A GROUP-LEVEL LEARNING APPROACH USING LOGISTIC REGRESSION FOR FAIRER DECISIONS

MARC ELLIOTT

BACKGROUND

Decision-making algorithms are permeating throughout society, where real-life data contains demographic imbalances

Advancements in algorithmic fairness are providing a layer of safety to the individuals being processed by AI systems, however, selecting an appropriate fairness notion can be challenging

Popular in-processing fair AI techniques utilise explicit fairness to achieve a desirable balance between fairness and performance

OUR OBJECTIVES

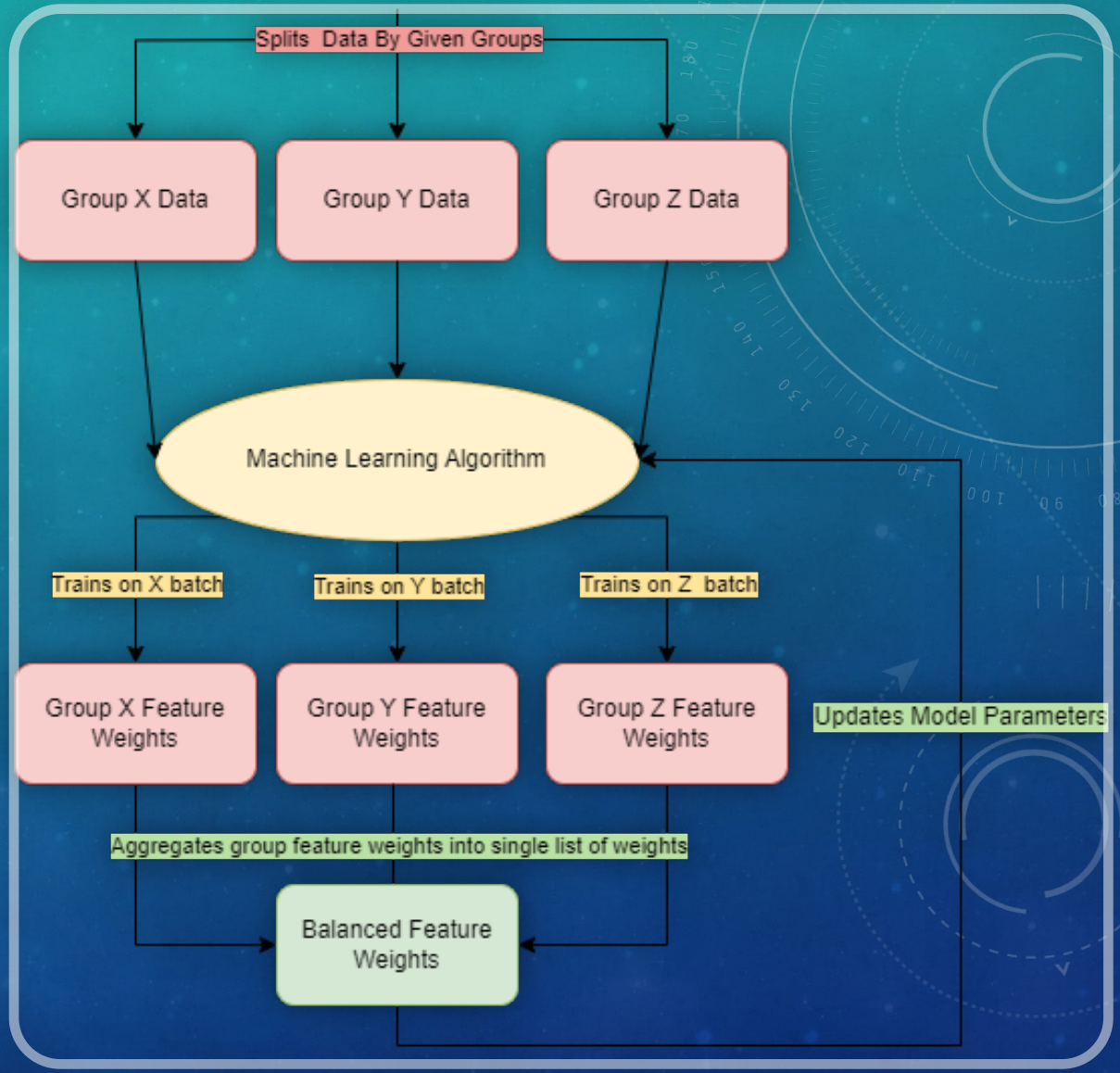
To provide minority demographic groups with greater influence in the algorithmic training process

To improve algorithmic fairness through an ‘implicit’ fairness approach

To maintain simplicity and convenience so it can be easily understood by a wider audience

