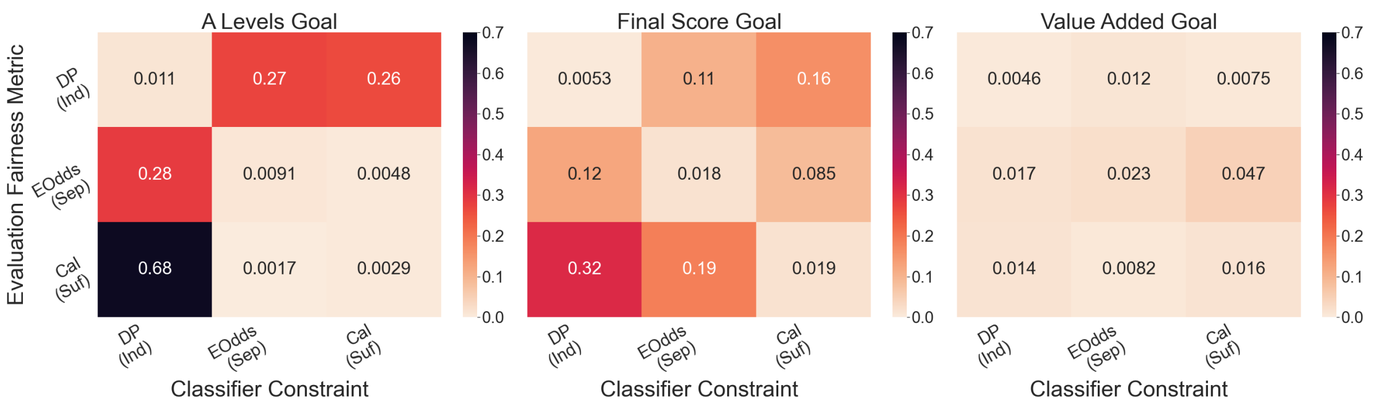
Fairness in Machine Learning, The Importance of Choosing the Right Goal

Student: Callum Ke, Supervisor: Dr. Laurence Aitchison, Project Type: Research

University of Bristol, Department of Computer Science

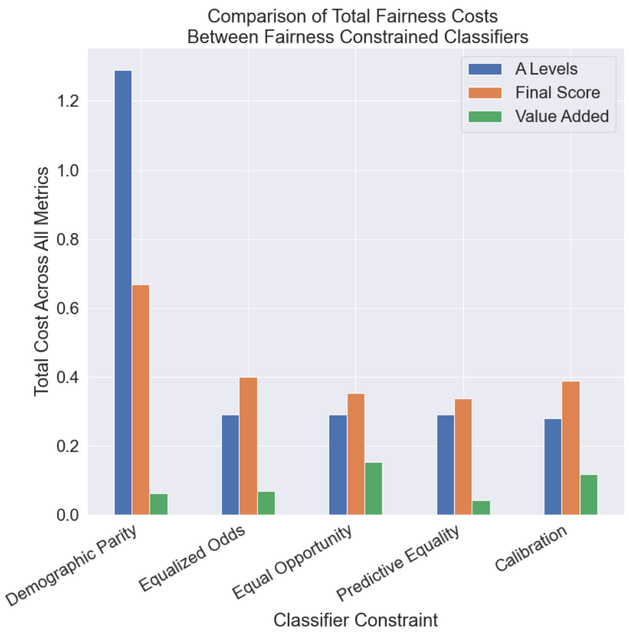
Motivation and Aim

- A fundamental issue within fairness research is the absence of an agreed upon formal definition of fairness that manages to capture all aspects of equality. This is enforced by the Impossibility Theorem that highlights the contradictory nature within commonly used statistical fairness definitions. This is problematic as each of these measures intuitively capture desirable notions of an equitable decision making system thus a fairness trade-off must be made.
- The aim of this thesis is to demonstrate that stakeholder within social decision making contexts, should prioritise their effort into choosing a 'fair' goal if achieving fairness and utility is an inherent aspect of their true objective. In conjunction, we show the futility of trying to optimise for fairness within the machine learning process via post-hoc manners(pre, in, or post-processing) if the model is trained to predict unfair targets.



