

Bias in AI: Implementation Project

fsdn71

Department of Computer Science
University of Durham
Durham, United Kingdom
fsdn71@durham.ac.uk

I. PROJECT PROPOSAL

A. Outline

The Adult Income dataset is a subset of the U.S. Census database, and is widely accepted as an accurate and predictive representation of the living circumstances of the U.S. population. It contains demographic, financial and employment-related information in both continuous and categorical forms, making it useful for training machine learning-based algorithms to predict various socio-economic features. However, it has been found to contain significant biases among race and gender attributes, persisting across multiple variables [1]. Therefore, this presents a suitable challenge for bias countermeasures to reduce bias shown in predictive models trained on the data. This project is proposed to analyse and implement state-of-the-art anti-bias methods on a classifier predicting whether an adult's annual income is over \$50,000, given data on employment class, capital flow and household attributes.

B. Motivation

The existence of gender and racial biases in the microdata is not unexpected - arising predominantly from historical imbalance in public policy. However, models which predict income or demographic variables can introduce or perpetuate bias, facilitating discrimination towards real-life individuals, when the outputs of the model are used to inform decisions on lending, hiring and other areas. Developing models which ensure fair outcomes for individuals is therefore a positive contribution to society, and one that this project aims to reproduce.

The Adult Income dataset is also appealing as an implementation project for fairness because of its wide coverage of individuals - the Census features every household in the U.S. Other datasets, such as those relating to online services such as Wikipedia or YouTube or educational outcome microdata, feature smaller subsets of the population, often with characteristics significantly different than the general population. The Census dataset is a suitable representation of the population, and includes weights for individual records, allowing a user to extrapolate directly onto the population. This means that algorithms applied onto the dataset will have some accurate insight into the biases present in the population as a whole, as well as how these could be altered with fairness regulation.

C. Adversarial Learning

The Adult Income dataset is well-established across the sub-fields of machine learning, with it being used to benchmark many different approaches to algorithmic fairness. In *Learning Adversarially Fair and Transferrable Representations*, Madras et al. outline a framework for explicitly discouraging discriminative behaviour within classification models [2] (a similar method is introduced in [3]). This is achieved by mandating that the inputs to a classifier be modified by a separate model (a neural network in the paper), such that an adversary cannot identify the sensitive attribute. This method has been selected as the fairness measure to implement due to its novelty, strong previous results and prior application on the dataset. Furthermore, the authors found that a representation-generation process that is accurate and unbiased on one task can be used as input to other predictive models using the same underlying data, improving fairness and retaining utility.

D. Planning

The implementation will involve the following steps:

- 1) Dataset pre-processing: decoding categorical features into one-hot representation, normalising continuous features.
- 2) Dataset analysis: producing descriptive statistics, broken down by sensitive attribute, to investigate the existing bias in the dataset, using Python libraries such as NumPy, Pandas and Plotly for data handling and visualisation.
- 3) Conventional implementation: using the TensorFlow [4] library, a neural network model comparable to the model used in [3] will be trained and tested on partitions of the dataset. Bias and utility will be reported.
- 4) Adversarial implementation: The model described in [2] will be implemented and applied to the conventional implementation training process. Changes in bias and utility will be reported, in comparison to the conventional algorithm.

The end result of the project will be a detailed analysis of existing bias in each of: U.S. survey microdata labels, outputs from an unconstrained neural network model, outputs from a model trained with adversarial fairness constraints. In addition, this project will extend the results by applying the adversarial approach onto a multi-variate sensitive attribute - the intersection of race and gender, as opposed to a single binary variable.

II. PROJECT PROGRESS

The implementation follows four major steps: analysis, conventional implementation and implementation of the adversarial model.

A. Analysis

The dataset contains a number of categorical features - these were transformed into one-hot encoded variables. The dataset does not contain null values, so no filtering was necessary. The sensitive attribute was selected as the intersection of race and gender - however, race was limited to White/Non-white values in order to keep the number of samples within each intersection similar. Additionally, when one-hot encoding categorical variables, categories with coverage of than 0.1% of the microdata were removed. The implementation notebook contains four tables, showing descriptive statistics for categorical and continuous variables for each demographic subgroup. All statistics and graphs are calculated using the census weights matched to the 1994 U.S. population.

The dataset shows significant bias in favour of both men and White individuals inclusively. Figure 1 shows the existing disparities in capital gains, loss and the high-income indicator, with the mean value for each subgroup as a percentage of the mean value for the population. It is clear that while bias exists along both racial and gender axes, gender-based bias is more severe. In addition, the distributions of continuous variables within demographic subgroups are different and may offer some explanation: Figure 2 shows the quantiles of weekly labour hours for each subgroup, indicating that the overall disparities are not caused by differences in the percentage working 40 hours (the median), but by the comparatively larger and fewer percentages among non-white or female individuals working low (20) hours and high (60) hours, respectively.