GROUP-LEVEL LEARNING LOGISTIC REGRESSION

- Modifies the training process to consider sensitive groups independently
- Learns co-efficient values at a group-level, to represent each considered group
- At each training iteration uses a median aggregation of all groups derivative co-efficient values
- Median values used for updating the model parameters – a more equal contribution from sensitive groups regardless of distribution



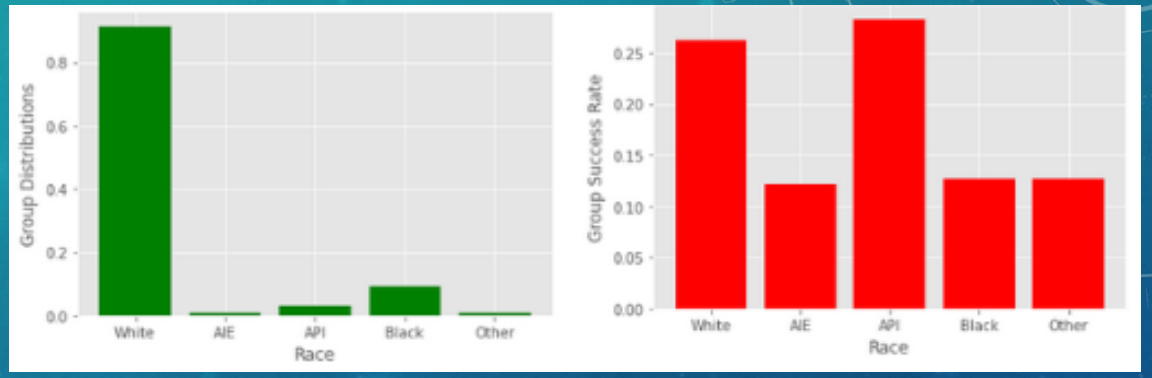
EXPERIMENT SETUP

- Two real-world datasets which contain imbalanced data: Adult & Open University
- Two tests for each dataset – a binary sensitive attribute and non-binary sensitive attribute
- Group-Level LR benchmarked against a baseline LR, and two state-of-the-art explicit Fair AI approaches
- Two commonly utilized fairness metrics measured: Demographic Parity (DP) and Equality of Opportunity (EO)

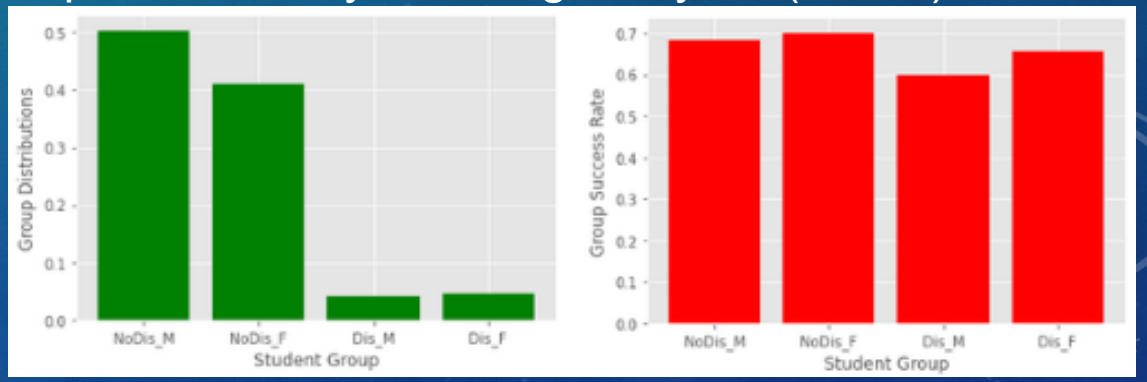
$$\mathrm{DP} \Rightarrow \mathrm{P}(\hat{y}= 1|\mathrm{S} = 0) \approx \mathrm{P}\ (\hat{y}= 1|\mathrm{S} = 1)$$

$$\mathrm{EO} \Rightarrow \mathrm{P}(\hat{y}= 1|\mathrm{S} = 0, \mathrm{Y} = 1) \approx \mathrm{P}\ (\hat{y}= 1|\mathrm{S} = 1, \mathrm{Y} = 1)$$