Experiments and Results

- We use a simulated experiment of University admissions to demonstrate our proposal: comparing the results of classifiers trained to predict acceptance based on: A levels, Graduation Score and a Value Added measure.
- We evaluate the 'value added' measure within as a 'fair' prediction target in comparison to A Levels or the final score. The base rates of each target variable can be seen on the right histogram.
- Our results clearly demonstrate the 'fair' goal: the value added target, can simultaneously obtain low costs across all test fairness metrics. Thus, showing significantly lower overall total fairness costs across all training fairness constraints in comparison to the other 'unfair' goals.



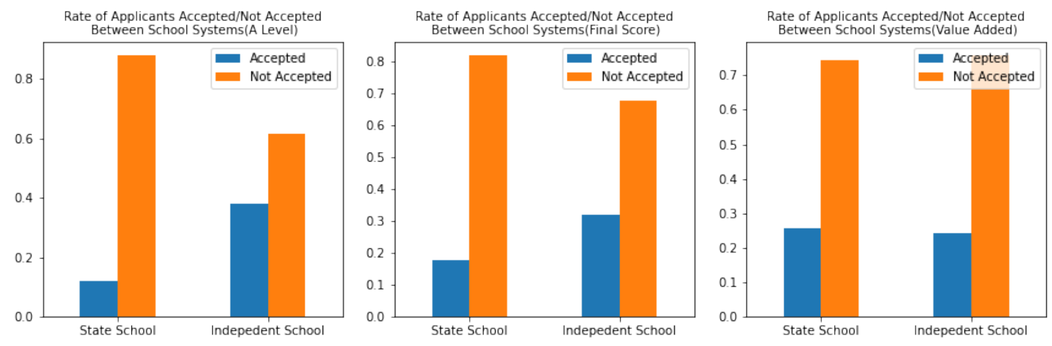
Future Directions

The fairness definitions within this thesis solely rely on equality of statistical metrics between groups. Other fairness definitions such as Causal, Counterfactual and Individual fairness go beyond group statistical fairness criteria that aim to evaluate discrimination within causal frameworks. It would be interesting to explore how choosing a 'fair' goal might also effect mitigating these alternative definitions.

Defining a 'Fair' Goal

In this thesis we define a 'fair' goal such that the underlying distribution has equal base rates of the positive outcome for each of the groups defined by the value of their sensitive attribute.

- Finding a 'fair' goal using this definition is not a trivial task for all decision making contexts, however, if a decision problem inherently prioritises fairness as an objective, then a goal that satisfies this criteria may be reasonably assumed to exist



Definitions of Fairness

$S \in \{a, b\}$ = Sensitive Feature, $Y \in \{0, 1\}$ = Truth labels, $\hat{Y} \in \{0, 1\}$ = Predicted Outcomes given test data.

1. Independence: $\hat{Y} \perp S$
   - Demographic Parity: Equal Selection Rates
   $$P(\hat{Y} = 1|S = a) = P(\hat{Y} = 1|S = b) = P(\hat{Y} = 1)$$
2. Separation: $\hat{Y} \perp S|Y$
   - Equalized Odds: Equal Model Error Rates(FPR/FNR)
   $$P\{\hat{Y} = 1|Y = 1, S = a\} = P\{\hat{Y} = 1|Y = 1, S = b\}$$
   $$P\{\hat{Y} = 1|Y = 0, S = a\} = P\{\hat{Y} = 1|Y = 0, S = b\}$$
3. Sufficiency: $Y \perp S|\hat{Y}$
   - Calibration by Group: Equal Precision Rates(PPV/NPV)
   $$P(Y = 1|S = a, \hat{Y} = y\} = P(Y = 1|S = b, \hat{Y} = y\}$$

The Impossibility Theorem

The Impossibility Theorem states that any pair of fairness definitions will be mutually exclusive at any one time. Except in the following scenarios:

- We can train a perfectly accurate predictor.
- Our predictor trivially assigns all predictions to a single value(0/1).
- The target variable achieves underlying equal base rates between sensitive groups.

Clearly, the third scenario is the only reasonable option in most use cases where utility is also important. We should also note that a perfectly accurate predictor is only sufficient to satisfy two of the three definitions of fairness simultaneously: Separation and Sufficiency only.

University of BRISTOL