DEPARTMENT OF COMPUTER SCIENCE

Mitigating the Conflict Between Fairness Constraints

Evaluating the Effect of Removing Target Selection Bias

Callum Ke

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Bachelor of Science in the Faculty of Engineering.

April 12, 2021

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of BSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.
This project did not require ethical review, as determined by my supervisor. Dr Laurence Aitchison.

Callum Ke, April 12, 2021

Chapter 1

Acknowledgements

I would like to thank my supervisor, Dr Laurence Aitchison, for his insightful and much valued support and guidance during this entire process.

Chapter 2

Abstract

Fairness in Machine Learning is a widely publicised and increasingly prevalent issue in modern society with the rise of AI being increasingly integrated in decision making contexts with large societal impacts. It is often the case that maximising accuracy is the sole priority within Machine Learning [28] processes, thus finding correlations between training features and the target are extremely important. An unfortunate consequence is the introduction of unnecessary biases and discrimination towards certain groups/features, thus calling for considerable thought into what makes socially responsible and fair AI that also provides utility at the same time. Such social decision making scenarios where fair decisions are crucial include: job applications, criminal justice and University admissions. This thesis aims to illustrate: stakeholders who chose to use machine learning models to make decision within such contexts, should prioritise their effort into choosing a 'fair' goal as it inherently removes the problem of fairness itself.

In conjunction, we show the futility of trying to optimise for fairness within the machine learning process via post-hoc manners given the Impossibility Theorem[10], a limitation in many proposed frameworks that aim to achieve overall (statistical) fairness [31], if the model is trained to predict inherently unfair targets. This impossibility result illustrates that certain observational fairness criteria are often mutually exclusive or inherently incompatible except in rare cases and/or with a significant cost in utility. Hence, even if we train a model to satisfy one definition of fairness a sacrifice in a conflicting fairness metric will occur. This is problematic as each of these measures intuitively capture desirable notions of an equitable decision making system. A fairness and utility compromise/sacrifice will have to be made in addition by the stakeholder: a accuracy-fairness trade-off.

Our experiments use a simulated University admissions problem case to demonstrate that choosing a 'value-added measure', a target that satisfies equal base rates between groups is the only reasonable scenario that circumvents the impossibility result and thus a 'fair' goal for a stakeholder to choose. The other two scenarios being: training a perfect predictor or trivially assigning predictions to a single label. Clearly, the latter two are either infeasible or come at a cost at a significant cost in utility. Our experimental results show that using our 'fair' target during our machine learning process does not compromise utility whilst yielding better results across all chosen measurements of fairness constraints. Mitigating the Impossibility Theorem and improving accuracy-fairness trade-offs.

We build upon previously proposed methods([18], [31], [1]) that constrain machine learning models to adhere to certain fairness metrics by treating them as convex constrained optimization problems[36]. We use three models, two of which demonstrate the futility of training a model using an 'unfair' target even if it satisfies a specific fairness constraint. The predictions from these models will ultimately violate or increase the cost with respect to another conflicting fairness metric[10]. In comparison, our model that has been trained to predict a 'fair' target that represents the stakeholders true objective, will yield optimal equity and utility. In doing so: maintaining low group disparities across all definitions of group fairness metrics and achieving reasonable model accuracy rates, even when the model is solely trained to satisfy a singular fairness constraint.

Contents

List of Figures

Chapter 3

Introduction

3.1 Motivation

Machine Learning processes and automated decision making is increasingly being relied upon in societal decision making contexts in a broad range of areas. Common and popular examples can be seen within directed adverting, loan applications and recidivism cases. It is a common misconception that algorithms are solely objective and free of bias, or that it is considered necessary for bias and discrimination to occur for a model to fulfill its purpose; to find correlations in the features in order to maximise the accuracy of its predictions.

However, there have been many cases where unnecessary bias and discrimination is introduced against certain groups. These groups are often defined by a socially 'sensitive' attribute such as race, religion and gender, where discrimination against is both objectively and morally unjustified. This fault cannot solely be attributed to the model itself, software and data is clearly not free of human and societal influence and as a result can reinforce human bias introduced from either the builder of the model and thus the construction of model and/or through the collection and processing of the data in order to fulfill a fair representation of reality.

A fundamental issue within fairness research that applies to both societal and computational problems is the absence of an agreed upon formal definition of fairness that manages to capture all aspects of equality. Formally in law, discrimination can be categorised as either disparate treatment(inequality in treatment) and disparate impact (inequality of outcome). However, there are direct conflicts between both categories portrayed in *Ricci v. DeStefano*[27] that are further enforced in an statistical context by the Impossibility Theorem proposed by the author in [10] with a variant proposed in [3]. Satisfying all fairness constraints is practically impossible even if all three are justified definitions of statistical fairness. A commonly cited case study that shows the manifestation of this problem manifests can be seen in the ProPublica study of the COMPAS recidivism prediction model which tries to predict the potential for recidivism by assigning a 'risk score' to individuals in the criminal justice process. The results show that if we define fairness as equality in error rates, the model was twice as likely to mislabel black defendants as higher risk than white defendants. However, when evaluated against a separate definition of fairness, the results were considered fair as individuals from both groups had equal probability of recidivism if they were predicted as high risk.

Autonomous decision making models within social contexts are commonly designed without enough consideration if the chosen prediction target appropriately reflects the stakeholders true objective [9]. This leads to implementations of Machine Learning models that may produce accurate predictions but are not appropriate within the given context if fairness is an inherent concern. This poorly representative goal often leads to algorithms that don't serve as objective and fair decision-makers, enforcing biases that are introduced/inherent within the goal itself; "[C]ertain kind[s] of biases are inherent in the selection of the goals or objective functions that automated systems will [be] designed to support."[4]. This can also be the case when the features are chosen which implicitly encodes sensitive information which a trained predictor could unknowingly discriminate against.

Unaware or attribute 'blind' classifiers have been proposed with the goal to eliminate 'disparate treatment', However these have shown to be futile in removing 'disparate impact' as sensitive attributes can often be encoded by proxy features. This ultimately leads back to the problem of the 'Impossibility Theorem' where even if we apply a certain fairness constraint to our machine learning model via pre, in or post processing - this will be futile as perfectly satisfying these multiple measures of fairness is

impossible given the biases encoded within the proxy features.

Increasingly more attention is being put into implementing fairness constrained models by removing discrimination within the training algorithm in classification and regression problems [26]. However, the process of debiasing or removing discrimination from a model often happens in a post-hoc manners via pre, in or post-processing. We show this is ineffective when the chosen target label is inherently 'unfair' and the objective is to prioritise fairness and maintain accuracy.

3.2 Problem Description and Thesis Contributions

In ideal scenarios, stakeholders should choose goals that are solely dependent on objective measurements that fairly represent the decision making criteria. Such goals should achieve equal base rates between groups defined by their sensitive attribute. Though the task of finding a goal with equal base rates is not a trivial exercise, it may be reasonable to assume that if a decision problem inherently prioritises fairness as an objective then a goal that satisfies this criteria should also exist. We use this issue as our approach/proposal that will allow us to optimise for fairness by satisfying all statistical definitions of fairness simultaneously and overcoming the Impossibility Theorem.

We demonstrate the effectiveness of choosing goals that have inherent equal base rates between groups, essentially removing bias from the target variable, in a common social decision making scenario. Constructing a predictor this way achieves the aim of obtaining low social and accuracy costs .

We simulate a data-set of applicants for university admissions and train supervised learning classification model to choose the best candidates according to a specified criteria. We compare two 'unfair' models, one that tries to predict an individuals A-level score and the other predicts an individuals final grade when graduating University. These values are then used as the basis for an individuals likelihood for admission. We define an individuals A-level score as positively correlated with their school system(protected feature) thus representing an unfair goal. The final score is generated with a high correlation with the individuals A-level score and thus by proxy is correlated with the school system, again acting as an 'unfair' goal. We compare these 'unfair' models to our proposed 'fair' model that uses a 'value-added' metric as its target. We show our 'value added' metric' satisfies equal base rates rates of admission between between groups and is 'fair' in the context of admissions decisions as it is uncorrelated with the protected feature. The comparison will demonstrate that predictors trained on the 'value-added' goal whilst constrained to satisfy one fairness metric, will simultaneously reduce the cost of fairness amongst other fairness metrics whilst maintaining optimal accuracy-fairness trade-offs. Thus, our proposition allows us to achieve the only reasonable requirement for mitigating the impossibility result. We repeat our experiments in different scenarios: where the sensitive feature is implicitly and explicitly provided during model training, to validate our hypothesis. Each is expected to give equivalent results:

1. The classifiers are 'blind' to the protected attribute but has access to proxy features that encode the protected attribute.

2. The classifiers have access to the protected attribute.

3. The classifiers have access to all features including the protected attribute.

3.3 Thesis Structure

The rest of the thesis will be separated into the following chapters.

Chapter 3: This initial chapter gives insight into the motivation behind our thesis and the problems we aim to solve. We then briefly summarise our contributions and provide a brief overview of the thesis structure and direction.

Chapter 4: This chapter aims to introduce related work, define the concept of discrimination and give insight into how automated decisions made by AI algorithms have been shown to discriminate unfairly in the past. We then formally define the chosen fairness metrics as statistical measures which will allow us to evaluate our models against disparate impact and thus discrimination. Furthermore, we provide the relation of these statistical measurements of fairness to notions of discrimination in law such as disparate treatment and disparate impact. We will then provide some background into the implementation of these fairness metrics as fairness constraints within machine learning algorithms that have been proposed in

recent literature. Finally, we assess the contradictions between the measures, providing an intuition and give proof to why it is impossible to satisfy all fairness notions at once - relating it to the Impossibility Theorem and the complexity of making fair decision in practise.
Chapter 5: We describe the details of generating our data-set for the University admissions problem case and how the training features differ for each experiment. Using data analysis and visualisation, we illustrate the underlying distributions of each of our features, relevant correlations between them and also how each of our target variables might be influenced by certain features. This leads to hypothesising the expected outcomes between our sensitive attribute and the model predictions trained on the three different target variables. We will then provide insight into the specific algorithms for each classifier, how we encode the fairness constraints as convex optimisation problems and the results we expect from each different training target and from each experiment.
Chapter 6: This chapter analyses the results of our experiments by the comparing each of the classifiers trained across different goals and fairness constraints. Using these results we aim to validate out proposal that a fairness constrained classifier trained with the value added goal maintains a low accuracy-fairness trade-off across all fairness metrics. In comparison, the classifiers trained on the 'unfair' goals that are also trained to satisfy each of the fairness constraints, incur large costs within the other fairness measurements that haven't been specifically optimised for and/or with a large cost in accuracy caused by being constrained to satisfy a single fairness definition. Finally, we repeat the following experiments using 'blind' classifiers but with the presence of proxy features and how they give equivalent results.
Chapter 7: Finally, we evaluate our results against our initial hypothesis, linking back to our aims and discuss the potential flaws of our proposal. We conclude with a summary of our thesis findings and discuss possible ideas/pathways for future work.

Chapter 4

Background

4.1 Discrimination in Law

Discrimination is a key part of machine learning, machine learning algorithms are designed to learn what features to discriminate in order to maximise accuracy in its predictions. However, this can often happen in unjustified manners whether the discimination provides no direct impact to performance and/or the discrimination occurs to certain sensitive groups where the end result becomes ethically unjustfiable. We will define the two groups of possible cases of discrimination in the law, **disparate treatment**[37] and **disparate impact**[4]. Such sensitive features may include race, education, religion, gender and a host of other social metrics.

We should note that both of these definitions are too abstract to be formalised in a machine learning model, but they provide a basis to objectively define fairness metrics using common statistical measures.

4.1.1 Disparate Treatment

Disparate treatment occurs when a machine learning model explicitly or intentionally (partially) considers any sensitive feature with its decision. An example of this is when the decision for an individual changes if the sensitive attribute of that individual changes or a model treats subgroups differently when making a decision . For example, the US Civil rights Act of 1964[4] does not allow employment decisions to be made on the basis of these sensitive features. This discrimination can either be direct or indirect via proxy features/quasi-identifiers that have high correlation with the sensitive attribute and can thus be deduced. An example of a proxy feature may be family income to predict the type of school (independent/state) a child went to.

We should note that both of these definitions are too abstract to be formalised in a machine learning model, but they provide a basis to objectively define fairness metrics using common statistical measures.

4.1.2 Disparate Impact

Disparate impact can be identified when an outcome disproportionality benefits or hurts certain groups defined by their sensitive attribute, independent of their true label. This can occur even if the process of making a decision may seem neutral. In order to minimise disparate impact we want to reduce the inequality of outcome([10], [3], [4]). One problem with disparate impact is identifying when the disparity is caused by unfair discrimination or that the outcome is in fact correlated with the sensitive attribute.

4.1.3 The Conflict Between Disparate Impact and Treatment

There have been many attempts to mitigate disparate treatment and impact. However, as mentioned there are direct conflicts between both categories which can be seen in the trial of **Ricci v. DeStefano**[27] where certain individuals were declined promotion even after passing an exam due to the fairness constraint of having proportionate race distributions in the successful candidates. These individuals (specifically one Hispanic and nineteen white) then sued under the basis of disparate treatment - i.e. race based discrimination.

Formally, it can be shown [23] that disparate learning processes that aim to mitigate both forms of discrimination will ultimately fail. For example, if the sensitive feature and nonsensitive features have

any correlation, any DLP will indirectly introduce disparate treatment causing a contradiction in solving both forms discrimination.

This can be linked to the formal Impossibility Theorem([10], [28], [3], [18], [6], [21]) amongst common observational fairness metrics which we later discuss in Section 4.5.

4.2 Sources of Bias

For automated decision making to take place, data must be obtained, processed and inputted to a user specified algorithm must be formalised to train on the data. The predictions given by this model must then be evaluated. However, data is not perfect and neither are the users behind creating and evaluating the algorithms themselves - thus allowing for multiple avenues for bias and discrimination, whether intentional or not, to be introduced into the machine learning pipeline[34](See Appendix F).

In this section we will provide a non-exhaustive list of avenues for bias to occur within the machine learning pipeline and then give a brief history of case studies that have shown how AI have shown bias and/or discrimination in decision making contexts that would be perceived as unfair or morally unjustified.

This thesis mainly focuses on impact of bias introduced by choosing inappropriate or inherently discriminatory target variables and thus causing a biased goal for a machine learning process. However, there are many other sources of bias that are either related or are a direct cause for bias introduced by the goal.

4.2.1 Data Bias

"Garbage in Garbage Out" is a commonly used phrase to describe the importance of correctly gathered and processed data. This is fundamentally the field of Big Data and Data Mining[30] where there has been extensive research [4] into the kind of inputs we want to provide machine learning algorithms to train on so it can accurately predict a target variable.

1. **Historical Bias.** Historical bias is the already existing bias and socio-technical issues in the world[34]. This can cause a feedback loop where past biases(that may not exist now) with regards to decisions replicates itself in an machine learning model even if the data is perfectly measured and selected. In this case, we are not introducing new biases into the system.

2. **Representation/Selection Bias** occurs when the data that is sampled lacks diversity (such as geographical) causing an under representation for some portion of the input space and thus generalizes poorly for some the minority group. This can cause conclusions to be more accurate and thus favor the majority group simply because there is more data. Consequentially, this causes higher uncertainty when predicting the inputs within the minority feature space. This is often attributed to inadequate sampling [34] that represents only a subset of the population or that the training data is not reflective of the data that the model will try to predict.

3. **Measurement Bias - Feature and Target Selection** occurs when the features and labels used do not provide full detail of the problem and would thus create a model that makes false assumptions simply because it lacks the full information with regards to the context. The opposite can also be the case when the granularity of data is too fine which can introduce noise that isn't actually relevant to the problem case.

    - Bias introduced in Feature Selection can also be attributed to proxies[10] that are used to try and predict the ground truth. These proxies can introduce false correlations between a protected class and the target variable via "redundant encoding" in which the protected attribute is encoded within the feature. The largest obstacle with proxies is the accuracy vs fairness trade-off. There isn't an obvious method to determine if a feature is relevant enough in order to predict the target, even if it provides a high correlation with the protected feature and thus the possibility of group discrimination. This can be summarised by the quote: *"[I]t is problematic [to remove a correlated attribute] if the attribute to be removed also carries some objective information about the label [quality of interest]* [8]
    - Measurement bias can often incur and be seen as the result of feedback loops[34]. This is the idea that a given bias is reproduced and enhanced and can be attributed to when the quality of the data is unequal between groups.

- **Target selection bias** (or the goal of our machine learning model) is also under the notion of measurement bias. A label must be chosen for a supervised ML model, however, often the case the label that is chosen can encode additional biases or may be an oversimplification of what the true purpose of the model is.

4.2.2 Algorithmic Bias

Algorithmic bias is purely added by the algorithm itself and is not actually present in the data. This can either be through prejudices that have been applied by the creator of the algorithm or by misinterpreting the data and thus teaching itself prejudices that are not reflected in reality.

1. **Aggregation Bias** occurs when false assumptions are applied to an aggregate population that comprises of many different subgroups , removing the potential of non-linearity between group and outcome. Specifically, this arises when *one-size-fits-all* models are used for groups with different conditional distributions to the target: $P(Y|X)$, thus assuming that the method of predicting an input to the target can be applied to all groups.

   (a) Aggregation can lead to a model that does not perform well for any group even when the groups are equally represented or only the majority group (if representation bias is also present).

   (b) **Simpson's Paradox** can be treated as a form of algorithmic, evaluation and aggregation bias. The definition is where,*" a trend, association, or characteristic observed in underlying subgroups may be quite different from association or characteristic observed when these subgroups are aggregated.* [9]

2. **Evaluation Bias** occurs during model evaluation using an inappropriate benchmark. Individual benchmarks may misrepresent the disparities of a model and oversimplify a success metric. Just as in Simpson's Paradox, a measure of success for the whole population may not be portrayed for individual subgroups. Benchmarks themselves may have data biases within them.

3. **Cause and Effect Bias** occurs in algorithmic systems when correlation is (mis) interpreted as causation. This can often happen when a set of features occur together frequently but are not necessarily causal.

Other forms of bias that are linked to the mentioned definitions above are population bias, social bias, emergent bias, deployment bias to name a few [26].

4.3 Definitions of Fairness

4.3.1 Abstract Definitions of Fairness

One problem that has been noted in past literature is that there is no fixed formal definition of fairness. The best we can do is provide a broad definition such as in [26] where fairness is defined as the *"absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics"*. In this thesis, we treat the type of school system, independent or state, as the binary sensitive or protected attribute and thus we try and show discrimination or lack of discrimination against this attribute.

This definition in itself is too abstract to formalise and is further obscured by the many possible subgroups of the population we can compare this against. For example, we can evaluate the same fairness constraint between individuals, groups or subgroups, each when applied, yielding different results.

Individual Fairness infers that we give similar predictions to similar individuals.

Group Fairness infers that we treat or infer outcomes for different groups equally such that we essentially treat all groups the same way.

Subgroup Fairness tries to merge group and individual fairness notations by checking if a fairness constraint for an individual group can be applied over a subset or collection of groups 47

In this thesis we will mainly focus on **group fairness** constraints which measure parity between specific observational fairness metrics to evaluate the cost of fairness and mitigate disparate impact. As we have mentioned, mitigating disparate impact is a commonly used fairness criteria in the law and in algorithmic scenarios to identify and mitigate discrimination. In [16] this is labelled as the *we are all equal* assumption. It implies that within the decision making process, observed differences between groups are not representative of the true distributional difference in our decision space. In a Universities admissions case, we would assume all differences in the observed feature space (such as A Levels) are inaccuracies and not representative of a students true innate potential. Discrimination is likely if one or more of the statistical measures(i.e. selection rate) is sufficiently different between groups[9]. However, it is not without flaws (discussed in section4.9). These flaws form the basis behind and the appeal of individual fairness and causal inference which we explore in Section 4.10.

4.3.2 Observational Measures

To formalise our fairness constraints, we must first provide the statistical criteria for which they will be defined as. We define the following variables:

Protected/Sensitive Feature $S \in \{a, b\}$ as the group membership of an individual or their protected attribute (in this case their school system).

Binary Decision $Y \in \{0, 1\}$ is the true label of an individual where $Y = 1$ is the positive outcome (the student is admitted) and vice-versa.

Input/Feature Vector $X \in R^F$ as a vector/set of unsensitive features of a particular individual that may or not be correlated with the protected feature S. This is given to a classifier in order to predict admission/rejection.

Predicted Outcomes $\hat{Y} \in \{0, 1\}$ as the predicted outcome of a classifier C given an input X.

Target Population is the joint probability distribution $P(Y, S, X)$

Below are commonly used metrics that can be inferred from the confusion matrix seen in 4.1.

		Predicted Label	
		$\hat{y} = 1$	$\hat{y} = 0$
True Label	$\hat{y} = 1$	True Positive	False Negative
	$\hat{y} = 0$	False Positive	True Negative

Table 4.1: Common Statistical Measures

4.3.3 Whole Table Rates

Definition 4.1 *The accuracy rate is the probability that a student is correctly classified which is defined by the ratio of True Positive + True Negative over the total population N.*

$$Accuracy\ Rate = \frac{TP + TN}{N}$$

Definition 4.2 *The prevalence rate is the pre-test probability of a positive condition within the population.*

$$Prevalence\ Rate = \frac{TP + FN}{N}$$

4.3.4 Row Wise - Conditional Procedure Metrics

Row wise conditional procedure metrics condition on the actual outcome and not the predicted outcome.

Definition 4.3 *The False Negative Rate (FNR/Miss Rate) is defined as the probability that a person who was actually admitted is classified as someone who wouldn't be admitted defined by the equation:*

$$FNR = \frac{Total\ Incorrect\ Negative\ Predictions}{Total\ True\ Positives} = \frac{FN}{FN + TP} = P(\hat{y} = 0|y = 1)$$

Definition 4.4 *The False Positive Rate (FPR/Fall-Out) is the probability that a person who wasn't actually admitted would be classified as someone who would be admitted and is defined by the equation:*

$$FPR = \frac{Total\ Incorrect\ Positive\ Predictions}{Total\ True\ Negatives} = \frac{FP}{FP + TN} = P(\hat{y} = 1|y = 0)$$

Definition 4.5 *The True Negative Rate (TNR/Specificity) is defined as the probability that a person would be predicted to be not admitted given that they weren't actually not admitted.*

$$TNR = \frac{Total\ Correct\ Negative\ Predictions}{Total\ True\ Negatives} = \frac{TN}{FP + TN} = P(\hat{y} = 0|y = 0)$$

Definition 4.6 *The True Positive Rate (TPR/Sensitivity) is the probability that a person would be predicted to be admitted given that they were actually admitted:*

$$TPR = \frac{Total\ Correct\ Positive\ Predictions}{Total\ True\ Positives} = \frac{TP}{FP + TN} = P(\hat{y} = 1|y = 1)$$

We should note that FNR parity is equivalent to TPR parity and FPR parity is equivalent to TNR parity.

4.3.5 Column-Wise Conditional Use Metrics

Column-wise conditional use metrics condition on the predicted outcome and not the actual outcome.

Definition 4.7 *The False Discovery Rate(FDR) is the probability that given a prediction of admission, what is the probability of the truth label being the opposite.*

$$FDR = \frac{Total\ False\ Positives}{Total\ Positive\ Predictions} = \frac{TN}{TN + FN} = P(y = 0|\hat{y} = 1)$$

Definition 4.8 *The False Omission Rate(FOR) is the probability that given a prediction of non-admission, what is the probability of the truth label being the opposite.*

$$FOR = \frac{Total\ False\ Negatives}{Total\ Negative\ Predictions} = \frac{TN}{TN + FN} = P(y = 1|\hat{y} = 0)$$

Definition 4.9 *The Negative Predictive Value(NPV) is the probability that a negative prediction is accurate or the ratio of correctly classified negatives over the total amount of negative classifications. This would correspond to the probability that someone was not be admitted given that they were predicted to not be admitted or the precision on the negative outcome.*

$$NPV = \frac{Total\ True\ Negatives}{Total\ Negative\ Predictions} = \frac{TN}{TN + FN} = P(y = 0|\hat{y} = 0)$$

When the Prevalence Rate p is known, we can calculate the NPV using Bayes Theorem:

$$NPV = \frac{FPR \times P}{FPR \times P + (1 - TPR) \times (1 - P)}$$

Definition 4.10 *The Positive Predictive Value(PPV) is the probability that a positive prediction is accurate or the ratio of correctly classified positives over the total amount of positive classifications. This would correspond to the probability that someone would actually be admitted given they were predicted to be admitted or the precision on the positive outcome.*

$$PPV = \frac{Total\ True\ Positives}{Total\ Positive\ Predictions} = \frac{TP}{TP + FP} = P(y = 1|\hat{y} = 1)$$

When the Prevalence Rate p is known, we can calculate the PPV using Bayes Theorem:

$$PPV = \frac{TPR \times P}{TPR \times P + (1 - FPR) \times (1 - P)}$$

We should note that FOR parity is equivalent to NPV parity and FDR parity is equivalent to PPV parity.