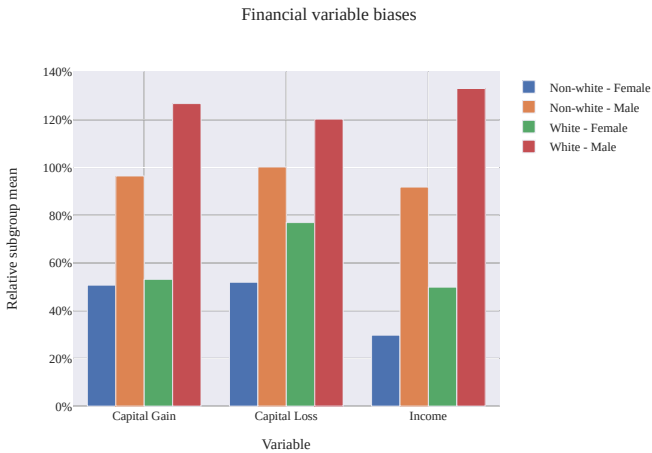


Fig. 1. Bias in financial variables

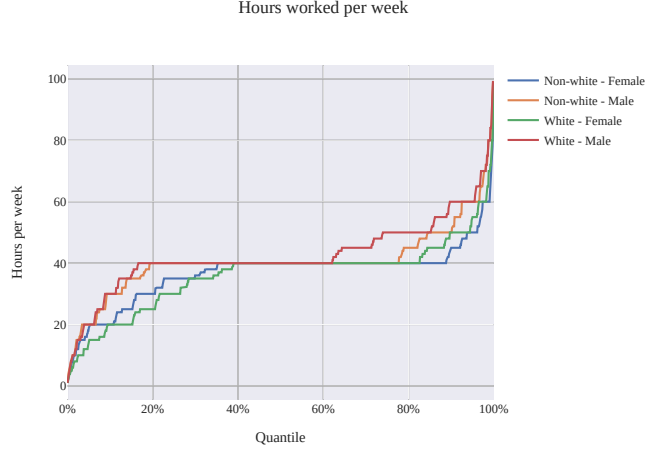


Fig. 2. Disparities in labour among subgroups

B. Conventional Implementation

After dataset pre-processing, we obtain three arrays of length 32,561. These are X , the features (79 features), S , the sensitive attributes (4 features) and Y , the labels (1 label). These are split into training, validation and testing partitions (with ratio 80 : 10 : 10). The modelling approach taken is to use a standard deep neural network, consisting of three fully-connected layers (of size 64, 32 and 16, respectively), each prepended with a dropout operation ($p = 0.02$), and a final output with sigmoid activation. The total number of parameters is 7,745.

The model successfully fit the underlying pattern of the data, achieving high classification accuracy on the testing set. Figure 3 shows the evolution of these performance metrics over the training session. Table I shows the test set accuracy metrics, including binary accuracy, precision, recall and F1 scores for the conventional model.

The model was also re-trained and re-tested on a demographically balanced version of the dataset. This was generated by calculating the percentages of the dataset records belonging to each of the four demographic groups, and using these to generate record weights ensuring equal proportions (i.e., generating a weight vector W such that for each demographic group, the sum of their individuals' weight values is one quarter of the original unweighted size of the dataset). An equally-sized dataset is then sampled from the original using the record weights as probabilities for individual rows. This approach was chosen over others, such as removing rows, in order to minimise unnecessary data loss. The model, when re-trained and re-tested, achieved similar classification accuracy metrics, shown in Table I.

However, as shown in Figure 4, the model replicates (albeit slightly less severely) existing bias within the dataset. Re-training the model on a sample with balanced demographic *presence* is likely to only improve the equalised odds metric,

rather than the actual outcomes of the model on the demographic subgroups, and therefore it is expected that in Figure 4 the conventional model trained on the balanced dataset does not improve disparate impact significantly.