Adult Income Dataset

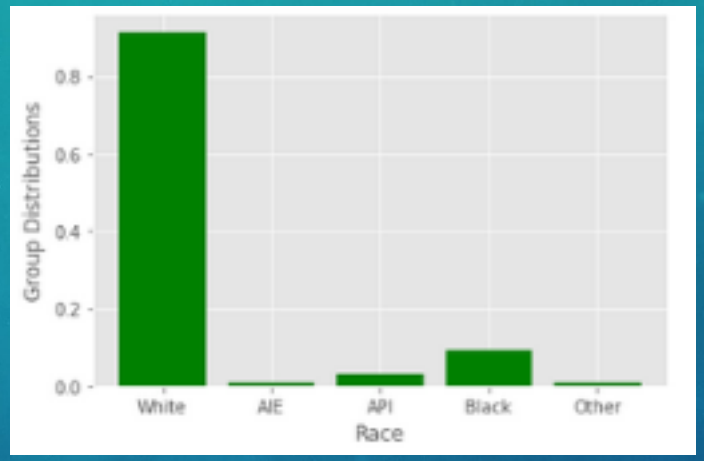


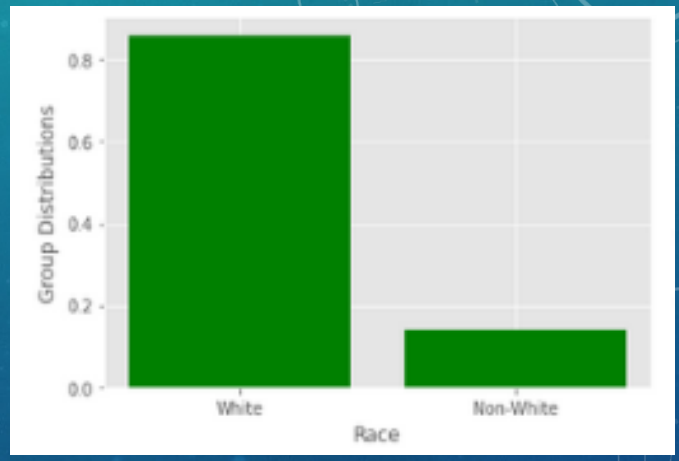
Open University Learning Analytics (OULA) Dataset



EXPERIMENT 1: ADULT INCOME DATASET

- Utilised the race attribute as the sensitive attribute
- Fairness assessed using two commonly used fairness metrics – demographic parity (DP) and equality of opportunity (EO)
- Two experiment scenarios: binary and non-binary sensitive attribute
- State-of-the-art explicit approaches achieve best in their constrained fairness metric
- Group-Level LR generally fairer than baseline LR at cost of accuracy

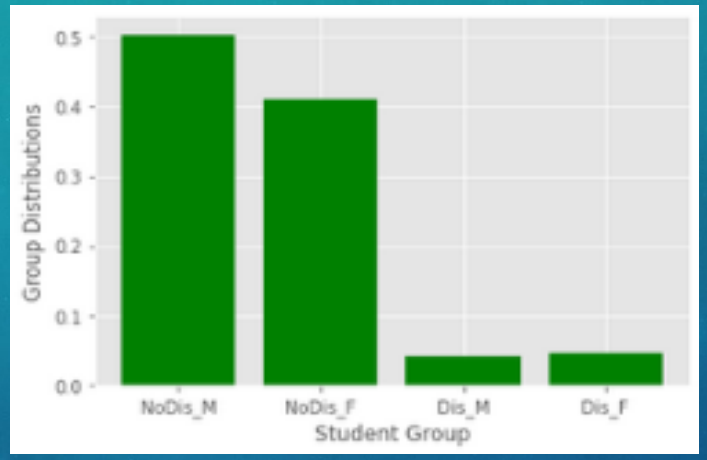


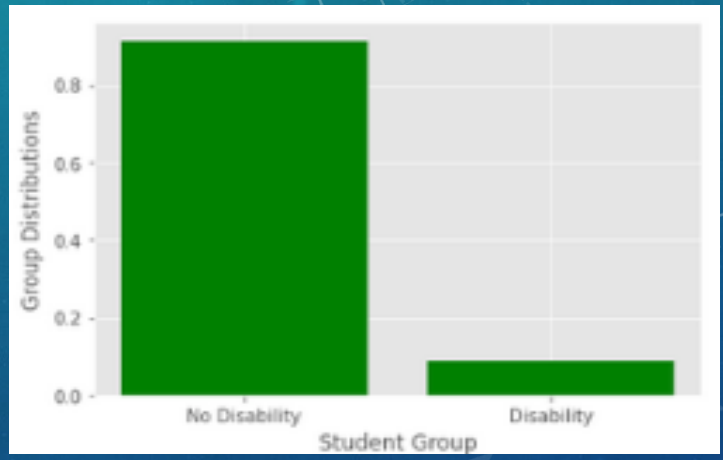


	Non-Binary Attribute			Binary Attribute		
Algorithm	ACC	DP	EO	ACC	DP	EO
Baseline LR	**0.8084**	0.1057	0.1687	**0.8084**	0.0312	0.0118
Fairlearn DP	0.7982	**0.0240**	0.1478	0.8011	**0.0059**	0.0502
Fairlearn EO	0.7563	0.0795	**0.0576**	0.8043	0.0255	**0.0028**
Group-Level LR	0.7535	0.0798	0.2370	0.7733	0.0230	0.0215