4.4 Formalisation of Fairness Metrics

Group fairness can often be abstractly split into three different criteria: *independence, separation* and *sufficiency* defined by Hardt et al[3]. There are many proposed definitions given to formalising group fairness within the literature (See Table 4.2) , however, it usually the case that they are direct equivalents or relaxations with similar intent of these three labels provided. We will focus on *independence, separation and sufficiency* for measuring observational bias within the rest of this thesis.

Name	Group Fairness Criteria	Relationship
Statistical Parity([14], [6])	Independence	Equivalent
Demographic parity [3]	Independence	Equivalent
Conditional statistical parity [12]	Independence	Relaxation
Equal opportunity [18]	Separation	Relaxation
Equalized odds [18]	Separation	Equivalent
Conditional procedure accuracy [6]	Separation	Equivalent
Predictive Equality [12]	Separation	Relaxation
Balance for Positive/Negative Class [21]	Separation	Relaxation
Avoiding disparate mistreatment[37]	Separation	Equivalent
Conditional use accuracy [6]	Sufficiency	Equivalent
Predictive (rate) parity[37]	Sufficiency	Relaxation
Calibration by group[3]	Sufficiency	Equivalent
Test Fairness [10]	Sufficiency	Equivalent

Table 4.2: Common Labels for Fairness Criteria in the Literature

Overall Accuracy Equality is a metric that may seem intuitive for defining fairness. It is defined by the equality of procedural accuracy between groups i.e. equality of accuracy rate. This would be a valid measure of equality assuming that true negatives are as desirable as true positives, which may not be the case in reality. This is not commonly used because it does not allow us to distinguish between accuracy for success and failures - which we may want to weight.

Total Group Fairness can thus be achieved when the independence, separation and sufficiency are all achieved[6]. We demonstrate in Section 4.5 that this is in fact impossible in practise.

4.4.1 Independence

Definition 4.11 *The Joint Distribution* $(\hat{Y}, Y, S)$ *satisfies Independence if* $\hat{Y}$ *is independent of* S*, this is denoted as* $\hat{Y} \perp S$ *where the likelihood of an individuals classification is independent of the individuals group*

$$P(\hat{Y} = y|S = a) = P(\hat{Y} = y|S = b) = P(\hat{Y} = y) = P(C(x) = y)$$

In the research literature, this is commonly referred to as Demographic Parity [18] or Statistical parity [6]. Specifically, where the positive outcome is desirable:

$$P(\hat{Y} = 1|S = a) = P(\hat{Y} = 1|S = b) = P(\hat{Y} = 1)$$

Intuitively, this can be seen as a method to mitigate *disparate impact* as the outcome is independent of the protected attribute which is often used as in law to imply fair/unfair treatment. We can apply this constraint within machine learning models via both pre, intra or post-processing[6].

The Flaws of Independence

Intuitively, this definition may seem fair as it is essentially blind to the protective attribute thus relating to *fairness through unawareness* which discuss in Section 4.4.4. As a consequence, independence incurs some of the same flaws as a 'blind' classifiers in maximising fairness and accuracy [14].

A classifier that satisfies independence will guarantee to mitigate disparate impact, however it completely removes any possible correlation between the protected attribute and the predictions. This implies that a perfect predictor that satisfies $\hat{Y} = Y$ cannot be satisfied when the base rates of our true labels are different i.e. both groups have different rates of outcomes (See Section 4.5). Thus constraining a machine learning model that is constrained to this fairness metric can seem to be counter intuitive if the goal of a model is to achieve high utility; not only satisfying the fairness constraint.

This idea of non-optimality is expanded on by Hardt [3] with his definition of '*laziness*'. Laziness can occur when the data set is skewed or disproportionate with respect to the protected attribute. The classifier will have more training examples and will more accurately be able to accurately predict the group that is represented more in the training data. This can lead to what is labelled as a *self fulfilling prophecy* [14] that occurs when it more likely to accept unqualified applicants from one group, which will thus further cause negative bias in future decision making decisions. An example of this would be in selecting candidates in employment decisions: we see that an enforced rate of interviews for individuals from the minority groups but there less care is taken to identify the best candidates to interview from the minority group Thus, leading to potential justifications for future discrimination.

[6] also emphasises the accuracy and fairness trade-off cause by Independence. If the goal is to train a predictor that provides utility. Then the training process must "capitalize on non-redundant associations" that the sensitive attribute has on the outcome and "excluding S will reduce accuracy".

4.4.2 Separation

Separation acknowledges and allows for the target variable to be correlated with the sensitive attribute and thus tackles one of the flaws of Independence by promoting the incentive to reduce errors uniformly between groups. It uses *row-wise* procedural measures to evaluate fairness. The formal definition of *separation* is:

Definition 4.12 *The joint probability distribution* $(\hat{Y}, Y, S)$ *satisfies separation if* $\hat{Y} \perp S|Y$ *i.e.* $\hat{Y}$ *is conditionally independent of* S *given the true labels* Y *- represented by the graphical model in 4.1*

In this case, since C *is a binary classifier, separation can also be defined by the following constraints for both groups:*

$$P\{\hat{Y} = 1|Y = 1, S = a\} = P\{\hat{Y} = 1|Y = 1, S = b\}$$
$$P\{\hat{Y} = 1|Y = 0, S = a\} = P\{\hat{Y} = 1|Y = 0, S = b\}$$

Thus we define *separation*(when the positive outcome is desirable) as requiring equality of the true positive and the false positive rate between groups - labelled as *true positive parity* and *false positive parity*. When the negative outcome is more desirable, this is equivalent to equality of false negative and true negative rates between groups. Intuitively, this can be interpreted as the probability of an individual being accepted(resp. denied) to be the equal independent of the group.

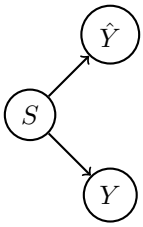

Figure 4.1: Graphical Model/Bayesian Network Representing the Conditional Relationship Between $\hat{Y}, Y$, S when Separation is satisfied

In the literature, separation is often labelled as Equalized Odds[18], Conditional procedure accuracy equality or Positive Rate Parity and has be applied via in([36], [37], [38]) or post processing([18]). The post-processing step can be visualised using *Receiver Operating Characteristic* (ROC) curves which is defined by the cost of true and false positive rates depending on a classifiers threshold to determine a binary outcome for a score(i.e. added value). The area under the intersection of both lines shows the achievable regions for classifiers that can achieve Equalized Odds, each defined by a threshold pair of acceptance for each groups. This is shown in Figure 4.2 demonstrating the trade-offs between true and false positive rates are not always the same/achievable for both groups. ROC curves allow us to visualise, choose and know the limitations of a classifier that can satisfy Equalized Odds; we need to choose a classifier that lies within the shaded region that minimizes the cost function with respect to accuracy.
Two possible relaxations of *separation* are noted in the literature: **Predictive Equality** and **Equal Opportunity** which we define respectively below.

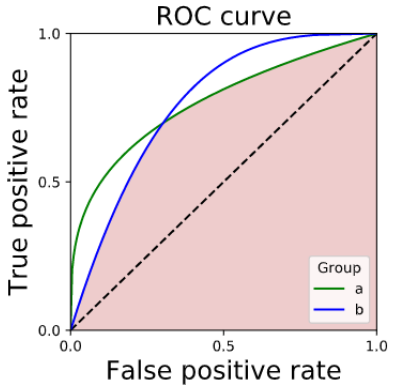

Figure 4.2: Receiver Operating Characteristic Graph [3]

Definition 4.13 ***Equal Opportunity*** *requires the first constraint (True Positive Parity) when the positive outcome (1) is desirable:* $P\{\hat{Y} = 1|Y = 1, S = a\} = P\{\hat{Y} = 1|Y = 1, S = b\}$

Intuitively, satisfying Equal Opportunity requires non-discrimination within the positive outcome group i.e. the individuals whom were admitted would have equal probability of being classified to also be admitted. This is a weaker definition of non-discrimination w.r.t Equalized Odds but can allow more flexibility and utility. It is also equivalent to False Negative Rate Parity

Definition 4.14 ***Predictive Equality*** *requires the second constraint(False Positive Parity):* $P\{\hat{Y} = 1|Y = 0, S = a\} = P\{\hat{Y} = 1|Y = 0, S = b\}$

Intuitively, satisfying Predictive Equality gives equal probabilities of acceptance for applicants whom wouldn't be admitted. A notable use of this definition can be seen within the COMPAS case study where a major criticism was identified to which the rate of false positives(where positive meant classified as being likely to re offend) was much higher among black than white populations[10]. It is also equivalent to True Negative Rate Parity

We can thus define *Equalized Odds* as a combination of when *Equal Opportunity* and *Predictive Equality* are both satisfied simultaneously by a predictor. The predictor is thus incentivised to perform well in both groups, not just the majority(as in the case of demographic parity).

Flaws of Separation

Separation ultimately measures *'model errors'* that *NorthPointe* [29] argues is of *'no practical use'* to a practitioner who is assessing the probability of an outcome actually occurring. Using separation as a fairness metric is contingent on knowing the ground truth label upon making a decision and thus can be argued as impractical at the time of making a decision if the decision maker does not know the truth label of an individual at the time of making a decision. This is clearly the case of recidivism predictions, where a practitioner has no information if the individual would reactivate or not and thus cannot measure fairness with respect to separation.

Given the definition of Separation relies on the equality of model errors, we expect to have accuracy comparable to that achieved for the protected group category for which accuracy is the worst. Implementing Separation requires reassigning prediction values based on some probability. However, the values chosen for the reassignment probabilities will need to be larger when the base rates between sensitive groups are more disparate. This is reinforced by how we define the cost of violating Separation in Section 4.6.2. Intuitively, when Separation is most likely to be violated, the damage to Separation will be the greatest with respect to accuracy and fairness. More classification errors will be made; as more positive labels will be treated as negatives and vice-versa. A consolation may be that all groups will be equally worse off [6].

4.4.3 Sufficiency

Sufficiency uses column-wise conditional use metrics to measure fairness such as positive predictive value and negative predictive values, conditioning on the classifiers predictions and defining rates based on

the likelihood of the classifiers prediction to be reflected in reality. In other words, this condition states that for every $\hat{Y}$, people in both protected and unprotected (private and state) groups must have equal probability of correctly belonging to the admitted set of candidates.

Definition 4.15 *The joint probability distribution (R, Y, S) satisfies separation if $Y \perp S|R$ i.e. Y is conditionally independent of S given a random variable R representing a score. This is represented by the graphical model in Figure 4.3*

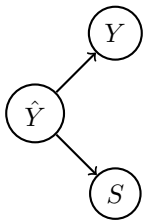

Figure 4.3: Graphical Model/Bayesian Network Representing the Relationship Between $\hat{Y}, Y$, S when Sufficiency is Satisfied

As a probabilistic equation, sufficiency is satisfied in a binary classifier if both the positive and negative predictive value hold. They are equivalent to *False Discovery Rate Parity* and *False Omission Rate Parity* respectively. Intuitively, this can be interpreted as the ground truth label Y of an individual to be independent of the individuals group membership given a predicted outcome $\hat{Y}$. When the positive outcome is desirable, Sufficiency requires the following two conditional probability statements to hold:

$$\text{PPV Parity} = P\{Y = 1|\hat{Y} = 1, S = a\} = P\{Y = 1|\hat{Y} = 1, S = b\}$$

$$\text{FOR(NPV Parity)} = P\{Y = 1|\hat{Y} = 0, S = a\} = P\{Y = 1|\hat{Y} = 0, S = b\}$$

Intuitively, this equalizes the chance of success given predicted acceptance. In the literature, this is often defined as Predictive (Rate/Value) Parity([18], [3]) and is equivalent to *calibration by group/well-calibration* which is a common alternative used in the literature([31],[3],[35]). Sufficiency is a slightly weaker notion than calibration but for all purposes are essentially equivalent notions which we see in Proof 4.1. The author in [10] introduces *Test Fairness* as an alternative name to calibration and shows that it implies equality in *positive predictive value* between groups as a necessary condition for test fairness and thus calibration.

Definition 4.16 *A classifier satisfies Test fairness if: $P(Y = 1|S = a, \hat{Y} = y\} = P(Y = 1|S = b, \hat{Y} = y\}$*

Definition 4.17 *A score R is well-calibrated if for all values: $r \in R$ we have $P\{Y = 1|R = r\} = r$*

Definition 4.18 *Calibration by Group is satisfied when $P\{Y = 1|S = s, R = r\} = r$.[3]*

Proposition 4.1 *If a score r satisfies sufficiency, then there exists a function $f : [0, 1] \to [0, 1]$ so that $f(R)$ maps R to a true label Y such that it satisfies calibration by group.*

Proof 4.1 *Fix any group $S = a$ and define $f(r) = P\{Y = 1|R = r, S = a\}$. Since R satisfies sufficiency then the probability $f(r)$ is the same across all groups. Consider the two groups a, b:*

$$\begin{aligned} r &= P\{Y = 1|f(R) = r, A = a\} \\ &= P\{Y = 1|R \in f^{-1}(r), S = a\} \\ &= P\{Y = 1||R \in f^{-1}(r), S = b\} \\ &= P\{Y = 1|f(R) = r, S = b\} \end{aligned} \tag{4.1}$$

Thus showing f(R) is calibrated by group and thus a binary classifier that thresholds a score R: $f(R) \in \{0, 1\}$ is also calibrated by group and satisfies sufficiency.

Calibration allows us to interpret values of a given score function as probabilities. It implies that the set of individuals instances that have been assigned a score of r will has a fraction of r of positive instances. This can can be treated as an accuracy score for a group but also as an unfairness metric if

calibration is only satisfied for a subset of all possible groups. It is worth noting that this does not imply a single individual within that set has a probability of r to have a positive outcome.

Intuitively, satisfying 'within' group calibration prevents probability scores from containing proxy information that may be contained within the sensitive attribute/group as calibration as it acts as a kind of normalization. It also implies that individuals, independent of group, with a certain score r will have equal probability being from the positive class which the probability must equal r.

Calibration(and in turn sufficiency) is satisfied if individuals who are predicted with a certain outcome, independent of the group, have the same probability of having their true label being from the positive class(admitted).

4.4.4 Fairness Through Blindness and Unawareness

A trivial approach to mitigating disparate mistreatment may be simply ignoring the protected attribute all together when training the model by removing it from the training data all together. Formally this is labelled as *fairness through unawareness.* A predictor satisfies *fairness through unawareness* if protected attributes are not explicitly used in the prediction process.

Intuitively, this may seem fair as the classifier has no explicit information about the protected attribute to train on. However, there are many drawbacks that limits the use of *unaware* classifiers in practical scenarios that make *fairness through awareness* a very weak fairness constraint, ultimately leading to a moral discussion of whether or not holding all groups to the same standard is fair even if there are real differences that might not be unfair to ignore. We show the ineffectiveness of 'unaware/blind' classifiers in Chapter 6.1.

Measurement bias via proxy variables is the most common criticism. As features are often correlated in real life scenarios, a classifier could infer the protected attribute as it may be implicitly encoded in the data. This is called *redundant encoding*[18]. Furthermore, if the protected attribute is in fact encoded, then the goal of the classifier will ultimately reduce to achieving high accuracy without a fairness constraint.

Another flaw with ignoring sensitive attributes is that it seems counter-intuitive to the goal of machine learning processes where we want to achieve accurate predictions. As *"removing all such correlated attributes before training does remove discrimination[through disparate treatment], but with a high cost in classifier accuracy."*[4] Fairness through blindness may remove meaningful information, causing a conflict with the goal to make accurate determinations.

Many real practises have been seen to use a 'race-blind' approach such as within housing applications, education admissions, credit scoring and criminal justice[17]. The ineffectiveness of *'blind classifiers'* of both maximising accuracy and fairness is shown using the FICO score data-set ([18], [10]) . We see that a race blind classifier where the objective is to maximise profit, achieves very high profit but clearly discriminates against a subset of races. This is due to the underlying measurement/historical bias that is introduced by using FICO scores as a proxy for default rates and thus a measurement for predicting profit. The FICO score and other training features acted as proxies for race even when it wasn't explicitly given to the classifier. The historical distributions of FICO scores between races also had inherent encoded racial discrimination. As the objective was to maximise profit, fewer loans were given to the group with historically lower FICO scores - even when both groups were equally qualified to obtain a loan. Consequentially, a company(the bank) who's goal is to maximise profit, will assume that their classifier is fair due to the blindness constraint and will also having little incentive to improve accuracy as they are achieving optimal profit. However, in reality discrimination occurs.

Fairness through Unawareness emphasises the conflict of*disparate treatment vs disparate impact* that we discussed in Section 4.1.3 and also the seemingly conflicting notions of fairness and utility.

4.5 Trade-offs Between Fairness Metrics: The Impossibility Theorem

The *'Impossibility Theorem'* is one of the key limiting factors of defining a 'completely' fair classifier with respect to observational fairness metrics which has been demonstrated in many recent papers ([21], [6], [31], [3]). This describes that no classifier that can achieve equality selection rate, false negative **and** false positive rates and precision rates simultaneously with the exception in trivial cases or a perfect/trivial classifier exists. Simply stating, the impossibility result demonstrates that a classifier cannot satisfy *well-calibration* in conjunction with *equalized odds* unless the base rates - the fraction of individuals with

positive class labels - are equal between protected groups differ, the classifier is an perfect predictor[10] or it trivially assigns all examples to a single class[38]. This is equivalent to showing a stronger notion that only two of three criteria: *Independence, Separation and Sufficiency* can be satisfied at once.

In practise, this leads to stakeholder's having to consider fairness trade-offs which may be acceptable in different contexts. For example, a judge may consider the the false positive rate(being incorrectly classified for recidivism) with much higher weighting than equality in selection rate between groups when making judgments in the criminal justice system as this would incur much higher social costs or simply being a much more serious error.

4.5.1 Trade-offs Between Independence, Separation and Sufficiency

More abstractly the impossibility results can also be proven using the three group fairness criteria: Independence, Separation and Sufficiency. It can be proven that all three cannot simultaneously satisfied except in degenerate or very rare cases[3]. In fact, any two of three are mutually exclusive at any one time. Intuitively, this is because all three constrain the joint distribution $(\hat{Y}, Y, S)$ in some way. Thus we may expect that imposing one of these constraints onto the distribution may only lead to trivial cases remaining if we were to satisfy the remaining two. Ultimately, these three criteria cannot be imposed as hard constraints which leads us to try and find trade-offs/relaxations between them which we explore in section 4.5.7. This idea is fundamental to our proposal of choosing 'fair' goals or targets. We want to choose goals that are close to these 'rare' instances where all three cases can be close to optimal. The proofs below are directly inspired by ([2] ,[3]).

We first define conditional independence between three random variables A, B, C as:

Definition 4.19 *Random variables A, B are independent conditional on random variable C when:*

$$(P(A|C)P(B|C) = P(A \wedge B|C)) \implies A \perp B|C$$

Intuitively, this means that given the event C has occurred, the event A occurring provides no information on the likelihood of B occurring.

4.5.2 Independence vs Sufficiency

Proof 4.2 *Assuming both independence and sufficiency hold - using the contraction rule[32]:*

$$(S \perp \hat{Y}) \wedge (S \perp Y|\hat{Y}) \implies S \perp (Y, \hat{Y}) \implies S \perp Y$$

Intuitively, this says that independence and sufficiency can only hold when the protected attribute is independent from the ground truth/target labels implying that both groups(defined by the sensitive attribute) should have equal base rates. Thus by contra-positive, unless equal base rates are satisfied, independence and sufficiency are mutually exclusive.

$$S \not\perp Y \implies (S \not\perp \hat{Y}) \vee (S \not\perp Y|\hat{Y})$$

.

4.5.3 Independence vs Separation

This proof requires that Y is binary, which in our problem case (admit vs not admit) is a reasonable assumption.

Proposition 4.2 *If we assume S is not independent of Y and $\hat{Y}$ is not independent of Y, then independence and separation are mutually exclusive:*

Proof 4.3 $(S \not\perp Y) \wedge (\hat{Y} \not\perp Y) \implies (S \not\perp \hat{Y}) \vee (S \not\perp \hat{Y}|Y)$ *Thus using proof by contrapositive:* $S \perp \hat{Y} \wedge S \perp \hat{Y}|Y \implies S \perp Y \vee Y \perp \hat{Y}$.

By the law of total probability:

$$P\{\hat{Y} = y|S = a\} = \sum_{y} P\{\hat{Y} = y|S = a, Y = y\} P\{Y = y|S = a\}$$

Applying our assumption $S \perp \hat{Y} \wedge S \perp \hat{Y}|Y$*:*

$$P(\hat{Y} = y) = \sum_y P(\hat{Y} = y|Y = y)P(Y = y|S = a) = \sum_y P(\hat{Y} = y|Y = y)P(Y = y)$$

Using our assumption that $\hat{Y} \in \{0, 1\}$ *i.e. a binary classifier, the above equation can only be satisfied if base rates are equal:* $S \perp Y$ *or that our predictions are independent of our target variables:* $\hat{Y} \perp Y$*. Intuitively, this is counterproductive towards our machine learning model that uses supervised learning to train our predictions as it will have no utility at all.*

4.5.4 Separation vs Sufficiency

Proposition 4.3 *Imposing both separation and sufficiency on the joint distribution will lead to a degenerate or rare solution space.*

Proof 4.4 *Assume all events in the joint distribution* $(S, Y, \hat{Y})$ *have positive probability. If* $S \not\perp Y$ *then separation and sufficiency are mutually exclusive (only one can hold).*

Using the Waserman Theorem: $(S \perp \hat{Y}|Y) \wedge (S \perp Y|\hat{Y}) \implies S \perp (\hat{Y}, Y) \implies (S \perp Y) \wedge S \perp \hat{Y})$

Therefore by taking the contrapositive: $S \not\perp Y \implies (S \not\perp \hat{Y}|Y) \vee (S \not\perp Y|\hat{Y})$ *thus completing the proof.*

This can also be proven using statistical measures directly relating to **Equalized Odds** and **Well-Calibration** which we discussed in Section 4.5.5.

Intuitively, this means that if the sensitive value is not independent from the target which means having unequal base rates, then only sufficiency or separation can hold but not both. This is the main cause for conflict between the *NorthPointe vs ProPublica* findings when evaluating fairness of COMPAS risk scores [29], [19]. NorthPointe argued the risk scores were considered fair on the account of sufficiency, whereas ProPublic argued for discrimination on the account of separation. NorthPointe suggested that due to unequal base rates within COMPAS scores between races, it was unrealistic to satisfy both sufficiency and separation and that sufficiency was contextually a better measure of fairness.

4.5.5 Incompatibility of Calibration and Equalized Odds

Though these two fairness metrics ultimately aim to achieve the same goal of having equal probability criteria between groups - and hence would ideally be satisfied simultaneously, the impossibility result demonstrates they are incompatible with each other, even in approximate cases. This clearly introduces a limitation to achieving perfectly fair classifiers(with respect to group fairness definitions). In practical scenarios, it will be unlikely to create a 'perfectly fair' classifier as we must either find a dataset that has equal underlying distributions between groups or create a clasifier is perfectly accurate.