Performance metrics by training epoch

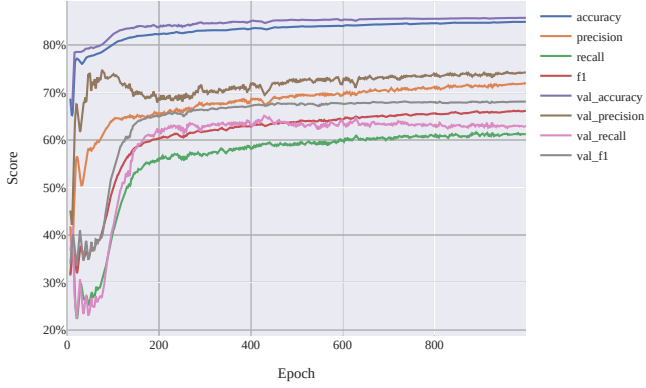


Fig. 3. Key performance metrics over training time

C. Fair Implementation

The adversarial fairness method outlined in [2] has a core method, and an adjustment for bounding demographic - this project implements the base fairness method. That is, a two-stage process: firstly, three models are implemented (an encoder, a decoder and an adversary). The encoder-decoder pair transform the input X to a representation Z and back, minimising reconstruction error and maximising adversary loss. The adversary predicts the sensitive attribute, minimising categorical cross-entropy. The classifier is excluded from this stage in order to maximise generalisability of the embeddings to different tasks. After this adversarial training, a classifier is trained to predict Y , the high-income indicator, from Z (we assume at this point that Z is both informative and unbiased, eliminating the need for further fairness measures). One parameter is relevant in the training process: the fairness-utility preference β , the relative value of the adversary loss against the encoder-decoder loss. This can be thought of as relative value of a decrease in how predictive the output is of the race-gender group, against an equal increase how predictive the output is of the true high-income value. For this experiment, we set $\beta = 2$ in order to test the wider application of the method outside its previous applications.

The models were trained for 1000 epochs each, and evaluated using the test set. Figure 4 shows the changes to key metrics: for each demographic group, the disparate impact as defined in Equation 1 (derived from the disparate impact formula proposed by Feldman et al. [5]), with the four-fifths rule highlighted.

$$DI(s) = \frac{p(Y = 1|S = s)}{p(Y = 1|S \neq s)} \quad (1)$$

TABLE I
ACCURACY RESULTS

	Accuracy	Precision	Recall	F1
Conv.	85.8	72.2	66.0	69.3
Conv. (Bal.)	85.8	71.9	66.8	69.2
Fair	78.9	69.8	45.9	60.8

The results show a large increase in fairness accompanied by a small decrease in accuracy, in line with prior findings. These outcomes are similar to those in [2], which reported a far stronger reduction in unfairness than other methods, though the results here obtain accuracies around 4-5% lower than in the paper. This is likely due to selecting a greater preference for fairness, as well as the approach to use race-gender subgroups as opposed to gender alone, increasing the power of the adversary and therefore reducing the utility of the fair embeddings in which race and gender must be more heavily disguised as a result (the utility of the embeddings is reduced as collateral damage). The most apparent effect on bias is the significant reduction in the advantage of White male records - this is to be expected: as the most populous demographic group, its divisor in Equation 1 has the lowest number of records, rendering it the most sensitive to changes in the mean for non-White or non-male demographic groups. Table I shows the breakdown of accuracy results for the fair model compared to the conventional models.

Changes to disparate impact

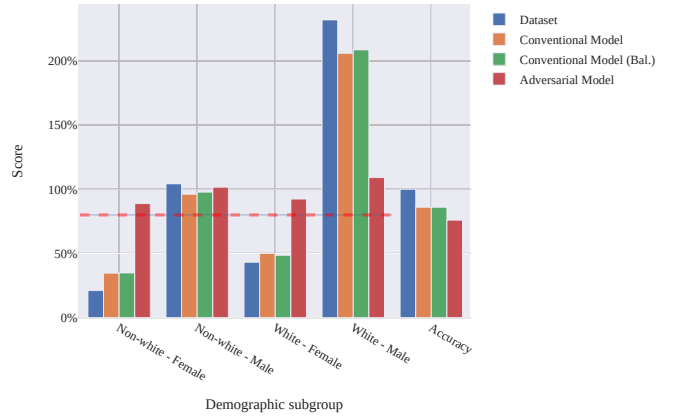


Fig. 4. Changes in metrics across models

The results of the adversarial model training underscore its effectiveness at reducing bias while maintaining model utility: as outlined in [5], the four-fifths rule is a widely-used measure of fairness in the U.S. with respect to disparate impact, and the adversarial method of generating ‘fair’ embeddings meets the criterion in all four of the demographic groups, in comparison to only one (the White-male intersection) in the dataset and conventional models, without a large decrease in accuracy, presenting a strong case for its wider use.

REFERENCES

- [1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *CoRR*, vol. abs/1908.09635, 2019. [Online]. Available: <http://arxiv.org/abs/1908.09635>
- [2] D. Madras, E. Creager, T. Pitassi, and R. S. Zemel, "Learning adversarially fair and transferable representations," *CoRR*, vol. abs/1802.06309, 2018. [Online]. Available: <http://arxiv.org/abs/1802.06309>
- [3] T. Adel, I. Valera, Z. Ghahramani, and A. Weller, "One-network adversarial fairness," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 2412–2420, Jul. 2019. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/4085>
- [4] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from [tensorflow.org](https://www.tensorflow.org/). [Online]. Available: <https://www.tensorflow.org/>
- [5] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," 2015.