EXPERIMENT 2: OPEN UNIVERSITY LEARNING ANALYTICS

- Open university is a real-world dataset, demonstrating real group imbalances between student demographics
- Utilised disability attribute for binary test, disability-gender for non-binary
- Similar trend in results – loss of accuracy for greater fairness than baseline LR model
- Explicit state-of-the-art achieves best fairness performance, but Group-Level LR often achieves competitive fairness





	Non-Binary Attribute			Binary Attribute		
Algorithm	ACC	DP	EO	ACC	DP	EO
Baseline LR	**0.8441**	0.0933	0.0843	**0.8441**	0.0899	0.0708
Fairlearn DP	0.8284	**0.0458**	0.0488	0.8331	**0.0488**	0.0343
Fairlearn EO	0.8062	0.1036	**0.0381**	0.8294	0.0514	**0.0286**
Group-Level LR	0.7875	0.0881	0.0501	0.7807	0.0501	0.0290

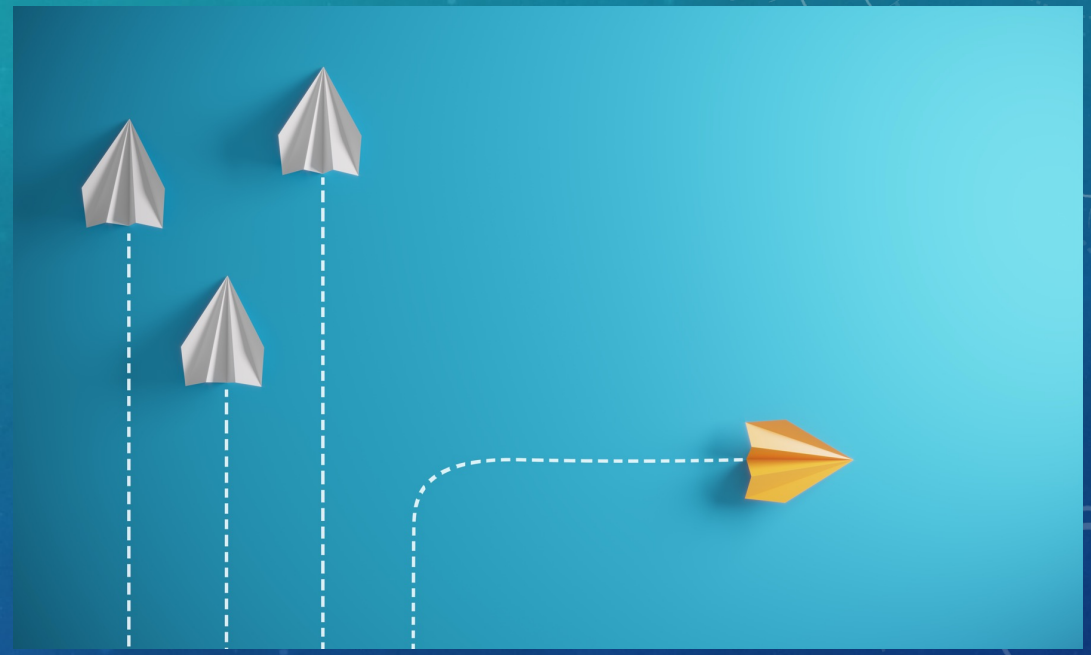
CONCLUSIONS

- The group-level algorithm may not yield the greatest fairness returns in one direction but points towards how **harnessing imbalanced data at the group-level can be used to improve fairness**.
- Our approach improved DP and EO against baseline LR, providing **minority demographic groups more control over model weight values resulted in more similar quality predictions for all groups**.
- Our method presents that **fairer predictions can be produced by implicit fairness approaches** by providing underrepresented groups with a more significant role in forming the algorithmic models which are being deployed and integrated into society.

FUTURE DIRECTIONS

The current approach signifies a small but crucial step towards novel implicit fair AI techniques, to further develop this area we aim to:

1. Adapt the group-level approach to other gradient-based models and be a more generalisable approach to fairer classifications
2. Integrate other group aggregation methods which are more semantically aligned with fairness may open new directions
3. Investigating methods of mitigating unfairness arising through data samples with outlier or extreme coefficient values



THANKS FOR LISTENING

QUESTIONS?