In the case of a binary classifier C, we assume that in non-degenerative cases(such as C trivially assigning all $\hat{Y}$ to one value), the predictions of C will have instances of both positive and negative classes. Thus an imperfect classifier will make at least one error and thus must have $FPR > 0$ and $TPR \geqslant 0$

Proof 4.5 *Assume unequal base rates(positive prevalence)* $S \not\perp Y$*:* $p_a = P(Y = 1|S = a), \neq P(Y = 1|S = b) = p_b$ *(imperfect classifier* C *such that for some* $x \in X$*,* $C(x_i) \neq Y_i$ *and positive rate parity(equalized odds).*

$$PPV_a = \frac{TPR \cdot p_a}{TPR \cdot p_a + FPR(1 - p_a)}, NPV_a = \frac{(1 - FPR) \cdot (1 - p_a)}{(1 - TPR) \cdot p_a + (1 - FPR)(1 - p_a)}, a \in S$$

In order to satisfy well-calibration/sufficiency both the following constraints must be satisfied:

$$PPV_a = PPV_b \implies \frac{TPR \cdot p_a}{TPR \cdot p_a + FPR(1 - p_a)} = \frac{TPR \cdot p_b}{TPR \cdot p_b + FPR(1 - p_b)}$$

$$NPV_a = NPV_b \implies \frac{(1 - FPR) \cdot (1 - p_a)}{(1 - TPR) \cdot p_a + (1 - FPR)(1 - p_a)} = \frac{(1 - FPR) \cdot (1 - p_b)}{(1 - TPR) \cdot p_b + (1 - FPR)(1 - p_b)}$$

For both to be satisfied, either $TPR = 0$ *or* $FPR = 0$ *(but not both since as our classifier* C *is imperfect).*

Given the above, either equality in PPV fails or equality of NPV fails due to the assumption of equalized odds($TPR_a = TPR_b \geqslant 0, FPR_a = FPR_b \geqslant 0$) and unequal base rates p. Implying that ***sufficiency/well-calibration*** *is mutually exclusive of* ***equalized odds*** *when $p_a \neq p_b$ i.e. unequal base rates.*

Intuitively, if base rates of acceptance between groups of independent and state school educated individuals, there will be an immediate fairness cost. If false positive and true positive rates are equal(Separation), due to the larger base rates for one of the groups, then a lower of proportion of the group with the higher base rate will be incorrectly predicted to be rejected(Sufficiency). This can easily be viewed as discrimination towards the group with the lower base rate. And, there is no doubt that accuracy will decline and will decline more when the probabilities of reassignment are larger.

A similar proof proposed in [10] shows an equivalent relationship:

$$FPR = \frac{p}{1-p}\frac{1-PPV}{PPV}(1-FNR)$$

This demonstrates that if p differs between groups, a score that satisfies equality in PPV(calibration) cannot have equal false and negative rates between groups.

4.5.6 Geometric Relationship Between Equalized Odds and Calibration

A geometric intuition of the relationship between calibration between groups and equalized odds is seen in [31] and in figure 4.4 below.

We can define all trivial classifiers as $\forall x : h(x) = c$ where $0 \leqslant c \leqslant 1$ is a constant.

The *generalized false-positive rate* of a classifier h_s for group $s \in S$: $c_{fp}(h_s) = E_{(x,y)}[h_s(x)|y = 0]$ where the protected group of $x = s$.

The *generalized false-negative rate* of a classifier h_s for group $s \in S$: $c_{fn}(h_s) = E_{(x,y)}[1 - h_s(x)|y = 1]$ where the protected group of $x = s$.

Thus *equalized odds* is satisfied if $c_{fp}(h_1) = c_{fp}(h_0)$ and $c_{fn}(h_1) = c_{fn}(h_0)$

Classifier h_s is perfectly calibrated if $\forall p \in [0, 1] : P_{(x,y)}[y = 1|h_s(x) = p] = p$. Thus for well-calibration to be satisfied, h_0, h_1 should be perfectly calibrated.

We define all trivial classifiers h^{μ_1}, h^{μ_0} as lying on the diagonal defined by $c_{fp}(h) + c_{fn}(h) = 1$ where the x axis is defined by the generalized false positive rate and the y axis is the generalized false negative rate.

The generalized false-positive and false-negative rates of all possible calibrated classifiers for each group H_1^*, H_0^* can be linearly separable by their groups base rates μ_s: $c_{fn}(h_s) = \frac{(1-\mu_s)}{\mu_s c_{fp}(h_s)}$.

h_1 lies on the line with gradient $(1 - \mu_1)/\mu_1$ and h_0 lies on the line with gradient $(1 - \mu_0)/\mu_0$. The perfect classifier for each group lies on origin of each line (respectively the trivial classifier lies on the diagonal). Thus for both groups, both error rates decrease as we move closer to the origin on the given line defined by a base rate μ_s implying a more accurate classifier.

We can now redefine the impossibility theorem using the geometric interpretation of all possible calibrated classifiers between each group:

Let h_1, h_0 be classifiers for group S_1, S_0 where the base rates differ:$\mu_1 \neq \mu_0$

h_1, h_0 can satisfy Equalized Odds and Calibration simultaneously $\iff$ h_1, h_0 are perfect predictors i.e. $c_{fn}(h_1) = c_{fn}(h_0) = 0$ = the origin point/intersection of both lines of calibrated classifiers of each group. This means unless we relax our constraints, perfect classifiers must be achieved to satisfy equalized odds and within group calibration.

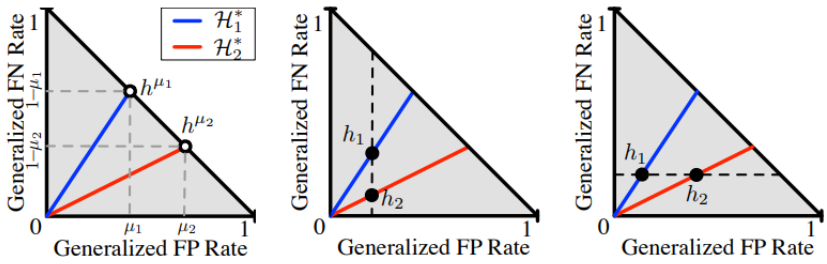

Figure 4.4: Graphical Representation of Calibration and Satisfying Relaxed Equalized Odds [31]

4.5.7 Relaxing Equalized Odds to Preserve Calibration

As we noted in Section 4.4.2, **Equalized Odds** can be relaxed to **Equal Opportunity** or **Predictive Equality**. We show that using these relaxations we can satisfy **Calibration by Groups** simultaneously with one of these two relaxations. This is relevant in practice as there is still value to exploring the best cost of fairness a classifier can achieve even if imperfect.

We define a cost function g_s to be linear function with dependence on the groups base rate:
$\mu_s : g_s(h_s) = a_s c_{fp}(h_s) + b_s c_{fn}(h_s)$ (Figure 4.5). a_s, b_s are non-negative constants that are specific to each group.
We assume at least one of $a_s, b_s > 0$ meaning if $g_s = 0 \implies c_{fp}(h_s) = c_{fn}(h_s) = 0$. This would imply we have no errors and thus a perfect classifier. This also shows that for calibrated classifiers, an increase in error rates (c_{fn}, c_{fp}) corresponds to an increase in cost g and an increase in the distance of the cost curve to the origin. The cost function g can be used as a proxy to the error rates.

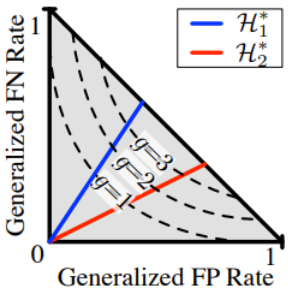

Figure 4.5: Level-order curves of cost. Low cost implies low error rates[31]

Proposition 4.4 *Given a cost function g_S, classifiers h_1 and h_2 achieve Relaxed Equalized Odds with Calibration for groups S_1 and S_0 if both classifiers are calibrated and satisfy the constraint $g_1(h_1) = g_0(h_0)$.*

The intuition is that we post-process our calibrated classifiers[18] to satisfy either *Equal Opportunity* or *Predictive Equality* as it is not a trivial task to encode calibration within the optimization problem themselves. The proof of optimality is demonstrated in [31].

We will assume that our classifiers h_0, h_1 are optimal with respect to their costs g_0, g_1 for the purpose of demonstrating calibrated and relaxed equalized odds constrained classifiers in a real data set. The experiment demonstrated in Figure 4.6 demonstrates the incompatibility of trying to satisfy Equalized Odds and Calibration by groups simultaneously. However, it also shows the effectiveness of trying to satisfy calibration with one of the relaxed Equalized Odds definitions. In this experiment we ensure equality of false negative rates (equal opportunity) between groups (Male and Female) whilst achieving calibration.

Looking at 4.6, we see the calibrated classifiers for each group lie on either the blue or red line.The coloured circles represent the optimal calibrated classifiers for each group. The diamonds represent the classifiers on the experiment constraint(Left: Equalized Odds, Right: Calibration + Equal FNR). On the left plot we see that we can approximately match the error rates and thus satisfying equalized odds

between groups but sacrificing calibration for group 1(blue line). On the right plot, we see that we can approximately equalize false negative rates between classifiers h_1, h_0 whilst maintaining calibration by groups. Another observation is that maintaining calibration will result in divergence of the FPR between groups and thus incurring a trade-off of predictive equality.

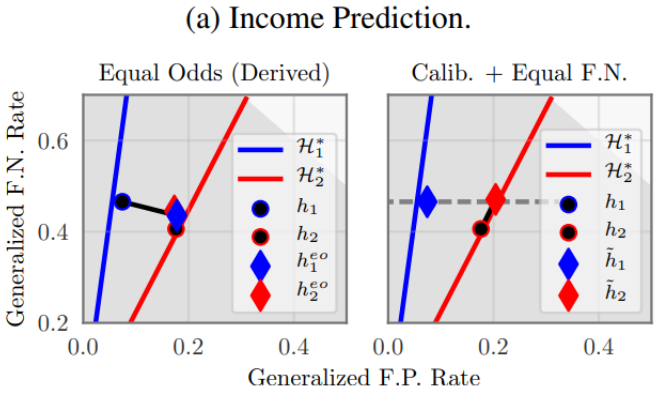

Figure 4.6: Left: Post-Processed Classifiers trained on Derived Equalized Odds(Diamonds). Right: Calibrated and Equalized Odds relaxation(Equal Opportunity) Constrained Classifiers via Post-Processing (Diamonds). Classifiers aim to predict Individual Incomes from the Adult Data-set[13] .[31]

4.6 Measuring Disparity of Group Fairness Constraints

This sections presents an abstract view of how to evaluate the cost of a certain fairness constraints that represents how much we have violated a certain group fairness criteria. The ideas in this section are formalised in Section 5.3 where we define each fairness constraint as optimization problems using a set of previously proposed algorithms that allow us to implement 'fair' classifiers(and thus solve fairness-aware classification problems) which will then be used within our experiments.

In order to implement group fairness constraints as optimization problems, we need to define them a cost metric. Intuitively, a method to give a sense of how fair/unfair a certain decision maker is by evaluating the difference of some metric between groups. In the literature, this is defined as a measure of disparate impact([10]), *immediate utility difference*[12] or the *risk difference*[36].

Each fairness cost can be defined using both the positive and negative predicted cases, within this thesis we solely focus on the metrics that use the positive case: $\hat{Y} = 1$. This is justified as a University admissions board is more inclined to prioritise equality of predicted acceptance(versus non-acceptance) between groups.

4.6.1 Independence: Demographic Parity

Definition 4.20 *We define the value of Demographic Parity Violation as the difference of probability of predicted admission between groups. Violating this property means that the probability of admission conditioned on each group is not equal.*

$$C_{DP} = |P(\hat{Y} = 1|S = 1) - P(\hat{Y} = 1|S = 0)|$$

This can also be rephrased as the probability of predicting a positive outcome for the whole population is independent of the protected feature. Therefore the cost of Demographic Parity can also be defined as:

$$C_{DP} = \max_a |P(\hat{Y} = 1|S = a) - P(\hat{Y} = 1)|$$

4.6.2 Separation: Equalized Odds

We recall from Section(4.4.2) that Equalized Odds is the combination of equality in False positive and true positive rates between groups given the priority on the positive outcome. These two measures define the relaxations of Sufficiency: *Equal Opportunity and Predictive Equality.* We can thus define

the measurement of violation of *Equalized Odds* as the maximum of the violation of the two relaxations. Intuitively, violating equalized odds means we are more likely to correctly predict one group to be admitted than the other or that we are more likely to wrongly admit one group than the other. This was the main account of discrimination in the *ProPublica* analysis as mentioned before.

Equal Opportunity

Definition 4.21 *The violation of Equal Opportunity is measured as the difference in true positive rates:*

$$C_{E\_Opp} = |P(\hat{Y} = 1|Y = 1, S = 1) - P(\hat{Y} = 1|Y = 1, S = 0)$$

Predictive Equality

Definition 4.22 *The violation of Predictive Equality is measured as the difference in false positive rates:*

$$C_{PE} = |P(\hat{Y} = 1|Y = 0, S = 1) - P(\hat{Y} = 1|Y = 0, S = 0)$$

4.6.3 Sufficiency

We define a classifier to satisfy sufficiency if it has equal *positive predictive* and *negative predictive values* between groups. This is sometimes called ' *Test Fairness*' as it shows the difference in precision rates between groups. Intuitively, a violation in sufficiency means that a positive prediction is more likely to be precise/accurate for one group.

Definition 4.23 *In the positive case, we define our measurement of violation of sufficiency in Section 6.3 as the maximum difference between the positive predictive value and the the false omission rate between groups.*

$$C_{Suf} = |P(Y = 1|\hat{Y} = 1, S = 1) - P(Y = 1|\hat{Y} = 1, S = 0)$$

Calibration by Groups:

Recall that Sufficiency is a slightly weaker relaxation of *Calibration by Groups*(Well Calibration) $Pr(Y = y|R = r, S = s) = r$. We proved in Section 4.4.3 that the same can be applied in the binary case $R \in \{0, 1\}$ and is thus equivalent to our definition of Sufficiency violation i.e. the maximum difference in *positive predictive value* and *false omission rate* between groups.

4.6.4 Risk Ratio:

A common method to define fairness violations is measuring the ratio of a statistical measure between groups. The closer the ratio is to 1, the more fair a classifier is with respect to a certain fairness metric.

$$RR = \frac{\text{Metric}_{S=0}}{\text{Metric}_{S=1}}$$

4.6.5 Approximated Costs

An approximated version of each of three costs, that can be seen used within practical decision applications for *disparate impact* law [7]. This is commonly used especially when within fairness constrained machine learning models in order to allow for flexibility to achieve better fairness-accuracy tradeoffs [38]. It is defined:

$$\frac{\text{Metric}_{S=0}}{\text{Metric}_{S=1})} \geqslant 1 - \epsilon \text{ or } |\text{ Metric}_{S=0} - \text{ Metric}_{S=1}| \leqslant \epsilon$$

This can be linked to the p-rule proposed by [33].

4.7 Techniques to Implement Fair Classifiers

There are various methods that have proposed to achieve fairness constrained classifiers. Each fall under the three general implementation categories: *pre-processing, in-processing and post processing*[26], [3]. We should note that in general, there is no 'one size fits all' algorithm within proposed techniques to improve fairness. Each should be tested in context and chosen based on data-set characteristics, implementation limitation and the stakeholders values.

Each of these three approaches have different strengths and weaknesses and in general requires information about the sensitive group within either training or test time - this can be seen as an obstacle within practical implications.

4.7.1 Pre-Processing

This category of implementation adjusts the feature space by transforming the training data. Commonly seen techniques create a representation of the data so that possible underlying sources of discrimination is removed or so that the feature space becomes uncorrelated with the sensitive attribute. Thus, we can use any machine learning technique with the assumption that all downstream discrimination is either justified or not present at all. If the *pre-processing* step satisfies any of the group criteria, then any deterministic training process on the new feature space will also satisfy the group criteria [3]. Algorithms that have been proposed using pre-processing techniques include *'learning fair representations'* that find a latent representation of the data that maps the data to a feature space where the protected attribute is obfuscated. A technique called *'optimized pre-pre-processing'* learns probabilistic transformations that edit the features and labels within the data whilst preserving as much information as possible. *Reweighing* is a processing technique that weights the examples in each group to ensure fairness [5]. Among these techniques, different options may be preferable depending on the problem case and whether or not the decision maker is allowed to change the values, ground truth labels or features within the data-set and which fairness metric the context requires.

A drawback to consider when using pre-processing is the ability to quantify and anticipate all the residual and causal effects of changing the feature space and specific interactions of changing feature values.

4.7.2 In-Processing

In-processing involves implementing the fairness constraint during training time by modifying the machine learning algorithm to explicitly remove discrimination. This can be done by changing the objective function directly or implementing a constraint. This can often lead to the highest utility as we are allowing the optimization of the model with consideration of the constraint. Algorithms that have been proposed include using a 'GridSearch'[1] approach that deterministically searches through a grid of trained classifiers with varying weightings and training parameters in order to find the model with the optimal accuracy-fairness trade-off. Another approach by [38] trains classifiers to maximise accuracy subject to disparate impact law that uses a regularization term that penalizes discrimination by measuring (un)fairness as the co-variance between the protected attribute and the distance of the features to the decision boundary.

As with all techniques, in-processing also has its drawbacks and obstacles to over come. Usually the loss function with fairness constraints can often be non-convex which means we may have to use context and implementation specific surrogate functions to find local optimums. A consequence of this is that some of the proposed algorithms are not able to be generalized across all fairness constraints and are specific to certain types of machine learning algorithms. This also often leads to undesirable outcomes that lead to either large trade-offs in accuracy or large costs to other forms of unfairness [36]

4.7.3 Post Processing

Post-processing involves adjusting the predictions of an already trained model that is optimized solely for accuracy. This technique often treats the model as a 'black box' where the training data and the optimization algorithm is fixed. [18] uses the approach of 'random' reassignment of class labels for the base predictions in order to satisfy certain fairness constraints. The technique assigns each predicted label a probability to change value. This probability can differ between the groups in order to minimize the cost of false positive or false negatives between groups. [31] offers approach to satisfy Calibration in conjunction with either one of the relaxations to sufficiency.

The advantage of post-processing techniques is that they are generalizable to any base classifier we choose to use - regardless of the technical implementations itself and does not require re-training. This is often the most practical use case as we may often not have access to the underlying decision maker and the training process. However, as a consequence we may inadvertently cause a significant cost of utility as we are reassigning the predicted labels that have been optimised for accuracy - attributed to the lack of flexibility within the accuracy-fairness trade-off. This can often be amplified when the inherent unfairness is large, thus more classification errors occur as more predicted labels will be switched in order to satisfy the fairness constraint. Thus, leading to all groups being equally 'worse off' [6].

4.8 Case Studies of Discriminatory Behaviour in Automated Decision Making

4.8.1 Assessing Candidates for College Admissions

A recent paper [39] examined a data-set from a large public University for assessing college schools admissions in order try and find a method to increase the amount of students that were from low-income neighbourhoods(¡$55,000). The results of this paper showed that choosing the right objective using a constraint optimization objective to admit students not only maintained academic requirements but satisfied the diversity criteria it was aiming for in comparison to unconstrained optimization that solely used(a mixture) of GPA or SAT scores to rank students. The prominent individual features used were high school GPA, SAT score, Composite GPA-SAT score, First-year college GPA and Zip-Code income.The data itself showed a positive linearity between an increase in income and GPA/SAT scores.

They compared their Constrained Optimization model(using Linear Programming), where the goal was to admit 25% of students from low-income zip-codes, to three other models that ranked students based on either GPA, SAT or Composite GPA-SAT.

The results showed that the models trained using unconstrained goals/targets to evaluate students created a large disparity between incomes of accepted and non-accepted students. This is a clear indication of *measurement bias* and the inherent discrimination introduced by the inappropriate choice of target variable. In comparison, the fairness-constrained or debiased goal using constrained optimization not only had higher representation of low-income students but also resulted in candidates with slightly higher average first-year college GPA scores. Another interesting finding is that these students had slightly lower average high school grades and substantially lower test scores showing the importance of choosing a target/goal that most appropriately represents the true intentions of an institution such as a University.

4.8.2 Unfairness in the Criminal Justice System and Predicting Recidivism

A commonly used case study for evaluating the effects of different fairness constraints on decision making is within the Criminal Justice System - more specifically a *ProPublica*[19] study of the COMPAS prediction tool that attempts to assign individuals a recidivism risk score based on various features such as arrest rates.

The disparate impact found in this study can be attributed to measurement bias and evaluation via inappropriate choice of proxies in feature selection to predict the target variable used to measure 'risk' and also the initial method of evaluating if the 'risk' score was fair. For example, prior arrests and arrests from close were used as proxy variables to measure the level of 'riskiness which can be considered as a false correlation. The use of prior arrests cannot conclude if an individual is at higher risk of causing a crime; the higher arrest rate was actually indicative of minority communities are policed more frequently [34].

The study found that the risk assessment system was "*almost twice as likely to mislabel a black defendant as a future risk than a white defendant.*" where the false positive rate was significantly higher for black defendants and the false negative rate for white defendants was also significantly higher compared to black defendants.([10], [6], [15]). Consequentially, white defendants were more likely to be mis-classified as low risk and black defendants were more likely to be mis-classified as high risk. This clearly violates equalized odds even when well-calibration between groups was satisfied[10] - relating us back to the Impossibility Theorem in Section (4.5).

4.8.3 Employment Decisions

Automated decisions plays a large and growing role within the employment and recruitment industry. Algorithms are being created to rate applicants in order to automatically identify potential candidates within the hiring process[30]. The intention for algorithms within hiring decisions is to remove discrimination or unconscious biases within human judgment and create objective measure of candidates. However, as with any system, the humans who create the system may impress their assumptions onto the model or the training data of the model contains systemic and historic biases that were created from past human hiring decisions[4] which has the potential to be perpetuated.

Such an example would be hiring decisions that consider academic credentials[25]. The models used were shown to assign a disproportionate weight to the reputation of the college or university from which an applicant has graduated. This can clearly be seen as measurement bias when an inappropriate proxy is used to falsely measure the competency of the candidate to the role - especially if the course isn't actually relevant to the job title itself.

An example of historical and target value bias may be applied when companies prioritise 'culture fit'[30]. Past decisions that were made to satisfy 'culture fit' where the workplace was majority white men might inadvertently recommend hiring more white men because the scored best for this goal/objective. This emphasises the importance of choosing the right (combination) of hiring goals.

4.9 Limitations of Observational Fairness Metrics

Group Fairness criteria constraints assumes that given a joint distribution $(\hat{Y}, Y, X \cup S)$ we should treat all groups $s \in S$ the same using passive statistical/observational statements of outcomes. Observational fairness metrics are often intuitive and appealing as they are easy to reason and are always measurable given samples from the joint distribution.

Figure 4.7: Two Identical Distributions $(\hat{Y}, Y, S)$ with Different Causal Graphs

However, they do not provide any insight into the 'why' or how decisions are made. They fundamentally lack *interpret-ability and explain-ability* of model intent and the features that influence the decision making process. For example, we can take two different scenarios that map to the same joint distribution but would intuitively and morally require the decision making process to lead to completely different interpretations on how to evaluate fairness given outcomes from both scenarios. In this case, our observational fairness statements would fail to measure the difference between both scenarios as they would both be equally fair/unfair. An example of this is shown in Figure 4.7 which is inspired from [3]. Given X_i are features, S is the protected attribute and $(\tilde{Y}, \hat{Y})$ are the optimal unconstrained score and optimal separated score respectively. We see $(\hat{Y}, Y, S)$ are identical distributions even with completely different causal graphs.

4.10 Related Work and Alternative Fairness Definitions

Within this section we briefly explore other possible methods for defining fairness and proposed methods that attempt to address some of the shortcomings of observational criteria. Appendix E illustrates the distribution of work done in past literature to formalise various fairness definitions within machine learning applications.

4.10.1 Individual Fairness

We describe in Section 4.4 the perspective of 'individual fairness' proposed by [14] which takes the view that 'similar people should be treated similarly' as an alternative to measuring fairness as equality between groups. Within this framework, it is assumed that there is an 'ideal' feature space to compute

inherent similarity between individuals and that "*determining whether two individuals are similar with respect to the task is critical, and assume that such a metric is given to them.*" Understanding this, we should be able to retrieve and maintain similarity of individuals within the observed data by applying domain/task specific measurements of (dis)similarity. This fairness constraint enforces a*Lipschitz property* on the classifier that finds a mapping from individuals over outcomes that minimizes expected loss conditions on the *Lipschitz* property. Intuitively, this condition defines a task-specific distance measure between two individuals x, x' as $d(x_i, x_j) \in [0, 1]$ such that when the individuals are mapped over to some new domain $M : X \to \Delta(O)$ we get similar distributions over outcomes.

The differences between the individuals within the new domain should be similar to that in the original domain. $D(M(x_i), M(x_j)) \leqslant d(x_i, x_j)$.

Formally, the optimisation problem is defined as: $opt(I) = min_{\mu_x, x \in X} E_{x \sim V} E_{a \sim \mu_x} L(x, a)$ subject to the constraints $\forall x, y \in V : D(\mu_x, \mu_y) \leqslant d(x, y), \forall x \in V : \mu_x \in \Delta(A)$ Where $L : X \times S \to R$ is an arbitrary loss function that achieves (D, d)-Lipschitz property for a metric $d : X \times X \to R$. I is an instance of our problem where our goal is to find the value of our optimal minimization problem $opt(I)$ with the mapping $M : X \to \Delta(S), M = \mu_{x, x \in V}, \mu_x = M(x) \in \Delta(A)$.

A similar notion is stated in [16] where the differences in individuals defined in the 'construct space' that have similar inherent attributes (such as intelligence, risk-adverseness and grit) should carry over to an 'observed space' that maps inherent attributes to observable measures (such as IQ Tests, Age and Duckworth Grit Scale). Ultimately, these individuals should also be treated similarly in our domain specific outcome distribution.

