model, such as a neural network. Neural networks can support non-linear solutions [7] and are therefore more flexible, having the potential to improve performance. Having said this, it would only be appropriate for the adult dataset as neural networks require large training sets [7]. Lastly, another direction is exploring other fairness metrics. Whilst equality of opportunity is intuitive, it lacks the stringency of other metrics [10], meaning models will appear fairer than they are in reality. Equalised odds for example adheres a more stringent policy whereby both true positive rates and false positive rates must be equal [10], therefore this metric provides a more accurate picture of model fairness. Having said this, the metric used should be assessed on a case by case basis, depending on the extent to which outcomes affects people's lives. If impact is large, very stringent metrics should be applied but if it is low, less stringent metrics would suffice.

Words = 2,740

References:

[1] Adrià Luz, 2017, Why you should be plotting learning curves in your next machine learning project,

https://towardsdatascience.com/why-you-should-be-plotting-learning-curves-in-your-next-machine-learning-project-221bae60c53

[2] Blum, A. and Stangl, K., 2019. Recovering from biased data: Can fairness constraints improve accuracy?. *arXiv preprint arXiv:1912.01094*.

[3] Brian Pietracatella, 2020, Are you dropping too many correlated features?, https://towardsdatascience.com/are-you-dropping-too-many-correlated-features-d1c96654abe6

[4] Briscoe, E. and Feldman, J., 2011. Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*(1), pp.2-16.

[5] Castelnovo, A., Crupi, R., Greco, G. *et al.* A clarification of the nuances in the fairness metrics landscape. *Sci Rep* **12,** 4209 (2022). https://doi.org/10.1038/s41598-022-07939-1

[6] Cornell University, Lecture 12: Bias-Variance Tradeoff, https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html

[7] Danny Varghese, 2018, Comparative Study on Classic Machine learning Algorithms, https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222

[8] Diamantidis, N.A., Karlis, D. and Giakoumakis, E.A., 2000. Unsupervised stratification of cross-validation for accuracy estimation. Artificial Intelligence, 116(1-2), pp.1-16.

[9] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[10] Garg, P., Villasenor, J. and Foggo, V., 2020, December. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 3662-3666). IEEE.

[11] IBM AIF360, 2022, AI Fairness 360 documentation, https://aif360.readthedocs.io/en/latest/index.html

[12] Johnson, T.B. and Guestrin, C., 2017, July. StingyCD: Safely avoiding wasteful updates in coordinate descent. In *International Conference on Machine Learning* (pp. 1752-1760). PMLR.

[13] Kleinberg, J., Mullainathan, S. and Raghavan, M., 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

[14] Lepri, B., Oliver, N. and Pentland, A., 2021. Ethical machines: The human-centric use of artificial intelligence. *IScience*, *24*(3), p.102249.

[15] Mahbubul Alam, 2020, Data normalization in machine learning, https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02

[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[17] Pessach, D. and Shmueli, E., 2022. A Review on Fairness in Machine Learning. *ACM Computing Surveys (CSUR)*, *55*(3), pp.1-44.

[18] Radovanović, S., Petrović, A., Delibašić, B. and Suknović, M., 2019. Making hospital readmission classifier fair–What is the cost?. In Central European Conference on Information and Intelligent Systems (pp. 325-331). Faculty of Organization and Informatics Varazdin.

[19] Rafael Alencar, 2019, Dealing with very small datasets, https://www.kaggle.com/code/rafjaa/dealing-with-very-small-datasets/notebook

[20] Rukshan Pramoditha, 2021, How to Mitigate Overfitting with K-Fold Cross-Validation, https://towardsdatascience.com/how-to-mitigate-overfitting-with-k-fold-cross-validation-518947ed7428

[21] UK Government, 2022, Discrimination: your rights, https://www.gov.uk/discrimination-your-rights

[22] Verma, S. and Rubin, J., 2018, May. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)* (pp. 1-7). IEEE.

[23] Zhao, T., Dai, E., Shu, K. and Wang, S., 2022. Towards Fair Classifiers Without Sensitive Attributes: Exploring Biases in Related Features.

Appendix:

[1] *Accuracy and fairness metrics for models trained on the adult dataset*

C Value	Accuracy	Fairness
100000	80.4%	-0.452
1000	80.4%	-0.452
10	80.4%	-0.452
1.0	80.4%	-0.452
0.01	80.5%	-0.449
0.0001	79.8%	-0.240
$1 X 10^{-6}$	79.6%	-0.231
$1 X 10^{-8}$	79.6%	-0.224

[2] *Accuracy and fairness metrics for models trained on the German dataset*

C Value	Accuracy	Fairness
100000	71.4	-0.300
1000	71.4	-0.300
10	71.4	-0.300
1.0	71.4	-0.300
0.01	72.0	-0.313
0.0001	66.0	-0.239
$1 X 10^{-6}$	66.0	-0.239

1 X 10^{-8}	66.0	-0.239

[3] *Accuracy and fairness metrics for reweighted models trained on the adult dataset*

C Value	Accuracy	Fairness
100000	78.9%	0.001
1000	78.9%	0.001
10	78.9%	0.001
1.0	78.9%	0.001
0.01	78.9%	0.001
0.0001	78.9%	0.005
1 X 10^{-6}	78.6%	0.010
1 X 10^{-8}	78.6%	0.010

[4] *Accuracy and fairness metrics for reweighted models trained on the German dataset*

C Value	Accuracy	Fairness
100000	70.7	0.007
1000	70.7	0.007
10	70.7	0.007
1.0	70.7	0.007
0.01	69.4	0.046
0.0001	67.0	-0.111
1 X 10^{-6}	67.0	-0.111
1 X 10^{-8}	67.0	-0.111

[6] *Accuracy and fairness metrics for selected models when evaluated on the test set when two attributes were used for both the adult (1-4a) and German (1-4b) datasets*

Model number	Accuracy %	Fairness score
1a	80.3	-0.464
2a	79.6	-0.223
3a	79.6	-0.158
4a	79.3	-0.019
1b	67.0	-0.332
2b	67.0	-0.332
3b	68.7	-0.038
4b	68.7	-0.038