However, a natural criticism of this perspective is that we have to assume we have purely objective measures to measure (dis)similarity between individuals within each domain. Thus, the responsibility relies on the decision maker to define unambiguous metrics of difference which can be possible avenues of bias and discrimination.

4.10.2 Causal Inference

Causal reasoning techniques, defined by causal models and graphs(often known as Bayesian Networks), have been proposed in order to solve fairness related issues in algorithms such as understanding 'why' decisions are made, thus allowing us to evaluate if a decision was fair or not. ([22], [20], [14]). We define a directed acyclic graph where each node represents observed attributes of an individual and edges represent causal relationships between features that encode conditional independence statements. Using the relationships within the graph, structural equations can be defined to demonstrate the influence of attributes onto the outcome i.e. the predicted scores/classifications. The *causal effect* of an event $X = x$ on a variable Y is the distribution of Y in the model where this event occurs: $E_{Model[X=x]}[Y]$. It is often the case in binary features that we are interested in the value of $E_{Model[X=1]}[Y] - E_{Model[X=0]}[Y]$ as a measure of differential treatment.

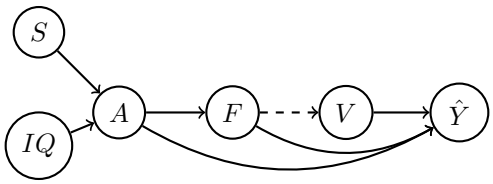

Figure 4.8: Causal Graph of Simulated University Admissions. Possible Decision Paths Corresponding to the Three Target Attributes/Goals.

We use the following labels as fairness definitions within causal models:

Proxy Discrimination

In a causal model, we define a *proxy* node/feature such that its value can be used to derive the value of another feature. Therefore, we define a causal model to potentially contain *proxy discrimination* if there exists an indirect path from the sensitive feature S to the predicted outcome $\hat{Y}$ such that there is a intermediate proxy variable X_i in-between that 'blocks' the path from S to $\hat{Y}$. In Figure 4.8, nodes

A and F representing A Level and Final Scores would be considered *proxy* features and thus this graph would contain proxy discrimination.

Unresolved Discrimination

In a causal model, we define a *resolving feature* X_i as a feature that is dependent on the sensitive attribute S in a non-discriminate way. This means the differences of the values of the *resolver feature* between different groups are considered non-discriminatory. Therefore, *unresolved discrimination* occurs if there is a directed path from the sensitive attribute to the prediction that does not contain an intermediary *resolver* node/feature. Thus the prediction $\hat{Y}$ is considered un-resolved. In Figure 4.8 we see that the node V representing the value added metric would act as a *resolver* feature and thus the decision path that involves S and V would not contain *un-resolved* discrimination. However, this graph does contain *un-resolved discrimination* as we can take a decision path from S that uses the unresolved features A, F.

Counterfactual Fairness

Counterfactual questions allow us to evaluate more than just observational outcomes. Using causal models, counterfactual's retrospectively ask how and outcome would change based on a change in a decision made in the decision making process. Counterfactual's provide a quantitative methodology to evaluate causal effects within a specific path of a causal graph; more abstractly it allows us to define 'fair' and 'equitable' decisions within a process that provides a desired outcome. We use the question of *'what would've happened'* if an individual had a different value for one of their attributes to define *Counterfactual fairness.* Thus a decision is counter-factually fair if, in both worlds, the same decision was made even if an individual changed their sensitive attribute/group. Intuitively this means that individuals with the same features but with different values of sensitive attributes should receive similar outcomes and is why counterfactual fairness is commonly grouped with individual fairness which we briefly summarised in Section 4.10.1.
Formally we define a classifier as being counter-factually fair if *"under any context* $\forall X = x, S = s$: $P(\hat{Y}_{S\leftarrow s}(U) = y | X = x, S = s) = P(\hat{Y}_{S\leftarrow s'}(U) | X = x, S = s)$. [22]. [3] proposes *interventions* using *do-calculus* that allows us to define and evaluate counterfactual's within a causal model by changing a certain value $S = s$ within all structural equations.

We can measure the *average-causal effect* of a substitution as $E[\hat{Y}|do(S = a), X = x] - E[\hat{Y}|do(S = b), X = x]$. Using this notion we can define *counterfactual demographic parity* such that every *do* assignment, the outcome is independent of the substituted variable S.

[3] also explores other methods to create analogous causal definitions to the observational group fairness criteria - thus connecting individual and group fairness notations.

We should note that evaluating a model for counterfactual fairness is not simply just flipping the value of an attribute - this does not generally generate valid counterfactual's. We know that within a causal graph, some features will be influenced by the flipped sensitive attribute which requires some change in other attributes down the path unless race does not have any descendant nodes i.e. all other features are independent of race - not a realistic scenario. This introduces the problem of understanding how variables interact - in real world scenarios, it is either intractable or simply too abstract to be able to evaluate each change downstream without full knowledge.

Fair Inference

Fair inference describes the legitimacy of a certain path taken to make a decision. Thus a causal graph satisfies fair inference if there are no illegitimate paths from the sensitive attribute to the decision[35]. This is ultimately reduces down to a moral and philosophical question as the decision maker themselves has to decide what a 'legitimate' path might look like and if some features in a path that may be a proxies for the sensitive attribute should be considered 'legitimate' or not. For example, a merit based world view may view A-Levels as an 'legitimate' target goal or an unbiased feature whereas as Group Fairness definitions may suggest otherwise.

Challenges of Causal Inference

The largest obstacle for using causal models is that the decision maker has to define a set of assumptions that are truthfully reflect the real world. The creator must have full quantitative understanding of how each feature is related to each other and is able to exactly infer a causal understanding of how certain (change in) variables may impact an individual. This is often not a reasonable assumption to make due to the complexity of social constructs and the lack of stable ontological and agreed upon structures in some domains [3]. This puts into question the feasibility of creating and identifying valid causal structures especially when the validity may be subjective to constantly changing social, human and scientific scrutiny. Ultimately, causal models are constrained by a set of bounded features and relationships within a specific fixed relational environment - where in reality these models will often contain various amounts of unaccounted noise and are constantly changing. Therefore what we can learn from a causal model is strictly bounded to the abstraction we choose to represent reality with - which can either to be too broad or too specific - each case leading to other possible avenues of bias and discrimination.

Chapter 5

Implementation

5.1 Data Generation

We use a simple simulated data-set for University Admissions criteria to illustrate our proposal. We sample $n = 10000$ total observations from the generation process and leave 70% for training and the other 30% to evaluate our models against.

$$
\begin{aligned}
\text{iq}_i &\sim \text{Normal}(5, 1) \\
\text{school\_system}_i &\sim \text{Bernoulli}(0.5) \\
\text{a\_level}_i &= \text{iq}_i + \text{school\_system}_i + \text{Normal}(0, 0.5) \\
\text{final\_score}_i &= \text{a\_level}_i + \text{iq}_i + \text{Normal}(0, 0.5) \\
\text{value\_added}_i &= \text{final\_score}_i - \text{a\_level}_i \\
\text{work\_experience}_i &\sim \text{Normal}(2 + 1.5 \cdot \text{school\_system}_i, 0.2) + \text{Normal}(2, 0.5) \\
\text{parent\_income}_i &\sim \text{Poisson}(20 + 10 \cdot \text{school\_system}_i) + \text{Normal}(50, 0.25) \\
\text{target\_a\_level}_i &= \begin{cases} 1 \text{ If a\_level}_i \text{ is within the top 25\% of all a level scores} \\ 0 \text{ Otherwise} \end{cases} \\
\text{target\_final\_score}_i &= \begin{cases} 1 \text{ If final\_score}_i \text{ is within the top 25\% of all final scores} \\ 0 \text{ Otherwise} \end{cases} \\
\text{target\_value\_added}_i &= \begin{cases} 1 \text{ If value\_added}_i \text{ is within the top 25\% of all value added scores} \\ 0 \text{ Otherwise} \end{cases}
\end{aligned}
$$

This data-generating process has the following properties:

1. The binary school system value represents what type of education the individual participated in state(0) and independent(1). This is also our protected/sensitive attribute.
2. A level scores are linearly dependent on the school system they attended to.
3. Final scores represent the individuals score after completing University. It is directly dependent on the individuals A level score and by proxy the school system they went to.
4. An individuals value added score is their final score minus their a level score. Thus $\text{final\_score}_i - \text{a\_level}_i = \text{iq}_i + \text{noise}_{\text{a\_level}} + \text{noise}_{\text{final\_score}}$. This means the final score is solely (linearly) dependent on the pupils base intelligence (IQ).
5. We use an acceptance rate of 25% to convert real valued scores to binary values $\in \{0, 1\}$ where accepted = 1, not accepted = 0.
6. Given that A Level Scores and Final Scores are dependent on an individual's school system, which is a source of historical bias, we would expect these two metrics to have inherent measurement/target variable bias if used as the target to predict if an individual would be admitted or not. Thus they would be 'unfair' goals.
7. 'Value Added' can be considered a 'fair' goal in a societal context as it is solely measures an applicants base intelligence and thus would be an unbiased way for predicting an individuals likelihood for admission.

8. We add noise to each of the following features; A Levels, Final Scores, Parent Income and Work Experience. We expect Final Score to be less correlated with the school system in comparison to A levels due to adding an additional layer of noise.
9. School System has a high weighting for calculating work experience meaning there should be a clear separation in distributions between groups.
10. Parent income is also calculated using a high weighting on the School System. However, we add a relatively large amount of noise which is why expect there to be a significant overlap of the distributions between groups.
11. During training and testing, we standardise the features due to the differences in scale.

5.1.1 Training Data for Experiment 1

This experiment uses an explicit encoding of the sensitive feature - in this case the school system - alongside the A-level score as features to train, predict and classify acceptance of an individual.

$$X_i = [\text{a\_level}_i, \text{school\_system}_i]$$

5.1.2 Training Data for Experiment 2

This experiment uses an implicit encoding of the sensitive feature using proxy variables(work_experience, parent_income that are highly correlated with the sensitive feature, alongside the A-level score as features to train, predict and classify acceptance of an individual.

$$X_i = [\text{a\_level}_i, \text{work\_experience}_i, \text{parent\_income}_i]$$

5.1.3 Training Data for Experiment 3

This experiment uses all features as predictors to train, predict and classify acceptance of an individual.

$$X_i = [\text{a\_level}_i, \text{work\_experience}_i, \text{parent\_income}_i, \text{a\_level}_i, \text{school\_system}_i]$$

Within each experiment, and individual has a set of three possible ground truth labels that are used dependent on which of the three classifiers, representing three different goals, are used for prediction:

$$Y_i = \{\text{target\_a\_level}_i, \text{target\_final\_score}_i, \text{target\_value\_added}_i\}$$

5.2 Data Analysis

Within this section we provide insight into our simulated data-set in the scenario when we assume the distribution of the population educated from independent and state schools are drawn from a Bernoulli(0.5). This gives us a roughly equal distribution of people from independent and state school.

5.2.1 A Level Distributions

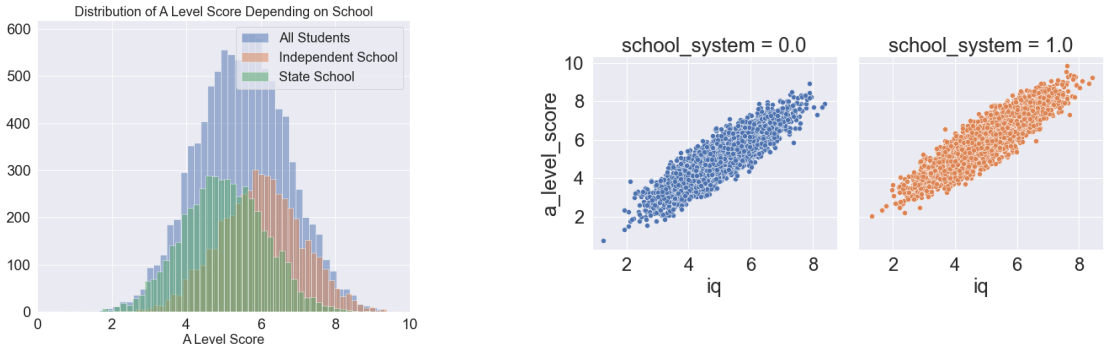

Figure 5.1: Left: Histogram of A Level Scores(All, Independent, State), Right: Scatter plot of IQ vs A level score for both groups

A level scores are linearly dependent on an individuals base intelligence. However, we also factor in an additional constant based the individuals school system(+0 for state educated applicants, +1 for independently educated applicants). Thus, we expect that people who were educated in independent schools to have higher a level scores on average. Given how we generate our data-set, we expect there to be a linear relationship with an individuals IQ and their A level score. However, we expect students from independent schools to have on average +1 A level score to individuals educated from state schools even with matching base intelligence values (IQ). We see this in Figure 5.1

5.2.2 Final Score Distributions

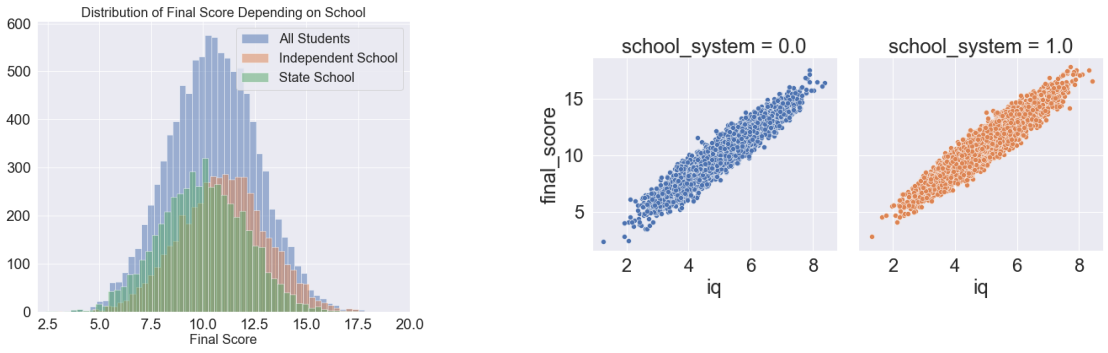

Figure 5.2: Left: Histogram of Final Scores(All, Independent, State), Right: Scatter plot of IQ vs Final score for both groups

Given the direct dependence of an individuals Final score and their A level scores we expect, by proxy/transitive properties, that the final score also has the same relationship/correlation with the individuals school system. Thus we expect that people who were educated in independent schools to have higher final graduation scores on average. However, since we add additional noise when calculating the final score, we expect the separation to be less pronounced than the difference in A level distributions between school systems. This can represent in reality how University causes a regression to the mean as individuals, no matter the school, will be taught in roughly the same manner as they have attended University. Given

how we generate our data-set, we expect there to be a linear relationship with an individuals IQ and their Final score with the expectation that students from independent schools to have on average higher Final score to individuals educated from state schools even with matching base intelligence values (IQ). Figure 5.2 clearly represents our expectations.

5.2.3 Value Added Distributions

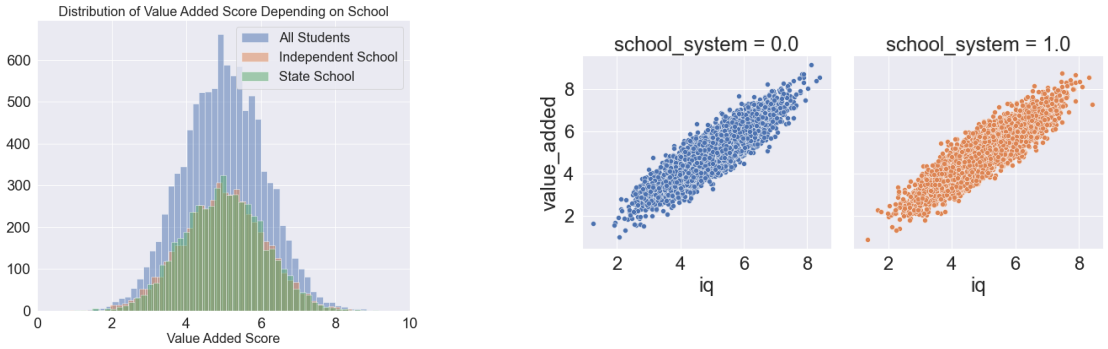

Figure 5.3: Left: Histogram of Value Added Scores(All, Independent, State), Right: Scatter plot of IQ vs Value Added Scores for both groups

As we calculate the value added score for and individual as their final score minus their a level score, the value added score removes any dependence/doesn't factor their school system. Thus we expect value added scores to be (solely) linearly dependent on the individuals base intelligence. Given how we generate our data-set, we expect there to be a linear relationship with an individuals IQ and their Value Added score. We expect the value added score to be roughly equivalent for individuals with matching base intelligence scores from both school systems. We see this in Figure 5.3.

5.2.4 Work Experience and Parent Income Distributions

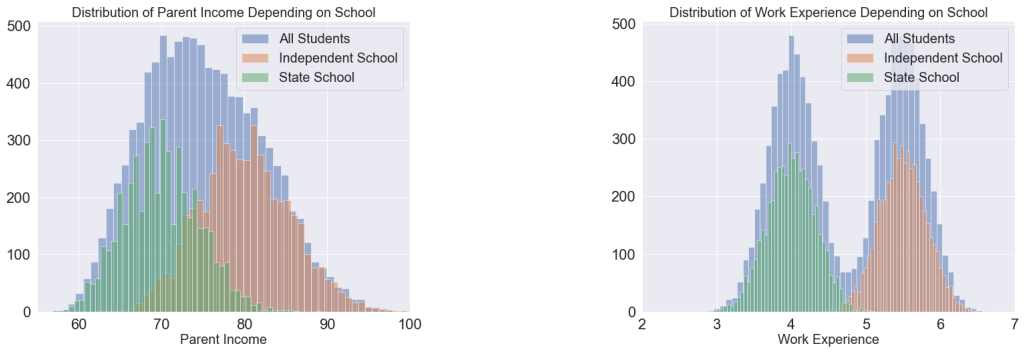

Figure 5.4: Acceptance Rate Between Groups Using Value Added Scores as Acceptance Criteria

Both of these features are used and treated as 'proxy' variables to the sensitive attribute and thus the school system. This requirement means that it implicitly encodes information about the school system. Figure 5.4 clearly shows how the distribution of both features(particularly work experience) are separated dependent on privileged and un-privileged groups. We expect during the the training process, a model will implicitly infer the school school system given the value of both the proxy features thus predict similar results for each label as if the sensitive attribute was explicitly included within the training features.

5.2.5 Base Rates of Acceptance

Figure 5.5 illustrates the base rates of acceptance/rejection between State and Independently educated groups given our three different goals/target labels to judge admission.

If a University Admissions board solely considered an individuals A level or Final score as their metric to measure worthiness of acceptance, we will expect that the rate of individuals accepted from Independent schools will be higher than State schools given that individuals from Independent schools will have on average higher scores for both of the mentioned admissions criteria. However, when an individuals Value Added score is solely considered as the admissions metric to measure worthiness of acceptance, we see that the rate of individuals accepted from Independent schools will be equal to individuals from State schools. This means the ***base rate*** of acceptance between groups is roughly equivalent, thus satisfying a condition that can be used to mitigate the Impossibility theorem in Section4.5 allowing us to satisfy all three fairness constraints at once.

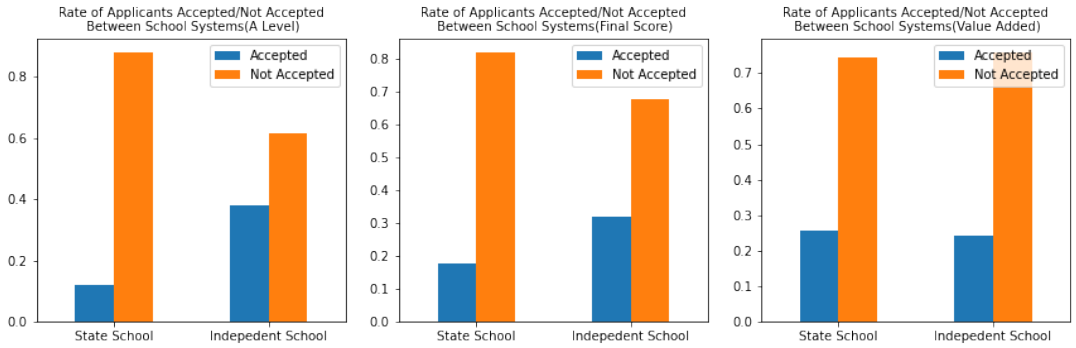

Figure 5.5: Comparing Base Rates of Acceptance Between Groups of Different Goals

5.2.6 Correlation Between All Features

Figure 5.6 shows the overall correlation of our features and possible metrics (target_a(A Levels), target_f(Final Score), target_v(Value Added)) to use as targets/admissions criteria. As expected, we A Level and Final scores to be positively correlated with an individuals school system they were educated in. Given that the school system is either 0 or 1, these scores are likely to increase if an individual is educated in an independent school. However, the Value Added score has close to 0 correlation with the school system, which is expected from a 'fair' metric that is by definition independent from the school system. All three metrics are heavily correlated with an individuals intelligence score which is expected. The reason they are not closer to 1 is because of the additional noise we added when generating these metrics. Each binary target value is also unexpectedly positively correlated with its real counterpart.

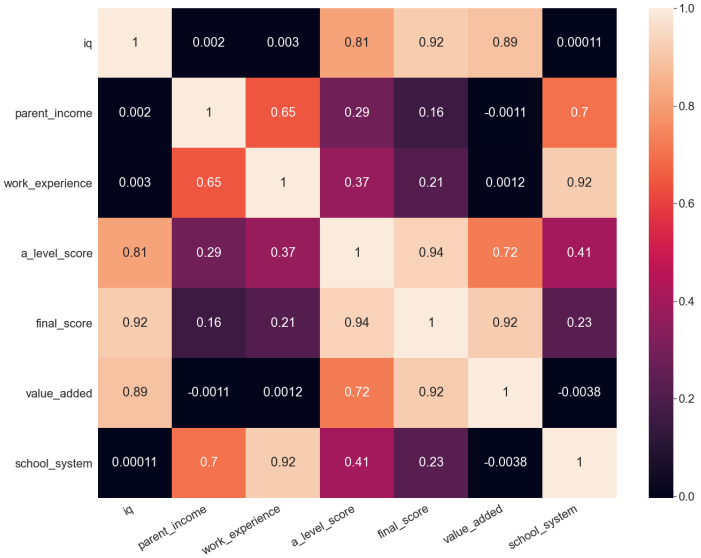

Figure 5.6: Correlation Heat Map Between Features.

5.2.7 Changing the Distribution of School System Amongst the Population

Appendix A provides similar visualisations for when we change our distribution of the sensitive attribute to Bernoulli(0.9) - corresponding to a more realistic representation of the population whom is educated from state and independent schools. In this case, $\sim 90\%$ of students are educated from state schools with $\sim 10\%$ from independent schools. Figure A.1 and A.2 demonstrate that the overall distribution clearly reflects the state school distribution - this can lead to Simpson's paradox and representation bias as we have less to data on the group trained within Independent schools. We thus assume that classifiers will be less precise and more error prone specifically when trained to predict this class simply because the model has less data to train on but also the change in base rates of positive outcomes between groups. This can lead to further discrepancy's when trying to generate fair outcomes between groups whilst maintaining an appropriate accuracy-fairness trade-off. The implications of this distribution change is discussed within the experiment results in Section 6.5.

5.3 Classification Models

5.3.1 Logistic Regression and Classification

Classification Problems

Before we go into detail describing proposed methods to formally implement fairness constraints within machine learning models, we must first understand the methodology and intuition behind any predictor for any binary classification task. A classification problem aims to learn a optimal classifier $f : X \to Y$ that maps feature vectors X to class labels $Y \in \{0, 1\}$ with the goal of minimising some classification loss function $L(\theta)$ such that $\theta_{opt} = \arg\min_\theta L(\theta)$ where θ is our model parameters or feature weights.

Within *'margin based'* classifiers, θ_{opt} can be seen as an optimal decision boundary within our feature space that is computed on a training data-set $X_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N_{train}}$ [38]. This decision boundary will then output predictions given a test-set $X_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{N_{test}}$ that is not seen by the model within the training process. The model predicts $f_\theta(x_i) = 1 = \hat{y}_i$ if the decision function $d_{\theta_{opt}}(x_i) \geqslant 0$ else $f_\theta(x_i) = 0$ where $d_{\theta_{opt}}$ represents the distance from the feature vector/data-point x_i from the decision boundary.

Logistic Regression

The standard Logistic Regression classifier represents our 'unconstrained model' where the sole purpose is to optimise for accuracy. However, this model also forms the basis of our fairness constrained classifiers that we use within this thesis. We optimise these 'base' classifiers for fairness either during training time or via post processing. We describe the details of these within Section 4.7.

It is often the case the directly solving the optimization problem for empirical loss: $\min_\theta L(\theta) = \min_\theta E_{X,Y}[1_{f_\theta(x)=y}]$ is intractable as the objective function is non convex. We thus often use surrogate functions that allow us to optimise θ by making the loss function convex.

Logistic Regression is a widely used classification model that allows us to take continuous feature vectors and map them as binary outcomes. It uses the *sigmoid* function to map the feature vector to the real domain and the *log* function as our surrogate decision function/indicator variable that makes the objective function convex. The *sigmoid* function maps each feature vector to the posterior likelihood of classifying that feature vector as 1:

$$h_\theta(x_i) = p(y_i = 1|x_i, \theta) = \frac{1}{1 + e^{-\theta^T x_i}}$$

Thus the *Log Loss* Function can be defined as:

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(h_\theta(x_i)) + (1 - y_i)\log(1 - h_\theta(X_i))$$

We then optimise for θ using *gradient descent* which is a technique that iteratively traverses down the loss curve to find the optimal model parameters that represents the local minimum of our loss function $J(\theta)$. We traverse the loss curve by calculating the current gradient at the point on the curve and move along the loss curve in the negative direction of the gradient at the current point.

$$\frac{\delta}{\delta\theta_j} J(\theta) = \sum_{-=1}^{m} (h_\theta(x_i) - y_i)x_{ij}$$

As the gradient decreases, this means we are iteratively getting closer to the local minimum of our loss-function. Thus, when the gradient reaches approximately close to 0, the next iteration will cause equivalently no movement along the loss curve and we stop iterating. After the last iteration, our algorithm will give us the optimal values for θ for which our cost function J is at a minimum.

Initialize parameters θ, ν, ϵ
while true do
  $\theta[t] = \theta[t-1] - \alpha \cdot \delta(\theta[t-1])$
  if $\delta(\theta[t]) < \epsilon$ **then**
    break
  else
  end
end

Algorithm 5.1: Vanilla Gradient Descent

The standard gradient descent algorithm is defined in Algorithm 5.1(however, we should note there are more efficient algorithms such as *stochastic gradient descent*). α is the pre-determined *learning rate* that defines how fast we traverse down the loss curve,ϵ is our convergence condition, θ are our model parameters/feature weights at iteration t and δ is the function to calculate the gradient of our loss function at the previous parameter setting.

5.4 Implementation of Fairness Constrained Classification Models

Within our experiments we use a *post-processing* techniques(See Section 4.7.3 proposed by [18] and [31] to implement each of three fairness constraints(*Independence, sufficiency and calibration*) onto a pre-trained Linear Regression model that is solely optimised for accuracy.

5.4.1 Separation Constraint

We take direct inspiration from [18] to derive a fair predictor $\tilde{Y}$ that satisfies *Equalized Odds* and thus *Separation* from predictions of an unconstrained binary classifier $\hat{Y}$ using a post-processing technique that does not require us to change the data or training process in any way. The definition of this derived predictor means that it only depends on the predicted outcomes of our unconstrained model $\tilde{Y}$ and the sensitive feature S. This means during construction of $\tilde{Y}$, the predictor solely requires the information from the joint distribution $(\hat{Y}, S, Y)$ and at prediction time only has access to $(\hat{Y}, S)$ with the possibility of introducing additional randomness. We should note that we can achieve separation is trivial by using a predictor that only predicts a single value(in this case only 0 or only 1). However, this is intuitively not going to be an optimal predictor and has a huge accuracy-fairness trade-off in most realistic cases.

Definition 5.1 *A derived predictor $\tilde{Y}$ is a possible randomized function of the random variables $(\hat{Y}, S)$ such that $\tilde{Y}$ is independent of X conditional on $(\hat{Y}, S)$. We consider a loss function $l : \{0,1\} \times \{0,1\} \rightarrow R$ that returns a real value that indicates the cost of our prediction in comparison to the ground truth label. Therefore we design a derived predictor $\tilde{Y}$ that minimises the expected loss $E[l(\tilde{Y}, Y)]$ constrained to one of our fairness metrics.*

Intuitively, the randomness introduced into the derived classifier $\tilde{Y}$ is from drawing Bernoulli random variables, indicating probabilities of flipping of some of the predictions of the original classifier $\hat{Y}$ for each group from positive to negative and vice-versa. A geometric solution is used to show that achieving each notion of separation can be solved/optimized as a linear program with four variables. We define the tuple $\gamma(\hat{Y})$ calculated from the joint distribution $(\hat{Y}, S, Y)$:

$$\gamma_a(\hat{Y}) = (Pr\{\hat{Y} = 1 | S = a, Y = 0\}, Pr\{\hat{Y} = 1 | S = a, Y = 1\})$$

Lemma 5.4.1 *This first component is the false positive rate and the second component is the true positive rate for the group $S = a$. This tuple γ can be interpreted as an expression of 'Equalized Odds' and each component defines 'Predictive Equality' and 'Equal Opportunity' respectively.*

1. *Equalized Odds* $\iff \gamma_a(\hat{Y}) = \gamma_b(\hat{Y})$.
2. *Predictive Equality* $\iff \gamma_{a,1}(\hat{Y}) = \gamma_{b,1}(\hat{Y})$ *i.e. they agree in the first component only.*
3. *Equal Opportunity* $\iff \gamma_{a,2}(\hat{Y}) = \gamma_{b,2}(\hat{Y})$ *i.e. they agree in the second component only.*

Lemma 5.4.1 allows us to define a two two-dimensional convex polytope: $P_0(\hat{Y}), P_1(\hat{Y})$ that can be used to demonstrate the trade-offs of false positive and true positive rates between groups. and thus defines the achievable region for classifiers satisfying *Separation* (See Figure 4.2). We can also prove(Proof 5.1 that $\hat{Y}$ is a derived predictor $\iff \forall a \in \{0,1\} : \gamma_a(\tilde{Y}) \in P_a(\hat{Y})$

Definition 5.2 *We define a convex hull $P(\hat{Y})$ of four vertices where $\{a,b\} = \{0,1\}$ are the binary classifications of each group: $P_a(\hat{Y}) = convhull\{(0,0), \gamma_a(\hat{Y}), \gamma_a(1-\hat{Y}), (1,1)\}$*

Proof 5.1 *$\hat{Y}$ is a derived predictor $\iff \forall a \in \{0,1\} : \gamma_a(\tilde{Y}) \in P_a(\hat{Y})$*
Since a derived predictor $\tilde{Y}$ solely depends on $(\hat{Y}, A)$ which are both binary random variables, then the feasible region containing $\tilde{Y}$ is completely described by four parameters in $[0,1]$ that correspond to the true positive and false positive rates for each group: $Pr\{\tilde{Y} = 1 | \hat{Y} = y, S = a\} \forall y, a \in \{0,1\}$. Each parameter choice thus defines the convex hull $P_a(\hat{Y})$ and thus the achievable region for each group where every point in the convex hull can be achieved by a specific parameter setting.

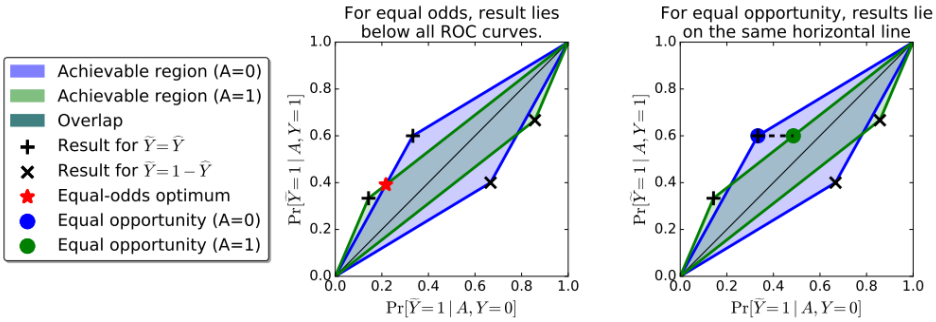

Figure 5.7: Convex Hulls Demonstrating Achievable Regions for Separation Satisfying Classifiers ($A \equiv S$) [18]

For simplicity, we visualise this using a simulated example taken from [18] in Figure 5.7 where the intersection of the the convex hull for both groups is the achievable region for satisfying *separation*.

We note that for *equalized odds* to be satisfied, the classifier must lie directly on a point of intersection between the convex hull P_a for both groups. Whereas for equal opportunity and predictive equality, we only need to lie on either a horizontal(matching true positives) or vertical line(matching false positives) connecting between the two respectively.Combining the above, we can derive the following optimization problem to derive the predictor satisfying equalized odds:

$$\min_{\tilde{Y}} \mathbb{E}[l(\tilde{Y}, Y)] s.t. \begin{cases} \text{Derived Predictor} : \forall a \in \{0,1\} : \gamma_a(\tilde{Y}) \in P_a(\hat{Y}) \\ \text{Equalized Odds} : \gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y}) \end{cases}$$

This gives us a linear program in four variables whose coefficients can be computed from the joint distribution $(\hat{Y}, S, Y)$ which gives us an optimal solution to for equalized odds predictor $\tilde{Y}$ derived from $(\hat{Y}, S)$ (See Proof 5.2).

Proof 5.2 *We show the objective function is a linear function of the four parameters describing the feasible region of the derived predictors satisfying separation:*

$$\mathbb{E}[(\tilde{Y}, Y)] = \sum_{y,y' \in \{0,1\}} l(y,y') Pr\{\tilde{Y} = y', Y = y\}$$

$$Where : Pr\{\tilde{Y} = y', Y = y\}$$
$$= Pr\{\tilde{Y} = y', Y = y | \tilde{Y} = \hat{Y}\} \cdot Pr\{\tilde{Y} = \hat{Y}\} + Pr\{\tilde{Y} = y', Y = y | \tilde{Y} \neq \hat{Y}\} \cdot Pr\{\tilde{Y} \neq \hat{Y}\}$$
$$= Pr\{\tilde{Y} = y', Y = y | \tilde{Y} = \hat{Y}\} \cdot Pr\{\tilde{Y} = \hat{Y}\} + Pr\{\hat{Y} = 1 - y', Y = y\} \cdot Pr\{\tilde{Y} \neq \hat{Y}\}$$

Each of the probabilities that do not use $\tilde{Y}$ can be computed from the joint distribution $(\hat{Y}, S)$ and the probabilities that do are simply a linear function of the parameters that describe $\tilde{Y}$.

The above proof for equal opportunity and predictive equality is equivalent except for the relaxed constraint $\gamma_{0,2}(\tilde{Y}) = \gamma_{1,2}(\tilde{Y})$ and $\gamma_{0,1}(\tilde{Y}) = \gamma_{1,1}(\tilde{Y})$. We note that classifiers within the achievable region require randomization.

Proof 5.3 *Classifiers within the achievable region require randomization.*
As an example, we choose the two classifiers corresponding to the coordinates $(0,0),(1,1)$. The first classifier does not classify anyone with the label $\tilde{Y} = 1$ whereas the second classifiers everyone with the label with the label $\tilde{Y} = 1$. Given an instance of the problem we construct a third classifier that randomly picks and uses the result of the first classifier at $(0,0)$ with probability p and the second classifier with probability $(1,1)$ with probability $1-p$. This classifier achieves both true and false positive rates p that lies on the central lie within the plot. We can create this third classifier using can be any other pair of classifiers lying on the line representing the intersection of both ROC curves - thus filling in the shaded achievable region. Due to the shaded region being convex, every point within the region must lie on a line segment between two classifiers that lie on the boundary of the intersection of both ROC curves(See Figure 4.2 for an abstract example).

For performance of an equalized odds predictor is determined by the minimum performance among all groups, thus we do not consider any classifiers within the region below the diagonal from $(0,0)$ to $(1,1)$. This is because this region will be strictly worse than any classifier above the diagonal as we will have always maintain lower rates of true positives with the same false positive rate and thus not desirable for any optimal predictor that aims to optimise accuracy and fairness. This means our classifier is incentivised to build 'fair' classifiers for both classes.

Derived Real Valued Predictors Using Thresholds

Even though are problem case doesn't strictly demand it, we should note that the same post-processing method can be applied to classifying real valued scores instead of strictly binary labels. We will not go into detail to prove the equivalency but the idea is to apply different thresholds t_s for each group to generate a binary prediction: $\tilde{Y} = I\{R > t_s\}$. This does not automatically sufficient and requires additional randomization as within the binary example. We define a threshold analogous definition to γ_a as $C_a(t) = (Pr\{\hat{R} > t | S = a, Y = 0\}, Pr\{\hat{R} > t | S = a, Y = 1\})$ that allows us to define the equivalent conditional ROC curves that allow us to satisfy equalized odds $C_a(t) = C_b(t)$. In this case, we can find every classifier by changing the threshold rates so that both ROC curves intersect. We repeat the same randomization process for choosing thresholds for each class to fill in the achievable region where $D_a = \text{convhull}\{C_a(t) : t \in [0,1]\}$ is an analogous to the convex hull P_a in the binary case.

We should note that in this case, D_a is no longer a polytope and thus we thus the problem is no longer a linear program, however, we can efficiently optimize for this problem numerically using ternary search.

$$\min_{\forall a : \gamma \in D_a} \gamma_0 l(1,0) + (1-\gamma_1) l(0,1)$$

$$\text{assuming} l(0,0) = l(1,1) = 0$$

5.4.2 Independence Constraint

Implementing classifiers to achieve independence takes a similar randomized approach as implementing separation. However, in this case we require the constraint $Pr\{\tilde{Y} = 1 | S = a\} = Pr\{\tilde{Y} = 1 | S = b\} = Pr\{\tilde{Y} = 1\}$ such that our derived predictor $\tilde{Y}$ does not depend on the protected attribute and the positive rate between groups are the same. We optimise based on the accuracy-selection rate trade-off(instead of false-true positive rate trade-off) by probabilistic-ally flipping predictions to achieve classifiers within the intersection of achievable region for each group - as in Proof 5.3 - such that we achieve achieved balanced accuracy and equal rates of positive outcomes between groups:
We define the accuracy score for a given group $a \in S$ as:

$$acc_a(\hat{Y}, Y) = \frac{1}{n_{\text{samples}_a}} \sum_{i=0}^{n_{\text{samples}_a}} 1(\hat{y}_i = y_i)$$

As with separation we define $\gamma_a(\hat{Y}) = (Pr\{\hat{Y} = 1|S = a\}, acc_a(\hat{Y}, Y))$ and derive our convex hull P_a as before:

$$\min_{\tilde{Y}} \mathbb{E}[l(\tilde{Y}, Y)] s.t. \begin{cases} \text{Derived Predictor} : \forall a \in \{0,1\} : \gamma_a(\tilde{Y}) \in P_a(\hat{Y}) \\ \text{Demographic Parity and Balanced Accuracy} : \gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y}) \end{cases}$$

5.4.3 Sufficiency Constraint

As mentioned in section 4.4.3, we can achieve sufficiency by satisfying calibration which is a slightly stronger notion. In practise, techniques such as *Platt Scaling* have been used to implement a calibrated predictor. *Platt Scaling* takes an un-calibrated score and treats it as a feature. It then trains a single feature regression model using this feature as the target variable.

Definition 5.3 *Platt Scaling aims to find scalar parameters a, b such that for a score R: the sigmoid function: $S = \frac{1}{1+\exp(aR+b)}$ fits the target variable Y with respect to the log-loss function: $-\mathbb{E}[Y \log S + (1-Y)\log(1-S)]$ which can be minimized give labelled examples drawn from (R, Y)*

Within our experiments, we use the proposed implementation by [31] to find a classifier that is well-calibrated between groups and allows us to achieve one of the relaxations of separation simultaneously. This follows from Section 4.5.6 showing the geometric relationship between calibrated classifiers and achieving separation. [31]) notes that it is not a trivial task to encode calibration within the optimization problem themselves. Therefore we post-process already calibrated classifiers (to satisfy either *Equal Opportunity or Predictive Equality*).The geometric intuition and proof of certain constraints and assumptions to achieve relaxed equalized odds with calibration is described in Section 4.5.7. We give a general description of the algorithm below to implement these constraints.

Recall that we define binary classifiers for each group h_1, h_2 such that each classifier outputs a probability that a sample x belongs to the positive class. We assume that these classifiers are optimally calibrated but are possibly discriminatory. These calibrated classifiers can be learned in methods such as *Platt Scaling*.

Classifier h_s is perfectly calibrated if $\forall p \in [0,1] : P_{(x,y)}[y = 1|h_s(x) = p] = p$. Thus for sufficiency to be satisfied, h_0, h_1 should be perfectly calibrated.

The *generalized false-positive rate* of a classifer h_s for group $s \in S$: $c_{fp}(h_s) = E_{(x,y)}[h_s(x)|y = 0]$ where the protected group of $x = s$.

The *generalized false-negative rate* of a classifer h_s for group $s \in S$: $c_{fn}(h_s) = E_{(x,y)}[1 - h_s(x)|y = 1]$ where the protected group of $x = s$.

We define a cost function $g_s(h_s) = a_s c_{fp}(h_s) + b_s c_{fn}(h_s)$ such that relaxed separation with calibration is satisfied if $g_1(h_1) = g_2(h_2)$

We derive a calibrated classifier $\tilde{h}_2$ such that $g_1(h_1) = g_2(\tilde{h}_2)$ given our assumption that this would be the optimal solution with respect to classifier cost. This cost constraint can be achieved using randomization by occasionally returning the second group's mean probability when asked for $h_2(x)$ for a subset of the second group.

$$\tilde{h}_2(x) = \begin{cases} h^{\mu_2}(x) = \mu_2 \text{ with probability } p \\ h_2(x) \text{ with probability } 1-p \end{cases}$$

The cost of $\tilde{h}_2$ is a linear interpolation between the costs of h_2 and h^{μ_2}(See Figure 5.8) such that $g_2(\tilde{h}_2) = (1-p)g_2(h_2) + pg_2(h^{\mu_2})$. We thus set $p = \frac{g_1(h_1) - g_2(h_2)}{g_2(g^{\mu_2}) - g_2(h_2)}$ to ensure $g_2(\tilde{h}_2) = g_1(h_1)$ whilst preserving calibration.

Algorithm 5.2 is considered 'optimal' as it arrives at a unique solution that is calibrated and satisfies one of the relaxations there are implications to consider.

For example, to find equalized fpr/fnr we must modify/derive a classifier $\tilde{h}$ so that it is strictly worse for one of the groups - this is an unavoidable outcome in this case. Another implication is that by withholding information for one group in order to add randomization during the fitting procedure of our derived classifier, we make the outcome inequitable for specific individuals within a group. This can cause a significant utility loss for one specific group which can be seen as a definition of unfairness in itself.

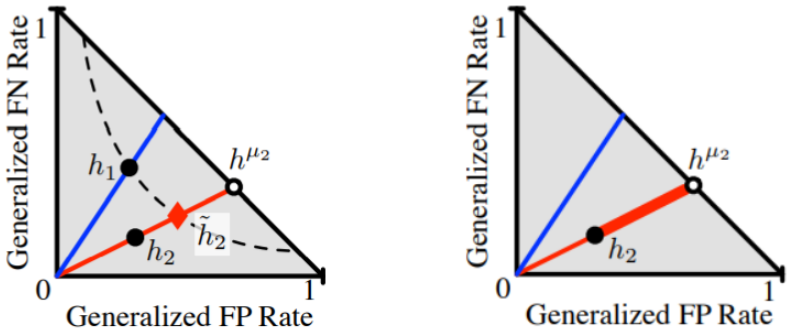

Figure 5.8: **Left**: Linear Interpolation of Derived Predictor $\tilde{h}_2$. **Right**: Possible Costs Linear Interpolation Between h_2, h^{μ_2} [31]

Input: classifiers h_1, h_2 so that $g_2(h_2) \leqslant g_1(h_1) \leqslant g_2(h^{\mu_2})$ and a holdout set X_{valid}
Determine base rate μ_2 using X_{valid} to produce a trivial classifier h^{μ_2}
Construct $\tilde{h}_2$ using p as an interpolation parameter.
Return $h_1, \tilde{h}_2$ which are calibrated and satisfy $g_1(h_1) = g_2(\tilde{h}_2)$.

Algorithm 5.2: Achieving Calibration and a Relaxed Separation Constraint

Chapter 6

Experiments, Results and Evaluation

6.1 Unconstrained Logistic Regression Estimator

We first show the results of an fairness unconstrained logistic regression estimator trained on each of the three target variables (Fig 6.1): A levels, Value Added and Final Score (target_a, target_v, target_f). These classifiers are designed to maximise accuracy only. The results show a large disparity in the cumulative distribution of predicted scores between groups when using A Level and Final Score as targets. We see that for the two 'unfair' goals, there are large disparities within cumulative frequencies of scores between groups. When using A Levels as the training target, we see that at scores between $0-0.5$ already $\sim 86\%$ of the pupils from state schools already fall under this score threshold. Whereas only $\sim 60\%$ of pupils from independent schools are included within this threshold. Similarly. when using Final Scores as the training target, $\sim 85\%$ of the pupils from state schools are predicted to be within this threshold whereas $\sim 70\%$ of pupils from independent schools are within this threshold. However, when using the Value Added measure as the training target, which we define as a 'fair goal, we see almost no discrepancy of cumulative frequencies of predicted scores between groups. Thus, using the former two 'unfair' goals - an unconstrained classifier that solely aims to optimise accuracy will generate predictions that give a higher likelihood of acceptance to the independent school group. Whereas using a 'fair' goal we see that the probability of achieving a high score is roughly equivalent between groups. These results are replicated across all three experiments (See Appendix B)

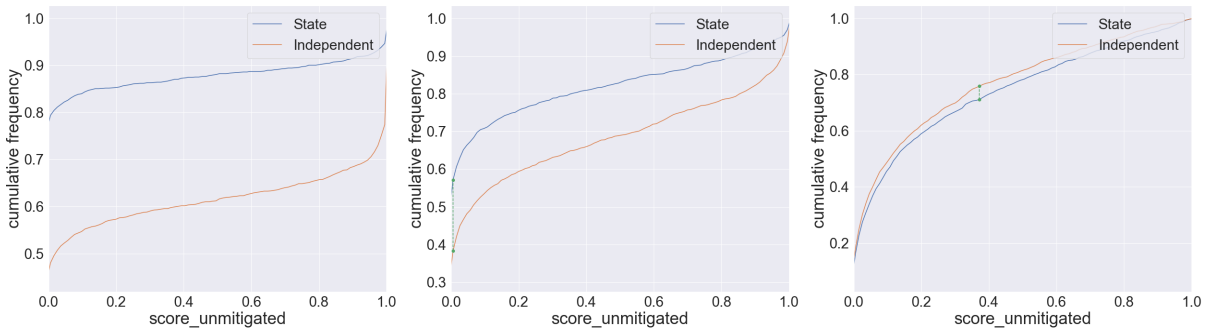

Figure 6.1: **Experiment 1**: Cumulative Frequencies of Predictions for a Fairness Unconstrained Classifier For Each Goal. **Left**: CDF Plots for Both Groups Predicted Scores Using A levels as the Training Goal). **Middle**: CDF Plots for Both Groups Predicted Scores Using Final Scores as the Training Goal. **Right**: CDF Plots for Both Groups Predicted Scores Using Value Added scores as the Training Goal

6.2 Comparison Between Different Fairness Constraints

Within this section we will reference each classifier using different goals with the following labels. A Level Goal: C_a, Final score Goal: C_f, Value Added goal: C_v. For each experiment, we train each classifier (corresponding to each of the three goals) under the same fairness constraint.

The prediction of each classification model $\{C_a, C_f, C_v\}$ for a particular individual X_i correspond to the set of predictions:

$$\hat{Y}_i = \{\hat{y}_i^a, \hat{y}_i^f, \hat{y}_i^v\}$$

We define the cost of a fairness metric to be the difference of the fairness metric between each group; defined in Section 4.6

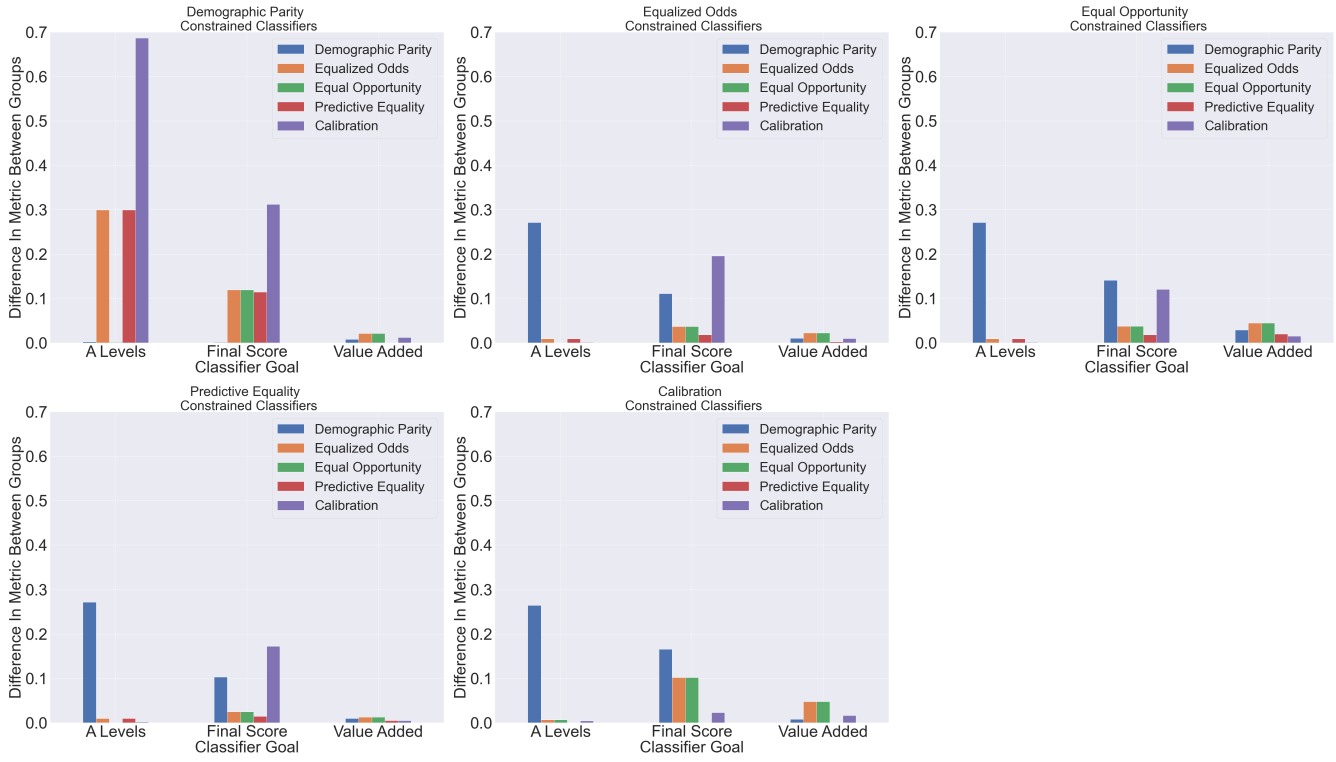

Figure 6.2: Costs of Various Fairness Definitions for Fairness Constrained Classifiers Trained Under Three Different Goals

6.2.1 Independence via Demographic Parity

These results allow us to compare three different classifiers C_a, C_f, C_v constrained to optimize for Demographic Parity and thus satisfy the Independence fairness constraint. We then compare the cost of $\hat{Y}$ of each classifier on the test set with each of the statistical fairness criteria. In Figure 6.2 we see that each of the classifiers are equally effective at mitigating disparity in Demographic Parity. However, both C_a and C_f incur a large disparity cost specifically in Calibration and Sufficiency. C_v has low cost and low variability amongst all fairness metrics.

The key takeaway is that the classifier trained with the Value Added Target alongside the Demographic Parity constraint manages to simultaneously reduce the cost of all other fairness metrics in comparison to the other two goals. However, we also infer that using Demographic Parity as a fairness metric incurs an increased cost within the other fairness metrics when an 'unfair' or possibly discriminate goal is chosen which can often occur in real scenarios. This references back to the Impossibility Theorem(See Section 4.5) and the possible limitations of using Independence as a fairness metric. The decision maker has to question if the benefit of equal likelihood of the positive outcome(selection rate) between groups is worth the sacrifice of increased disparities in error rates and precision amongst predictions, assuming the underlying base rates between groups is unequal given the measurement criteria/goal.

6.2.2 Separation

We split the results for this constraint into three subsections such that we train the classifiers C_a, C_f, C_v to be constrained to optimize for Separation but also two of the relaxed definitions. As before, we compare the cost of $\hat{Y}$ of each classifier on the test set with each of the statistical fairness criteria.

Figure 6.2 allows us to evaluate how Separation as a fairness constraint relates to its relaxed definitions and vice versa. Recall that we defined Separation as an equivalent to Equalized Odds and in Section 4.6.2 we defined the cost of Separation as the maximum of the cost of its relaxations: Predictive Equality and Equal Opportunity. Within these results, we see that each of the definitions of Separation give similar results amongst the costs of each fairness metric. Where for each classifier the Equalized Odds cost is the maximum of the cost between Predictive Equality and Equal Opportunity as expected. We see throughout each of the Separation definitions, C_a produces a high cost in Demographic Parity and C_f produces a relatively larger cost in Demographic Parity(Independence) and Calibration(Sufficiency). C_v again demonstrates low cost across all metrics when constrained to all three Separation definitions.

We note that satisfying a relaxation of the Separation constraint is not mutually exclusive to satisfying the other one - in fact by definition we should be able to satisfy both if we were to satisfy Separation at all. Which is why we see a low cost of Equalized Odds even when a classifier is solely constrained to one of Predictive Equality and Equality of Opportunity. Given the above, a decision maker must decide if the equality of model error rates between groups is worth sacrificing for an increase in disparities of selection rate and/or precision rate, assuming the underlying base rates between groups is unequal given the measurement criteria/goal.

6.2.3 Sufficiency

These results allow us to evaluate the classifiers C_a, C_f, C_v that are constrained to satisfy equality of Calibration between groups, a slightly stronger notion of Sufficiency. Recall from Section 4.4.3 that in the binary case, Calibration by groups and thus Sufficiency can be defined as equality of positive and negative predictive values between groups. We see that again, C_a incurs a high cost in the Independence fairness metric whilst C_f incurs a moderate cost in disparities of True Positive Rates as well as Selection Rates(Independence) between groups. C_v again demonstrates relatively low fairness costs across all metrics when solely constrained to satisfy Sufficiency.

It is worth noting that within C_f, we simultaneously satisfy low costs in Calibration and Predictive Equality, a relaxation of Separation, but not Separation(Equalized Odds) itself as proposed in Section 4.5.7. We discuss the nuances of this implication as well as the reasoning behind the simultaneous low disparities of TPR, FPR, PPV, NPV between groups for the classifiers C_a constrained under Separation and Sufficiency in the next section.

6.3 Fairness Costs Across Goals

Figures 6.3 and 6.4 provide a different perspective that helps us visualise the relationships of each of the fairness constraints and also demonstrate the effectiveness of our proposal of choosing a 'fair' goal to mitigate the Impossibility Theorem. The heat-maps allow us to evaluate C_a, C_f, C_v ,each trained to satisfy various fairness constraints, against the cost of certain fairness metrics that might be used to judge if a predictor is fair or not. We will refer to these heat maps as '*cost grids*' We use Figure 6.3 to show the relationship of our three core fairness definitions defined in Section 4.4: *Independence, Separation and Sufficiency* in the context of 'fair' and discriminative goals/targets to predict. Figure 6.4 is equivalent to the previous figure however it provides further insight into how the relaxations of Separation may interact with the other fairness constraints.

A Levels

The cost grid representing C_a clearly demonstrates the mutual exclusivity of satisfying Independence simultaneously with either Separation or Sufficiency - demonstrated by the large between group disparity costs of $0.28, 0.68$ respectively in order to maintain a low disparity of selection rate between groups. This is expected as Demographic Parity(Independence) is the sole fairness metric that doesn't take into account/isn't conditional on the ground truth labels Y, in fact we recall from Definition 4.11 that Independence is reliant upon the predicted outcomes being independent from the protected attribute or the ground truth.

However, C_a can satisfy low costs in Sufficiency and Separation at the same time, demonstrated by the neutral coloured 2x2 cost grid within the bottom right of the entire grid. This may seem to contradict the Impossibility result. This is expected, as we recall that C_a is a classifier that tries to predict A levels scores, using $target\_a$ as the training labels and that each of the classifiers uses a_levels as a feature in the feature vector when training. This corresponds to the information we have of an individual during time of predicting admissions. Given that $target\_a$ is directly correlated with a_levels, the trained classifier C_a should come close to being a perfect predictor for predicting A level scores and thus a perfect classifier for classifying an individual to be either accepted or unaccepted when using A levels is used as the admissions criteria. This is reinforced by Figure 6.5 where the classifiers trained to predict A levels have an MSE of approximately zero when using $target\_a$ as the target labels. We should also take into account that when generating our data-set and sampling from the distribution of A levels, we only add a small amount of noise and assuming that the A level values from our test set are sampled from the same distribution as our training set.

We recall from Section 4.5.5 that the 'rare' requirements to satisfy both Sufficiency and Separation is to either have equal inherent base rates of the target variable between groups or that we are able to train a perfect predictor. In this case, the latter is satisfied as we have described that C_a is approximately a perfect predictor for $target\_a$, thus allowing us to satisfy both Equalized Odds and Calibration at the same time - even when the classifier is only trained to satisfy one of the definitions.

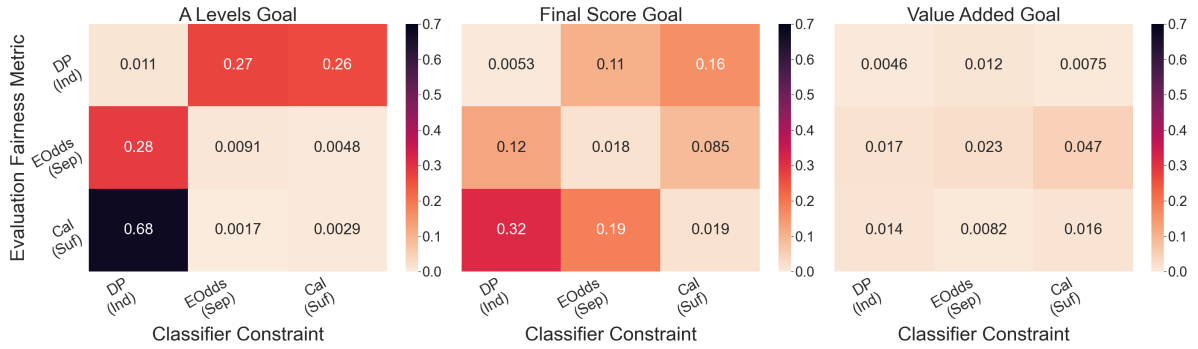

Figure 6.3: Cost Grid of All Fairness Constrained Classifiers Trained w.r.t Cost of Three Fairness Metrics

6.3.1 Final Score

The cost grid representing classifiers C_f is a clear example of the Impossibility Theorem defined in Section 4.5. Each possible pair of fairness constraints given the three core definitions will be mutually exclusive at any one time unless we are able to satisfy either having equal base rates of positive outcomes between groups for our target variable, in this case $target\_v$, or train a perfect predictor. Clearly within these results, we can achieve neither as the only region where we achieve low cost within the heat map is the diagonal, indicating that the classifier only performs well against the test fairness metric it is what trained to satisfy. The difference between the results/the cost grids forC_a and C_f, even though their target labels $target\_f$, $target\_a$ are highly correlated is the slight difference of generating the features used to also generate the target labels for the training set. Recall from Section 5.1, and individuals $final\_score$ value is directly generated from the individuals a_level value but with additional noise. This is important to note as an individuals final score can be seen as a proxy to the value of their A level score, however, a perfect predictor cannot be obtained(as seen in the results for C_a) due to the additional noise when generating the values for the target variable $target\_f$ in the training set - even though the a_level value is used as a training feature and is a proxy for an individuals final score and thus $target\_f$. We may perceive Final Scores to be a fairer goal than A levels as the additional noise potentially introduces more uncertainty into the associations of the sensitive attribute to the admission metric. This would for the mitigation of some of the inherent discrimination within the data or goal that a trained model may emphasise - we see this within our cost grid for C_f where the fairness costs are less extreme than for C_a. However, this additional uncertainty can potentially reduce the accuracy of C_f - thus incurring an accuracy-fairness trade-off which we discuss further in Section 6.4.

Again, we see that C_f under Demographic Parity and thus the Independence constraint causes the largest conflict within the other test fairness metrics causing a Sufficiency disparity cost of 0.32.

6.3.2 Value Added

This result demonstrates the validity of our thesis proposal. We see in the right cost grid in Figure 6.3 representing the costs of the test fairness metrics for C_v trained under the three core fairness metrics. This cost grid shows a consistent neutral colour across all grids - representing a low fairness cost across all test metrics across all possible fairness constraints that C_v has be trained to satisfy. This is clearly directly conflicting with the Impossibility Theorem, where each of the fairness metrics should be mutually exclusive with each other. However, in this case using our 'fair' goal we have been able to satisfy each of the fairness metrics: *Independence, Separation and Sufficiency* simultaneously.

The reasoning behind this is that the classifier C_v trains to predict the target label *target_v* which is generated based on an individuals *value_added* scores. As we've demonstrated in Figures 5.2.3 and 5.5, the underlying base rates of acceptance between groups are equal if we used the value added metric as the core success criteria. Knowing this, we see that an individuals value added score is completely independent of the sensitive attribute. The feature is in fact solely correlated with the individuals base IQ with two additions of noise, which is generated from the same distribution no matter the group. We recall from Section 4.5 that having equal base rates between groups given the target metric/goal is one of the unique scenarios that allows us to combat the Impossibility Theorem. Thus, even without being able to train a perfect predictor this goal allows us to satisfy low disparities within all three core fairness constraints and is can thus be considered a 'fair' goal.

We should note, that given that we have added two additions of noise, our trained predictor C_v will be fairer than the other classifiers C_a and C_f but will incur some trade-off in accuracy. We discuss this further in Section 6.4. Ultimately, the decision maker will have to decide if this trade-off is worth it in order to satisfy all fairness definitions. There is obviously not a strict answer for each problem case and the decision maker will have to take into account the context of each individual and also whether the cost of disparities within a certain fairness metric is equivalent between all fairness metrics.

6.3.3 Fairness Performance of Relaxed Definitions of Separation

The cost grids in Figure 6.4 is equivalent to the previous results using the three-by-three cost grid. However, we use the relaxed definitions of Separation to demonstrate the propositions in Section 4.5.6 and ideas by [31]. Recall Proposition 4.4 and Algorithm 5.2 where we introduce the idea of deriving a predictor that is well calibrated between groups but also satisfies one of the relaxations of Separation at the same time: equal true positive or false positive rates between groups. In this case, we see for each of our calibrated constrained classifiers, we satisfy equal false positive rates between groups equating to a 0 cost in predictive equality. The most demonstrative case is again the middle cost-grid of C_f. We are able to train a classifier that satisfies a low cost in Calibration and Predictive Equality(FPR) but notably has a relatively high cost in Equal Opportunity(TPR). Clearly, this validates the proposal that we can satisfy Sufficiency and one of the relaxations of Separation simultaneously assuming we do not satisfy one of the two unique cases where the Impossibility Theorem does not necessarily apply. Within the other two cost-grids, we are able to satisfy both Sufficiency and both relaxations of Separation due to the reasons we have mentioned in Section 6.3.

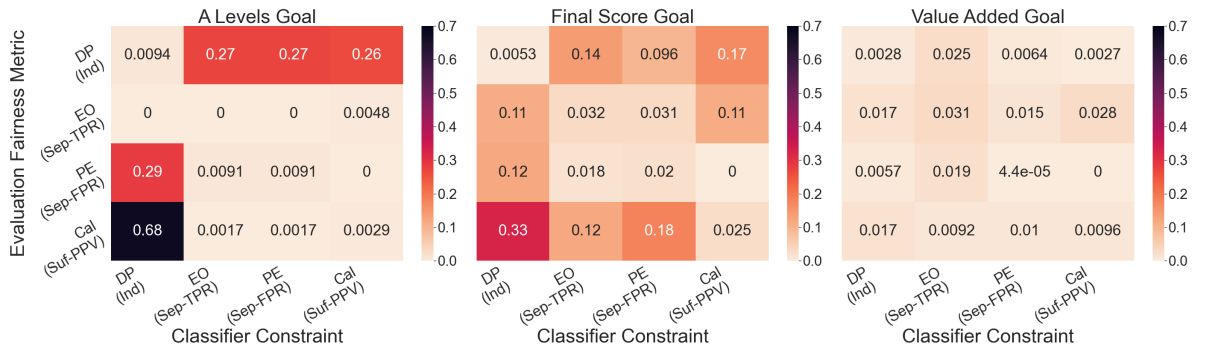

Figure 6.4: Cost Grid of All Fairness Constrained Classifiers Trained w.r.t Cost of Three Fairness Metrics Including Relaxations

Intuitively, we can infer equivalencies between both Figures 6.3 and 6.4. Given that we define the cost of Separation as the maximum between the cost of Equal Opportunity and Predictive Equality, it is easily

to see that the maximum values of the center 2x2 grid within each of the 4x4 grids is equivalent to the value of the center grid within each of the 3x3 grids in the previous section. This also demonstrates that satisfying Equalized Odds is equivalent to satisfying both Equal Opportunity and Predictive Equality. This most clear demonstration is within the middle cost grid of C_f where we have a neutral 2x2 grid representing a low cost to both, equivalent to the neutral coloured center grid of the 3x3 grid for C_f within the previous section.

Other notable observations is that within C_a we are able to achieve perfect equality of true positive rates for each of the fairness constraints with the exception to Sufficiency which optimises false positive rates alongside Calibration constraint. The intuition behind the results of C_a under Demographic Parity is interesting to explore further as it provides insight into why Independence may not be the most appropriate fairness constraint if we want to also care about utility. Given that Demographic Parity(Independence) solely cares about about equality in rate of positive outcomes between groups, then it can trivially assign positive outcomes by flipping coin as long as it satisfies the constraint. However, as our method of achieving Demographic Parity using Threshold Optimisation (See Section 5.4.2) requires that we also optimise for accuracy, we iterate through a range of probability thresholds to maximise accuracy. In this case, our threshold that maximises accuracy allows for correct prediction of all individuals with underlying positive ground truth labels. However, at the same time we must satisfy approximate equal selection rates between groups by the definition of Demographic Parity. Recall that for our A level goal, the underlying rates of positive outcomes between group(base rates) is not equal(See Figure 5.1) implying that if we correctly predict true positive outcomes for the group with the largest base rate of positive outcomes, then we are also directly cause an increase in false positive rates within all groups with the lower base rate of positive outcomes. Thus causing an increase in negative predictive value and model error rates via false positives for those groups which is the reasoning to the increase in Predictive Equality and Calibration costs. We expect this to be reflected in the MSE of C_a under the Independence constraint.

Referencing back to the previous results in Section 6.3, it is no surprise that C_a when constrained to Sufficiency and Separation achieves low disparity costs in Equal Opportunity, Predictive Equality and Calibration. We should be able to satisfy equality in true positive and false negative rates if we were to satisfy Equalized Odds and by proxy Separation. Given the explanation for the ability to train a perfect predictor using A levels as a goal for C_a, we can also satisfy Sufficiency simultaneously. This naturally allows us to hypothesise that these fairness constraints will also allow us to maximise accuracy for C_a by definition of Separation and Sufficiency not being mutually exclusive and/or the fact that this is the implication of being able to train a perfectly accurate predictor. We explore if this is assumption is valid in Section 6.4.

6.4 Fairness vs Accuracy Trade-off

The trade-off between accuracy and fairness is important to consider for a decision maker as it defines how we balance the importance of utility and equity within our decision. Clearly, there is no clear-cut rule or law that provides strict guidelines for this as this decision is contextually dependent on the values of the stakeholders. These results aim to demonstrate the potential trade-offs we may have to make with respect to accuracy and fairness depending on which classifier out of C_a, C_f, C_v to choose.

Within Figure 6.5 we see that Demographic Parity is the only fairness constraint that does not allow us to achieve a perfect predictor with any of the target labels/training goals. This reinforces our hypothesis from the previous section and also from Section 4.4.1. We explain that satisfying Independence is guaranteed to cause a trade-off of accuracy as our predictions ignore the ground truth labels and also incurs the same flaws as implementing a sensitive attribute blind predictor which we describe in Section 4.4.4. The result is a larger number of false negatives and/or false positives, representing an increase in model errors, and a decrease in model precision. This is supported by the experiments within the previous section(See Figures 6.3, 6.4). Figure 6.6 demonstrates that Demographic Parity incurs the highest total fairness cost over all metrics for our classifiers C_a, C_f trained under 'unfair' goals. This means that Independence may further increase discrimination in contexts where the problem case may already include inherent discrimination. Classifier C_v representing our 'fair' goal has the highest trade-off in accuracy in order to satisfy Independence, however, we see that it incurs a dramatically lower total cost in fairness under all metrics.

The MSE graphs for each of the classifiers constrained under Equalized Odds, Equal Opportunity and Predictive Equality are approximately equal. This is to be expected as Equal Opportunity and Predictive Equality are relaxations of Separation and can be combined to satisfy Equalized Odds. Therefore it

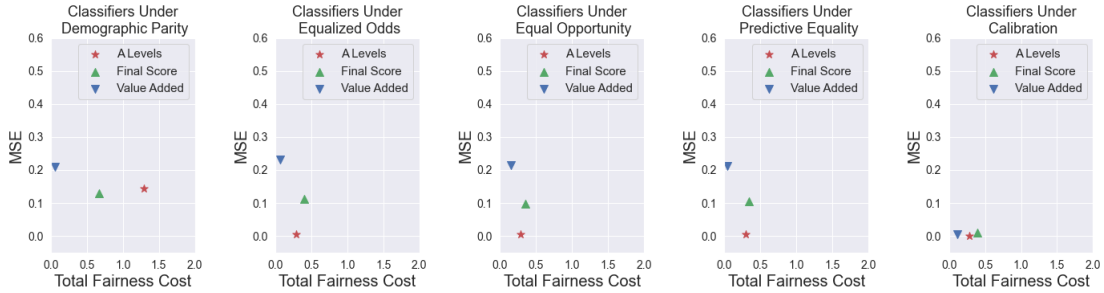

Figure 6.5: Fairness Constraint Cost vs Mean Squared Error for Different Goals Trained Under the Same Constraint

is expected that the classifiers trained to constrain Equalized Odds will not perform worse than the maximum error between the two relaxed definitions, which in this case incur the same errors amongst all three classifiers. We are able to validate our assumption of being able to train C_a to be a perfect predictor constrained under either Separation and Sufficiency whilst being able to satisfy both simultaneously. C_f, C_v constrained to these fairness definitions both incur a trade-off to accuracy in comparison to C_a with C_v being the least accurate. This is expected due to the additional noise introduced when generating the training labels $target\_f, target\_v$ and the presence of an individuals A level score within the training features, allowing for C_a to be trained to be an approximately perfect predictor of $target\_a$. Intuitively this means if our sole objective was to maximise accuracy C_a can be considered the classifier that offers the most utility and C_v being the least. However, taking into account the equity of each classifier (See Figure 6.6), C_v clearly performs better than the other two classifiers with approximately half the fairness costs of both. It is worth noting that if the value added measure was available at prediction time, we can assume that we would be able to create an approximate perfectly accurate and overall fair predictor. This is obviously not feasible in real University admissions scenarios as this feature by definition cannot be available at the time of admission.

Finally, we see that all classifiers have equally low MSE when constrained to Sufficiency, this is expected. Recall that Sufficiency conditions on the predicted score and evaluates if the probability of success (or failure) is the same across groups. Recall that constraining a classifier to Sufficiency requires the calibration of each group. This calibration process is equivalent to requiring equally high precision for each group. This means that constraining to Sufficiency also requires the proportion of incorrect positive/negative predictions to be low for both group which directly translates to training an accurate predictor as well. We note that this fairness definition is commonly used in practise as we don't always have the ground truth labels to condition our predictions in order to evaluate model errors.

The results above demonstrate the importance of taking into account the accuracy-fairness trade-off, as optimising for both is often infeasible. [6] states that even when base rates are the same between groups, one could 'not achieve perfect fairness while also getting perfect accuracy'. We see that the classifier C_v that is trained on a 'fair' is far superior to the other classifiers trained on 'unfair' goals when the priority is to mitigate the Impossibility Theorem and satisfy all the fairness metrics simultaneously - thus validating our proposal(See Figures 6.12, 6.4). However, achieving low disparity costs between groups for each fairness metric comes at a cost of accuracy. For instance, achieving a low disparity cost for all fairness definitions does **not** imply that the cost of the fairness metric is low for each group. A classifier can satisfy equality in false positive rates and thus Predictive Equality whilst simultaneously having a high rate of false negatives/false positives for each group. This would correspond to a high model error rate and thus a trade-off in accuracy. The same can be said for the Sufficiency constraint where each group has equally low precision rates, resulting in satisfying the fairness constraint but with a cost in accuracy. This reality can be seen in Table D showing the metric values within the experiment for each classifier under each constraint. C_v consistently has higher TPR, FPR and lower PPV,NPV values compared to the classifiers trained to predict the other goals, resulting in higher model errors and lower precision rates as expected.

Ultimately, a stakeholder is left with the decision of how much utility is worth sacrificing for satisfying all fairness definitions but also how the definitions are weighted in comparison to each other. This decision is inherently context specific to the problem itself. We provide possible following scenarios that show the possible viewpoints and reasoning behind choosing each of three goals in the University Admission problem case:

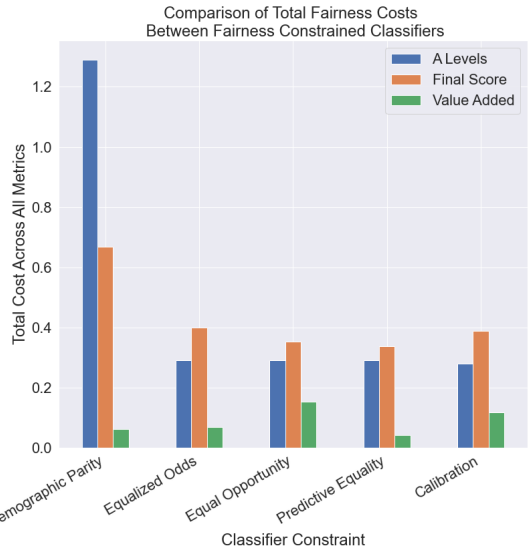

Figure 6.6: Comparison of Total Cost of Fairness Across All Three Goals Across All Fairness Constraints

- If a stakeholder solely wanted to maximise equity of a decision, then the classifier C_v is the clear winner as it is able to at least half the total fairness cost of the next best classifier. However, the stakeholder will have to accept at least a two-fold increase in accuracy cost to the next worse performing classifier in terms of accuracy C_f(an increase of ~ 0.1 MSE).

- C_a may be the most attractive option if the stakeholder solely cared about the Separation and Sufficiency fairness constraints and also wanted to maximise accuracy. However, this is requires the ability to train a perfect predictor. In most real world scenarios, training a perfect predictor will be infeasible as it will require full knowledge of the underlying distribution of the features, being able to account for all possible sources of noise and perfectly predicting the ground truth of data that our trained predictor has not seen before. This is clearly often not a reasonable assumption to make. Another problem with this classifier/goal occurs when the stakeholder priorities satisfying Demographic Parity. This would result in a significant cost to the other fairness metrics as well as accuracy, demonstrated by our previous results. Often the large disparities for error and precision rates between groups would be considered to be unacceptable and a cause for disparate treatment.

- C_f representing the Final Score goal may seem like the least attractive choice for a stakeholder. Even though the total fairness cost when satisfying Demographic Parity is less extreme than C_a, the disparity cost is till at least seven times larger than C_f under the same constraint. This classifier/goal also performs worse in terms of total fairness costs than the other two options - notably at least a two-fold increase fairness cost compared to C_v and a $0.05 - 0.1$ increase compared to C_a. The accuracy of C_f does perform better than C_v but is relatively just as costly with respect to accuracy when compared to C_a.

Another perspective to consider when choosing which goal to use is when we weight the value of the cost in accuracy equally to the same cost of fairness. Taking this perspective, the far smaller fairness cost of C_v compared to the other classifiers with the ability to satisfy all fairness metrics far outweighs the increase in MSE/trade-off in accuracy. We should also note that choosing a 'fair' goal with equal base rates that allows us to satisfy all fairness constraints may be a more feasible option in real world scenarios compared to training a perfect predictor. Even in the case that we are able to train a perfect predictor, we will only be able to satisfy Separation and Sufficiency simultaneously with a significant trade-off in Independence. Thus for a stakeholder whom is willing to accept a small/moderate trade-off in accuracy in order to significantly reduce the cost of discrimination, the choice is obvious. Choosing the classifier trained under our 'fair' goal represented by our target label: *value_added* is the best option - thus validating our proposal and achieving our thesis aim.

6.4.1 Acceptance Rates Amongst Outcomes of Fair Classifiers

Figure 6.7 illustrates the acceptance rate for both education groups within the predictions of each classifier on our test set after being trained to satisfy each of the fairness constraints. Given the definition of Demographic Parity, the acceptance rates between groups within our classifier predictions is expected to be roughly equal under this constraint. We see this result for all three classifier goals. However, for our two classifiers trained to predict our proposed 'unfair' goals C_a, C_f, we clearly see a large disparity in rates of acceptance between groups within the predicted labels of all other fairness constraints. This can be interpreted and a common case for citing disparate impact and also portrays that if the priority is to achieve equal selection rates(Independence) then the other fairness constraints may not be suitable - again reinforcing the mutual incompatibility of Satisfying Independence with either Separation and/or Sufficiency. As we stated before, Final Score and thus classifier C_f may seem to be a 'fairer' goal than A

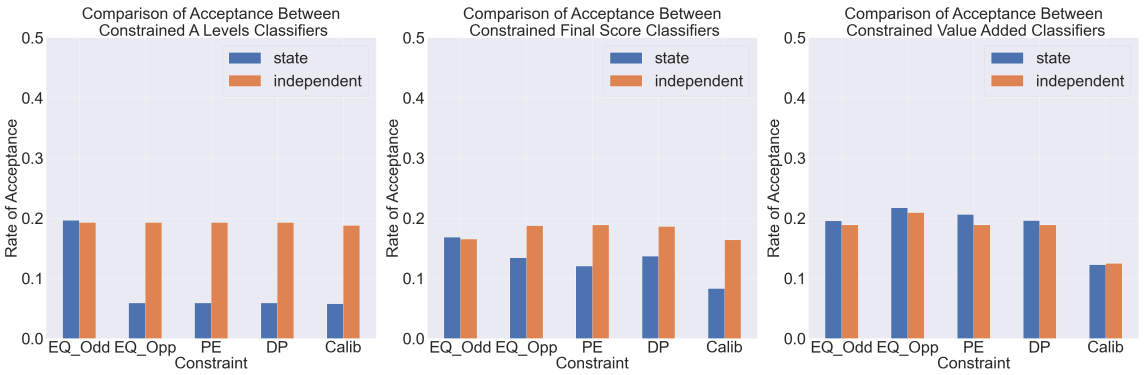

Figure 6.7: Rate of Acceptance Between Groups for Each Goal w.r.t Each Fairness Constraint

levels. In the case of prioritising balancing selection rate, this perception may hold true demonstrated by the reduced disparity of selection rate between groups across all classifier fairness constraints compared to C_a. Though far from equal. It is important to keep in mind that this reduction in fairness cost comes at a cost in accuracy as described in the previous section but also that solely reduces the cost in one fairness metric, Independence, and overall has a higher total fairness cost than C_a and C_v. As expected, the classifier C_v trained to predict our proposed 'fair' goal, value added, clearly performs the best when mitigating the disparities of selection rates between groups, even when optimised to satisfy other fairness metrics.

6.5 The Effects of Changing Population Distribution Rates of School Systems

We now evaluate the effects of changing the proportion of the population educated in Independent and State schools. Recall from Section 5.1 that when generating our data-set, for ease, we allocated both groups with roughly equal size by sampling the feature from a Bernoulli(0.5) distribution. This clearly does not reflect real life, in fact approximately only 7% of the British population is educated within the Independent school system [11]. To reflect this, we repeat the above experiment using a Bernoulli(0.93) distribution.

We first evaluate the base rates of acceptance for each group for each target variable. Figure 6.8 shows that the disparities still occur within the 'unfair' goals $target\_a, target\_f$ whereas the base rates are almost equal between groups for our 'fair' goal $target\_v$. Thus, we can expect similar results from the previous data-set with equal group sizes but with two noteworthy points to consider.

First, compared to our previous data-set the disparity between base-rates for the A level target variable can be inferred to be even more pronounced as the base probability of admission for independent school is in fact higher than the probability of rejection. In comparison to our initial data-set, we saw that each of the target variables had base rates of acceptance being lower than the rate of rejection. Ethically, this is especially questionable and would definitely be a case for discriminatory behaviour. Not only does the group educated from an Independent school system has higher probabilities of admission than the group educated from the state school, they also have higher within group probability of being accepted than rejected.

The second noteworthy point is that the base rate of acceptance when using the value added target variable within this data-set is slightly higher to the privileged group. Though they are very nearly equal between privileged(independent) and unprivileged(state) groups, the disparity between the two is slightly higher than the previous data-set within the same experiment. We discuss the implications of these two observations below.

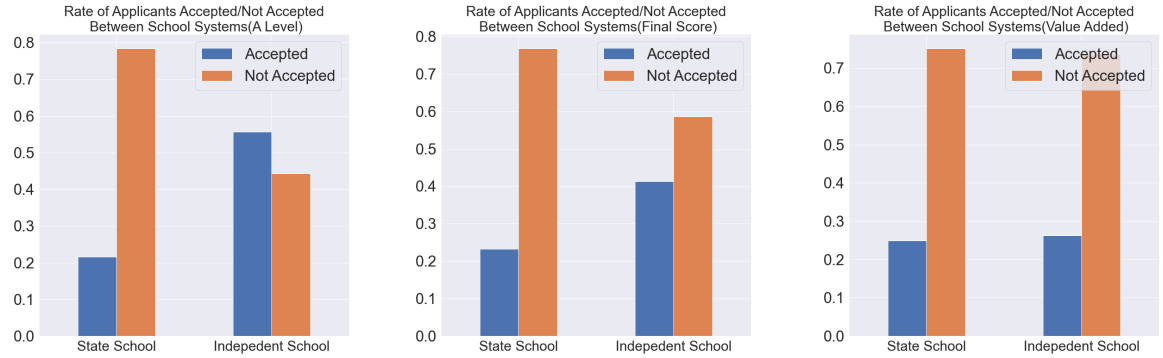

Figure 6.8: Histograms of Base Acceptance/Reject Rates Between Groups Depending on Target: school_system ∼(Bernoulli(0.93))

Due to the larger disparity with base rates for all target variables, we expect there to be an more significant cost to the model error rates(FPR, FNR) and with precision rates when satisfying Demographic Parity. The larger the disparity in base rates of acceptance, the more errors we will have to make in order to satisfy equal selection rates between groups - this applies to both groups. This hypothesis is validated in Figures 6.9 and 6.12 in comparison to the equivalent visualisations in the previous experiment where the total fairness costs for each classifier in order to satisfy Independence is $1.62, 0.8, 0.21$ compared to $1.31, 0.68, 0.08$ respectively. Specifically, we see a large increase in the cost of Equal Opportunity/disparities of true positive rates within each of our classifier predictions(See Appendix D.2). As with the previous data-set, we are able to satisfy Separation and Sufficiency simultaneously due to being able to train a near perfectly accurate classifier C_a.

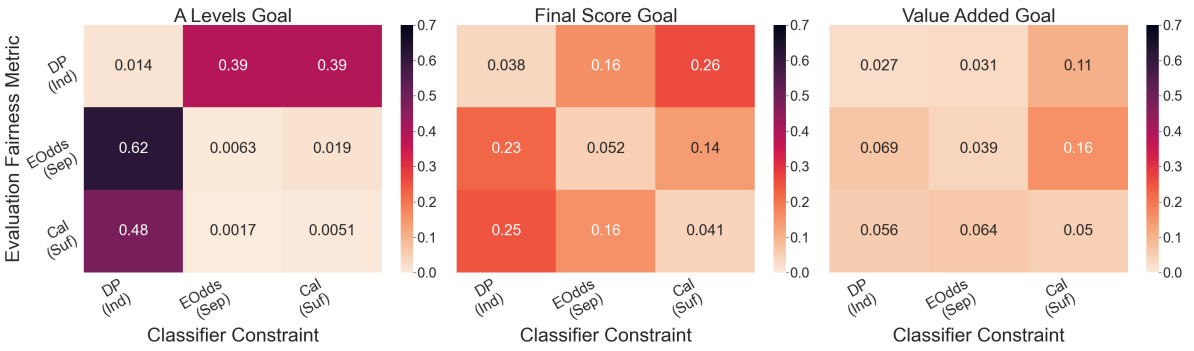

Figure 6.9: Cost Grid of All Fairness Constrained Classifiers Trained w.r.t Cost of Three Fairness Metrics

An interesting observation is that instead of having a large disparity in Predictive Equality, we now have a large disparity in Equal Opportunity for classifiers satisfying Independence (See Figure 6.10). This result can be explained by the classifier having access to a smaller sample size of the privileged group to train on. This would cause an increase in uncertainty in the ability to identify true positives for this group. Conversely, the classifier has a large set of data for the unprivileged group to train on, thus being able to more accurately identify the associations between the features and the target in order to accurately predict positive outcomes for this group. This disparity is emphasised for the classifier C_a due to the ratio of acceptance(base rate) within the training set for each group when using A levels as the admissions metric. The privileged group has a higher proportion of individuals accepted than rejected, therefore we would expect a classifier to be less able to identify the nuances that attribute an individual from this group to be classified as accepted. In contrast, due to the significantly smaller portion of the

unprivileged group being accepted within the training set, we would expect a classifier to perform better at identifying the associations of an individuals features to a positive classification for this group. In this case, this would be an individuals IQ value.

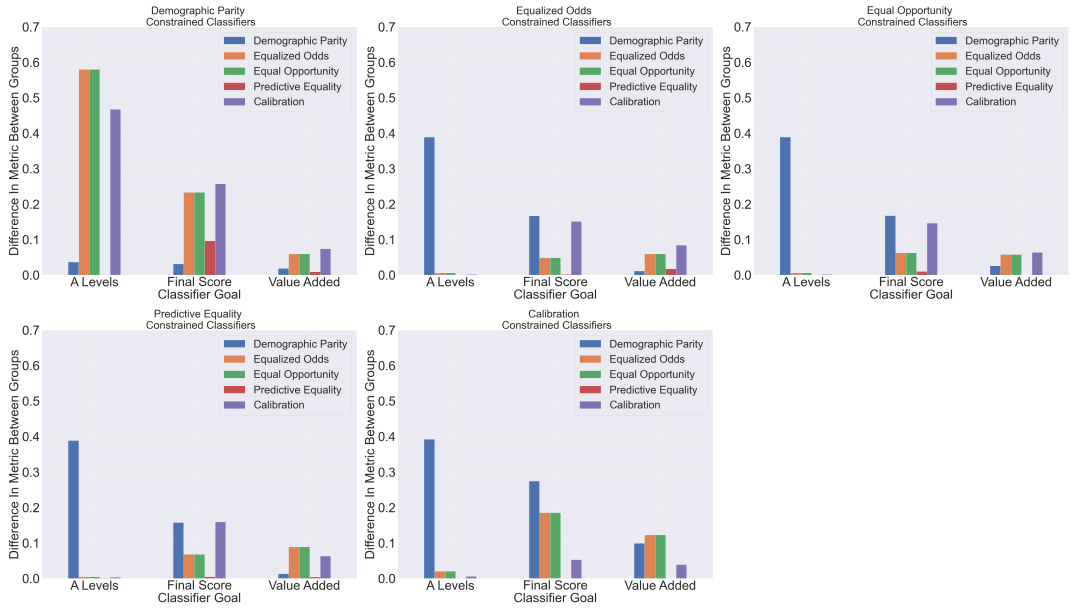

Figure 6.10: Fairness Constraint Cost vs Mean Squared Error for Different Goals Trained Under the Same Constraint

The same reasoning can be applied to the resulting cost grids for C_f, C_v in Figure 6.9. A notable outcome is that we see a significant increase of the Separation cost for C_v under the Sufficiency constraint. C_v nearly has an equal fairness cost to C_a though still less. This can also be attributed to the slight increase within the difference in base rates of acceptance between groups within this data-set for the value added acceptance metric. Thus causing C_v to perform slightly worse to mitigate the Impossibility result. However, we see that in Figure 6.11 that under Sufficiency, the accuracy of each of the classifiers are approximately the same, thus this slight increase in fairness cost does not incur a large accuracy-fairness trade-off compared to the other classifiers.

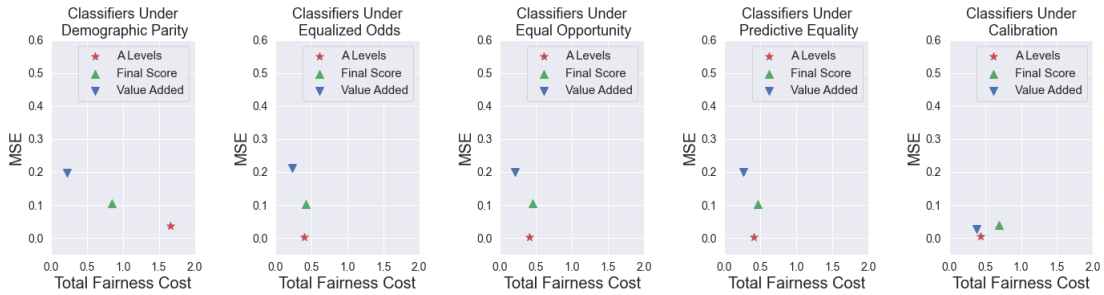

Figure 6.11: Fairness Constraint Cost vs Mean Squared Error for Different Goals Trained Under the Same Constraint

The accuracy results for each of the classifiers is similar to the previous data-set(See Figure 6.11). We are still able to obtain the perfect predictor C_a under the Separation and Sufficiency constraints and the relative costs between classifiers are also similar between for all constraints. A notable result is that under the Independence constraint, the fairness cost of C_a and C_f are significantly larger than before which we reasoned for above.

Overall, we come to the same conclusion that the classifier C_v trained under our proposed 'fair' goal incurs a significantly lower total fairness cost over all fairness definitions(See Figure 6.12). The same reasoning with respect to the accuracy-fairness trade-off can be applied from our previous experiment. The value added goal still has a positive to equal accuracy-fairness trade-off under all fairness definitions, assuming the stakeholder weighs the cost in fairness equally to the cost in accuracy. One could argue

under Sufficiency, the total cost of fairness between C_a and C_v can be considered relatively equal, however in this scenario the MSE/accuracy cost between these two classifiers are also approximately equal. Thus C_v still performs better than C_a as the increase in fairness is still proportional to the slight decrease in accuracy.

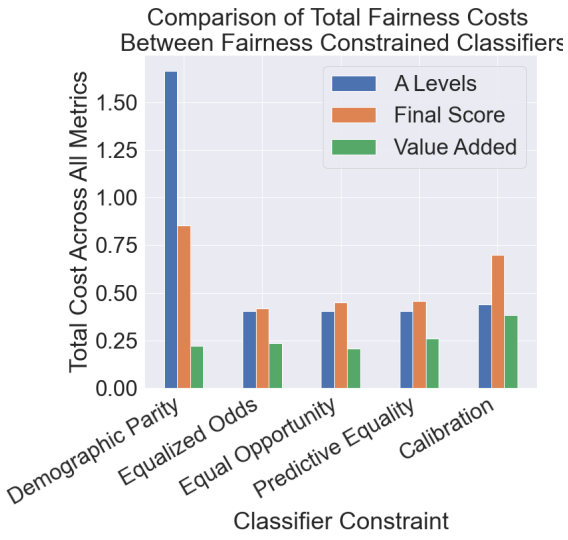

Figure 6.12: Fairness Constraint Cost vs Mean Squared Error for Different Goals Trained Under the Same Constraint

6.6 Sensitive Attribute Blind Classifiers with Proxy Features

Our final experiment aims to demonstrate the futility of 'blind' classifiers in both aspects of accuracy and fairness when their is inherent target variable bias within the classifier goal which we discussed in Section 4.4.4. We emphasise that the results from the previous experiments carry over to scenarios where the classifiers are blind to the sensitive attribute with the presence of proxy features that introduce 'proxy' discrimination. We conclude that in this scenario our classifier trained to predict the value added measure, representing our 'fair' goal, is still the best option in terms of mitigating the impossibility result and in terms of the accuracy-fairness trade-off.

Section 5.1.2 describes the feature set we use within the training set of this experiment for each classifier. Even though we do not include the school system explicitly within the training features, the features: work_experience and parent_income provide enough information to the classifier to allow it to learn the value of the sensitive attribute by proxy as the values are clearly separable by group(See Figure 5.4). Thus the classifier can learn the association of the sensitive attribute to the target label/goal without the need of the sensitive attribute itself. The results shown in Appendix C.1 validates our assumption that a 'blind' classifier is equivalent to the classifiers within our previous experiment where the sensitive attribute is made explicit due to the presence of proxy features. We are able to obtain very similar cost grids for each classifier as well as equivalent total fairness and accuracy costs across all classifiers thus obtaining the same accuracy-fairness trade-offs for each. The results of this experiment are important as in many real world scenarios, we may not have access to an individuals sensitive attributes due to data protection laws but also the presence of proxy features that can be combined to infer the value of the sensitive attribute. Taking this into account we still conclude that choosing an inherently fair goal, such as our value added metric, remains the optimal choice for a stakeholder whom cares about optimising fairness whilst maintaining high accuracy of a trained predictor.

6.7 Training Classifiers with Access to Full Feature Space

The purpose of our final experiment is to show the results of classifiers that have full information/access to all features. Recall the features used within training the classifier from Section 5.1.3. The classifiers now have access to all available features. The classifiers are now provided full information to ascertain

the association between the feature space and the target/goal. Intuitively, we may presume that the classifiers may perform better in-terms of accuracy whilst maintain the same fairness results due to the additional information of the underlying distribution of the data. However, we expect that the addition of the proxy features do not increase the accuracy of any of the classifiers as they provide they encode the same information as each other. Thus the sensitive attribute being both explicitly and implicitly encoded via the proxy features and the feature itself should not improve performance. Using the same reasoning, we do not expect that there will be additional discrimination introduced due to the additional proxy features as they do not provide additional information with respect to the relationship between the sensitive attribute and the goal. Thus we expect to achieve the same results as the two previous experiments with respect to the accuracy-fairness trade-offs within each of the classifiers corresponding to each of the target labels. The results in Appendix C.2 validate our reasoning.

Chapter 7

Conclusion

In this thesis we proposed and demonstrated the effectiveness of choosing non-discriminate goals within machine learning models in order to mitigate the 'Impossibility Results' found by [10], [3] and explained in Section 4.5. The success of accomplishing this aim illustrates the importance for stakeholders to prioritise putting considerable thought into choosing a goal that represents their true objective and achieving equal base rates between groups, in contrast to solely relying on optimising for fairness after-the-fact. This is especially relevant in scenarios where machine learning models are used to generate predictions within social decision making contexts, where the objective is to optimise for fairness and accuracy.

The experiments use the context of University admissions to train a classifier to predict an classify individuals from a simulated data-set as accepted or rejected. We use three different classifiers trained on three different target variables representing the decision goal: A-level score, final graduation score and value added score where the two former classifiers are considered to trained to predict 'unfair' goals and the latter being a 'fair' goal. The results of the experiments within this thesis validate our proposition and achieves our aim: we demonstrate that choosing a 'fair' goal for a predictor is superior in achieving overall fairer outcomes, mitigating the impossibility result in addition to maintaining accuracy - in contrast to mitigating for fairness after-the-fact(via pre, in or post processing).

7.1 Summary of Contributions

Within our first two chapters, we provide insight into our thesis proposal and the overarching aim as described above. The second chapter introduces various fairness group criteria, abstractly defined in the Law as Disparate Impact and Treatment and formally as our three main observational metrics: Independence, Separation and Sufficiency. In conjunction, revealing methods of how fairness constraints have been implemented and measured within machine learning models proposed in recent literature. We also discuss how discrimination might be introduced within decision making contexts leading ultimately to real-world scenarios where such criteria have been applied and used as a case for discriminate practise. Notably, this chapter also introduces the Impossibility Theorem that has been proven within recent fairness literature showing the incompatibility and seemingly contradictory nature of satisfying all stated fairness metrics simultaneously, unless in certain rare and/or trivial cases where overall utility of the model may be put into question. Subsequently, we evaluate the weaknesses of our chosen observational fairness criteria and give an intuition in which to consider if 'overall' fairness cannot be achieved and what fairness trade-offs that will incur. We also describe possible relaxations that can be made to allow for more than one definition to be satisfied at once. Finally, we delve into other possible fairness definitions such as Causal and Individual Fairness that may highlight the shortcomings of observational fairness metrics that this thesis uses to evaluate fair outcomes.

The following chapter formalizes our problem case as a machine learning classification problem, giving a brief outline of a standard classifier model and the methodology of how we implement fairness constraints within the optimization problem of a standard linear classifier. This chapter also describes the process in which we generate our simulated data corresponding to our University admissions scenario, providing analysis of the underlying distributions and laying out the expected results of each classifier given the distributions in which the target labels are sampled. We use this chapter to show the results and evaluate the performance with respect to accuracy and fairness within three experiments: the training data contains the sensitive attribute, the training data contains the sensitive attribute via implicit encoding of proxy features, the training data has access to all features. We discover that each exper-

iment provides the equivalent result: the classifier trained to predict the 'value added' goal generates the lowest total fairness cost by allowing us to satisfy all three fairness definitions at once, mitigating the impossibility result whilst achieving the best accuracy-fairness trade-off compared to the other two classifiers.

Crucially, the results of all three classifiers reinforce the impossibility theorem proposed in the literature by showing the fairness trade-offs that occur in constraining a model to satisfy a certain fairness definition. Simultaneously, the results demonstratw the specific scenarios that may allow us to satisfy two or more definitions at once.

7.2 Future Directions

7.2.1 Causal Inference and Individual Fairness

Within this thesis, we solely use observational/statistical group fairness metrics as definitions for both training 'fair' classifiers but also as our method of measuring fairness costs within the predictions as well as evaluating the accuracy-fairness trade-off of our classifiers.

Recall in Section 4.9, discussing the limitations of solely relying on statistical measures of fairness and as a solution within Section 4.10 we propose possible alternative definitions of fairness such as Causal, Counterfactual and Individual Fairness that aim to combat the suggested short-comings. Using this, an obvious extension of this thesis would be to include these additional fairness constraints to train classifiers on and measure fairness costs against. Methods introduced within [14], [22], [20] would allow us to implement each of these definitions within the optimisation problem during model training and metrics such as the 'Lipschitz Condition' can be used to measure the fairness cost between individuals.

This would be an interesting extension as the impossibility result only includes mutual exclusivity between statistical fairness metrics but does not necessarily state if it holds with respect to any of the alternative definitions. Though some of the alternatives have been able to defined as equivalencies to observational metrics such as Demographic Parity [20].

7.2.2 Human in the Loop and Perceived Fairness

While we are able to use disparities within some measurement criteria as a representation of (un)fairness between groups, however as we have mentioned, fairness cannot be defined within strict formal boundaries and is often the case that certain definitions may be more relevant in different scenarios. Thus deciding whether or not a prediction made by a machine is fair will ultimately lay upon the human stakeholders which will be very much contextually dependant on the specific problem case. We thus ask the question and/or suggest the possibility of using a 'human in the loop' as a fairness definition in itself that is able to be used as a feedback mechanism within a machine learning optimisation problems as well as a method of measurement of (un)fairness in itself. [24] uses the University admissions case to provide insight between what he describes as 'perceived'(human) and 'factual'(formal) faces of fairness. The paper focuses on different definitions and measurements of 'perceived' fairness and how perceived fairness may fluctuate depending on the context of when and on what basis it is being asked(exit, voice, organizational reputation).

7.2.3 From Simulated to Real Scenarios

Although we included experiments that modified our simulated data-set in order to reflect the distribution of independent and state school students within the UK, the feature generation process was entirely trivial and assigns very assumptive and most likely unreliable correlations/relationships between features and the targets. The data-set itself can only be used as a *very loose* abstraction of reality that removes a significant amount of complexity and information that would apply to a real University admissions scenarios simply from having a very small set of features and relatively low amounts of noise. Clearly, a clear future direction for this thesis would be able to closely replicate the results within a real data-set of University admissions. However, this comes with challenges that may make this experiment infeasible. Ethical issues concerning data-privacy arise considering the large amount of sensitive information regarding individual students and their socio-economic background may be included within the data.

7.3 Concluding Remarks

To conclude, we provide some additional remarks regarding the results of this thesis and what they can and cannot be interpreted to show. As we have mentioned, our thesis successfully demonstrates the proposition of using 'fair' targets in order to mitigate the impossibility results of satisfying multiple statistical group fairness definitions simultaneously whilst maintaining utility. However, the impossibility result and the fairness definitions used within this thesis are all methods to measure (un)fairness rather than proving fairness itself and cannot be treated as an overarching proof of fairness. Similarly, a disparity in a certain fairness metric is also not sufficient to prove for unfairness given the fluctuation of what is considered fair for different contexts. We cannot treat any single definition as an appropriate definition of what is 'fair'. This is emphasised by the short-comings of some of the fairness definitions described in earlier chapters. Ultimately, solving fairness related issues within certain scenarios requires domain-specific knowledge that should be task-specific, where-as our thesis only uses implementations of measurements of what is fair in fixed statistical terms. Evaluating what is an appropriate accuracy-fairness trade-off can also be misleading as it is almost infeasible to account for the consequences of fairness constraints in real-world decisions, demonstrating the difficulty in being able to measure the trade-off itself, especially if certain trade-offs are weighted differently. For example, the fairness of a decision may depend on/account for the number of people become unemployed in trade for a specific reduction in disparity between admission rates between Independent and State school educated groups. How would an admissions board ethically decide what are reasonable accuracy/fairness trade-offs in the first place given that someone has to be worse off? Clearly, deciding whether a decision is fair or not can quickly become intractable given the complexity of the decision itself. The short-comings of formal fairness definitions does not invalidate our thesis proposal itself. Demonstrating the effectiveness of choosing appropriate target variables within automated decisions opens the door for creativity and further exploration of how to make 'fair' machine learning models. Though the task of finding a goal with equal base rates is not a trivial exercise, however, it may be reasonable to assume that if a decision problem inherently prioritises fairness as an objective, then a goal that satisfies this criteria should exist. Our results give incentives for stakeholders to consider and experiment with goals that maximally proximate their actual motives, ultimately yielding fairer and more equitable outcomes to both the decision maker and also the individuals whom the decisions directly impact.

Bibliography

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification, 2018.

[2] Solon Barocas and Moritz Hardt. Fairness in machine learning nips 2017 tutorial — part i, 2017.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[4] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

[5] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018.

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art, 2017.

[7] Philip Bobko and Philip Roth. The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. *Research in Personnel and Human Resources Management*, 23:177–198, 06 2004.

[8] Toon Calders and Indre Zliobaite. Why unbiased computational processes can lead to discriminative decision procedures, 01 2013.

[9] Lu Cheng, Kush R. Varshney, and Huan Liu. Socially responsible ai algorithms: Issues, purposes, and challenges, 2021.

[10] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

[11] Social Mobility Commission. *Elitism in Britain*. UK Government, 2019.

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230, 2017.

[13] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[14] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. *ArXiv*, abs/1104.3913, 2012.

[15] Anthony Flores, Kristin Bechtel, and Christopher Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.". *Federal probation*, 80, 09 2016.

[16] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness, 2016.

[17] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning, 2018.

[18] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[19] Surya Mattu Julia Angwin, Jeff Larson and ProPublica Lauren Kirchner. Machine bias there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[20] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning, 2018.

[21] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016.

[22] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2018.

[23] Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ml's impact disparity require treatment disparity?, 2019.

[24] Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. Implications of ai (un-)fairness in higher education admissions: The effects of perceived ai (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 122–130, New York, NY, USA, 2020. Association for Computing Machinery.

[25] Richtel Matt. How big data is playing recruiter for specialized workers. *The New York Times*, 2013.

[26] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[27] Charles E. Mitchell. An analysis of the u.s. supreme court's decision in ricci v. destefano: The new haven firefighter's case. *Public Personnel Management*, 42(1):41–54, 2013.

[28] Arvid Narayanan. Tutorial: 21 fairness definitions and their politics. Youtube, March 2018.

[29] northpoint. Compas risk scales : Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county. In *volaris*, 2016.

[30] Executive Office of the President. *Big Data: a Report on Algorithmic Systems, Opportunity, and Civil Rights*. White House, 2016.

[31] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration, 2017.

[32] Greg Restall. Substructural Logics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2018 edition, 2018.

[33] Debi Saini. Book review: Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing, by dan biddle (gower publishing limited, alderhshot, 2005). *Vision—The Journal of Business Perspective*, 10:113–114., 01 2006.

[34] Harini Suresh and John V. Guttag. A framework for understanding unintended consequences of machine learning, 2020.

[35] S. Verma and J. Rubin. Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7, 2018.

[36] Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness-aware classification: Criterion, convexity, and bounds. *CoRR*, abs/1809.04737, 2018.

[37] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, Apr 2017.

[38] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2017.

[39] Rebecca Zwick, Lei Ye, and Steven Isham. Using constrained optimization to increase the representation of students from low-income neighborhoods. *Applied Measurement in Education*, 32(4):281–297, 2019.

Appendix A

Data-set Analysis of Changing Education System Distribution

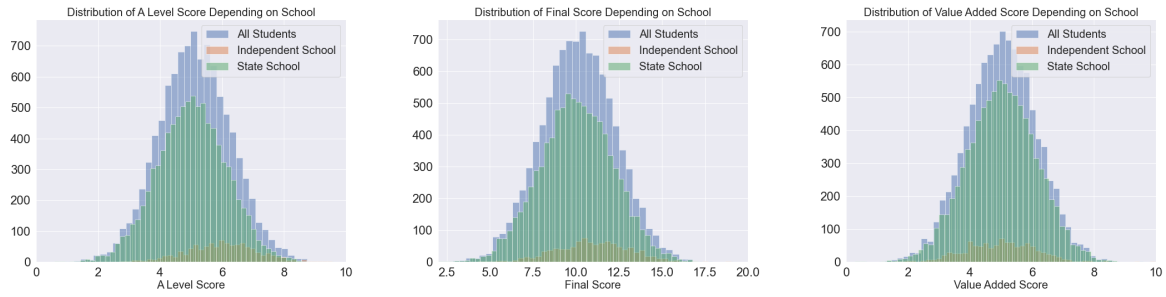

Figure A.1: Distributions of A-Level, Final Score and Value Added Scores Using a Bernoulli(0.9)

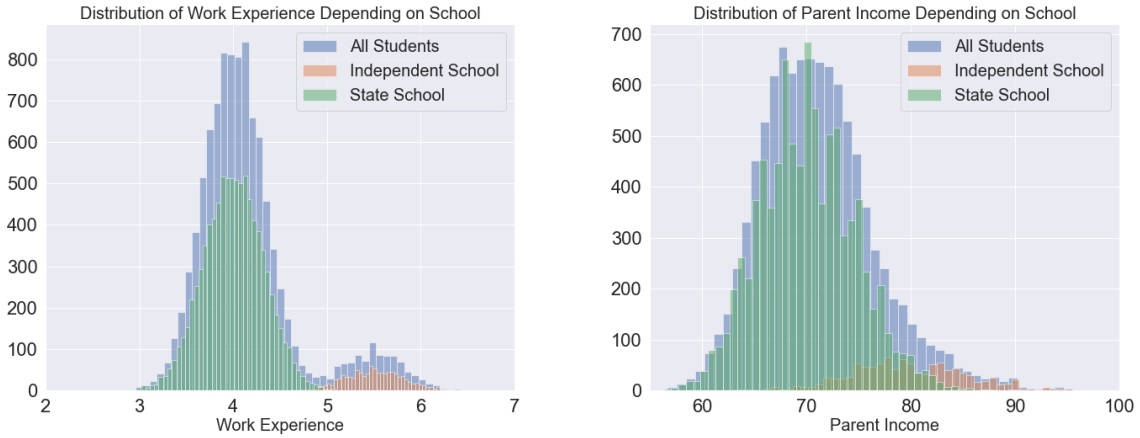

Figure A.2: Distributions of Work Experience and Parent Income Using a Bernoulli(0.9)

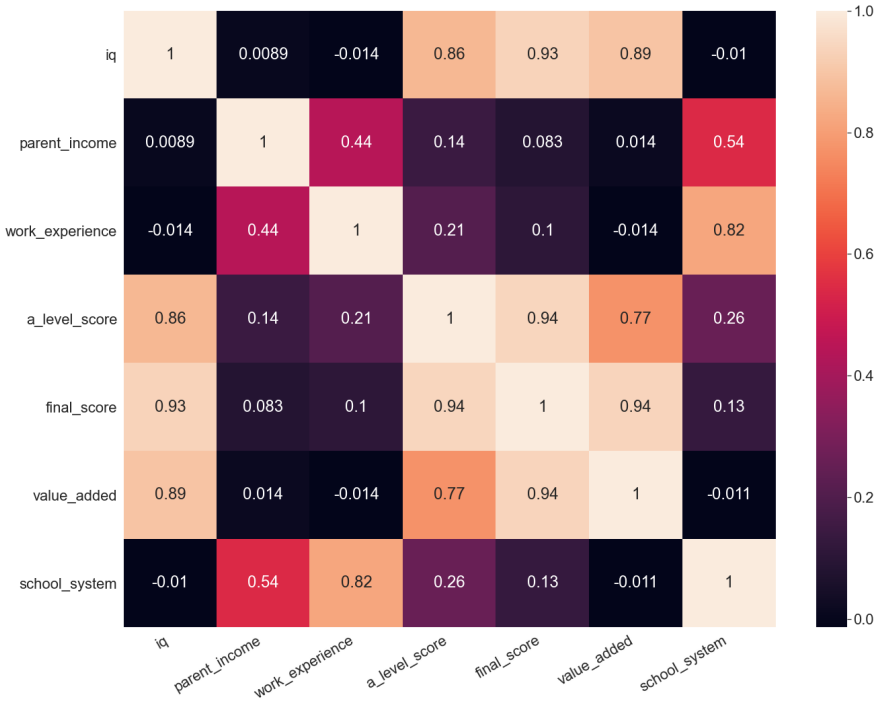

Figure A.3: Correlation Heatmap of Features

Appendix B

Unconstrained Classifier CDF Plots

B.1 Experiment 2

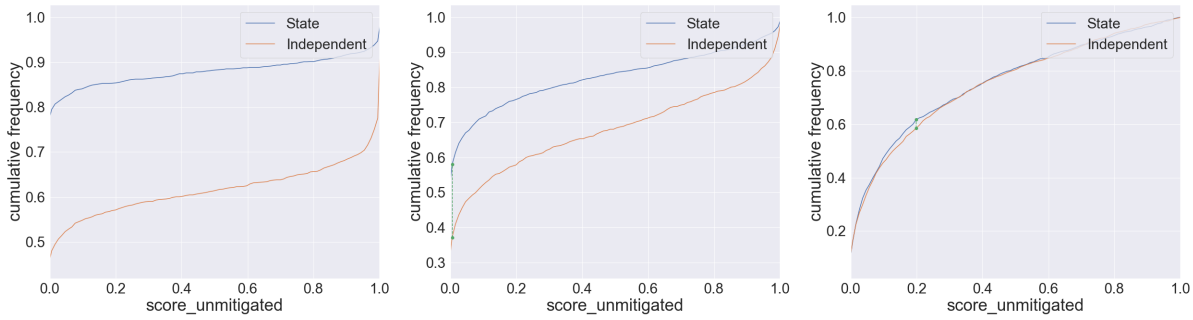

Figure B.1: Left: CDF Plots for Both Groups Predicted Scores Using A levels as the Training Goal), Middle: CDF Plots for Both Groups Predicted Scores Using Final Scores as the Training Goal, Right: CDF Plots for Both Groups Predicted Scores Using Value Added scores as the Training Goal

B.2 Experiment 3

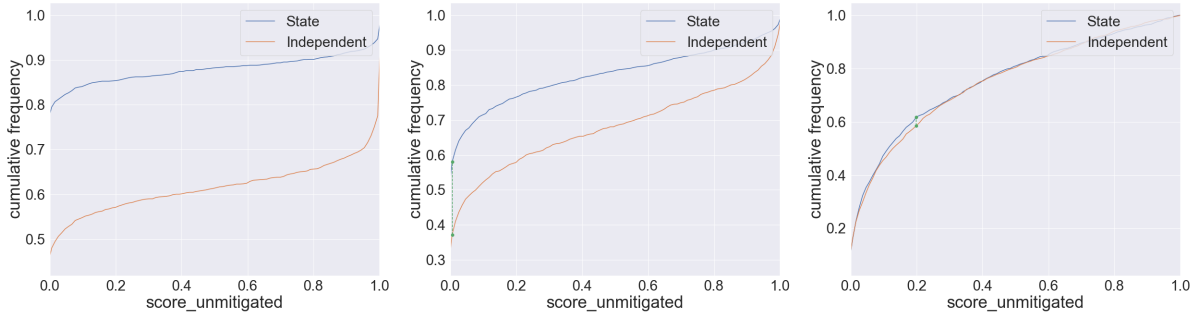

Figure B.2: Left: CDF Plots for Both Groups Predicted Scores Using A levels as the Training Goal), Middle: CDF Plots for Both Groups Predicted Scores Using Final Scores as the Training Goal, Right: CDF Plots for Both Groups Predicted Scores Using Value Added scores as the Training Goal

Appendix C

Experimental Results of Experiment 2,3

C.1 Experiment 2

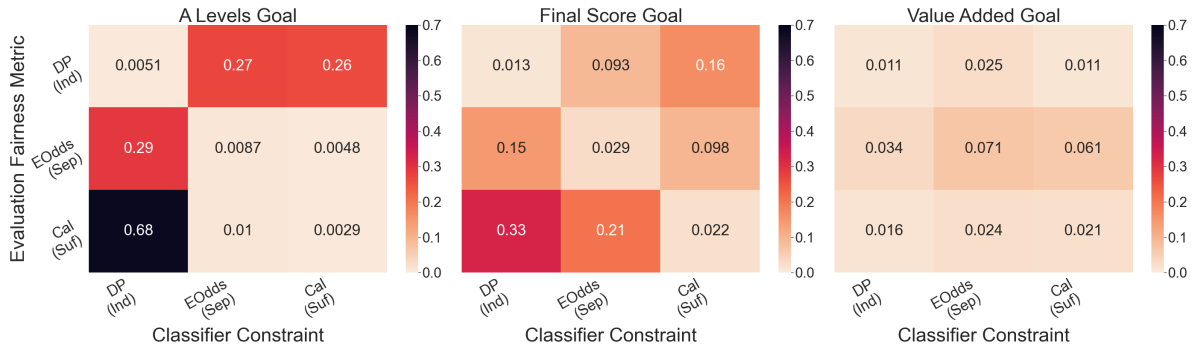

Figure C.1: Cost Gird of All Fairness Constrained Classifiers Trained w.r.t Cost of Three Fairness Metrics

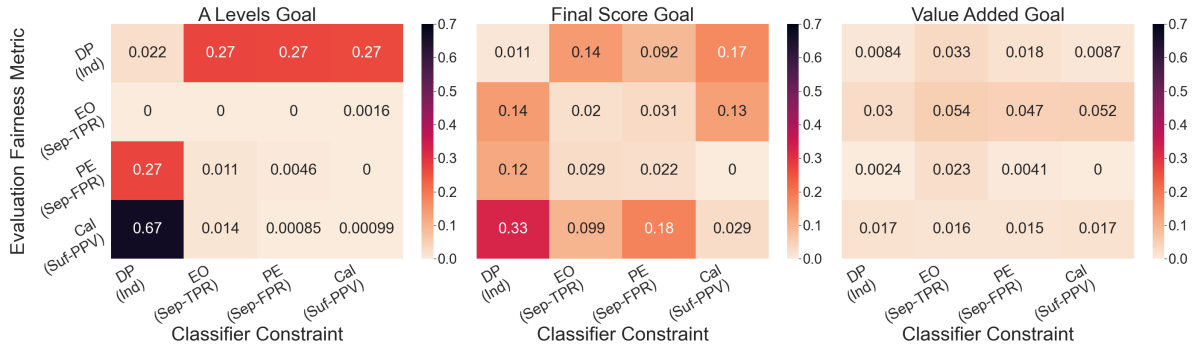

Figure C.2: Cost Grid of All Fairness Constrained Classifiers Trained w.r.t Cost of Three Fairness Metrics Including Relaxations

C.2 Experiment 3

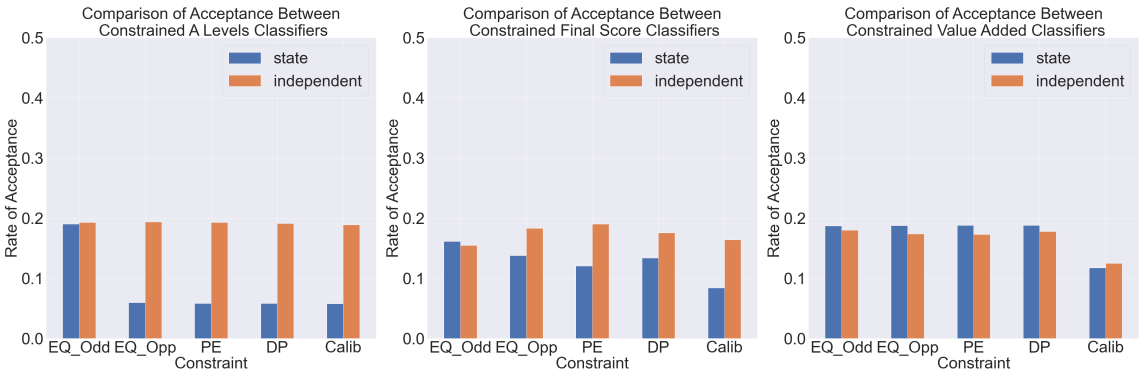

Figure C.3: Rate of Acceptance Between Groups for Each Goal w.r.t Each Fairness Constraint

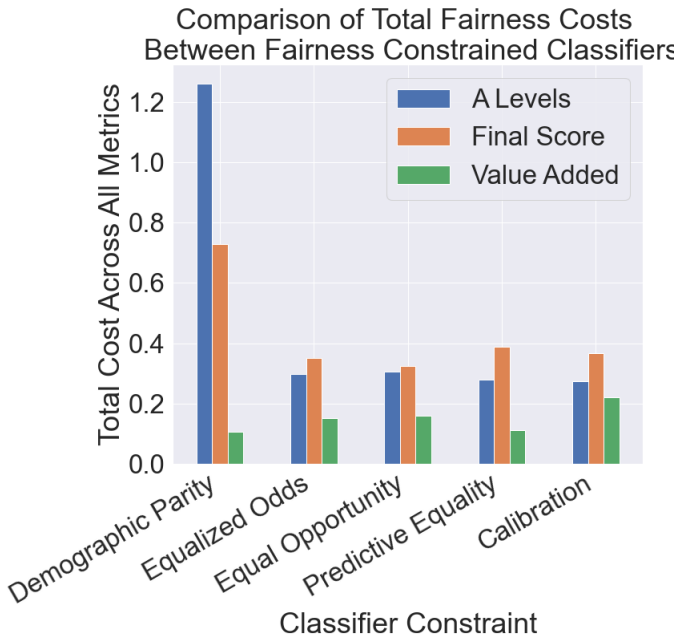

Figure C.4: Comparison of Total Cost of Fairness Across All Three Goals Across All Fairness Constraints

Figure C.5: Fairness Constraint Cost vs Mean Squared Error for Different Goals Trained Under the Same Constraint

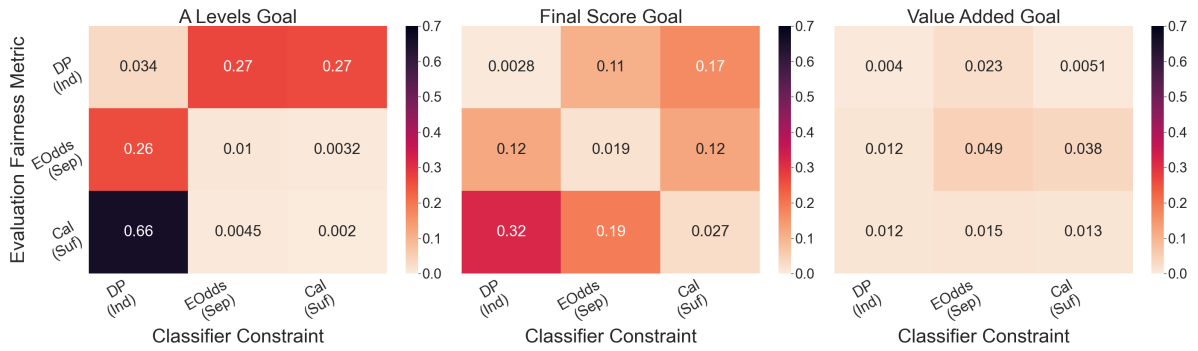

Figure C.6: Cost grid of All Fairness Constrained Classifiers Trained w.r.t Cost of Three Fairness Metrics

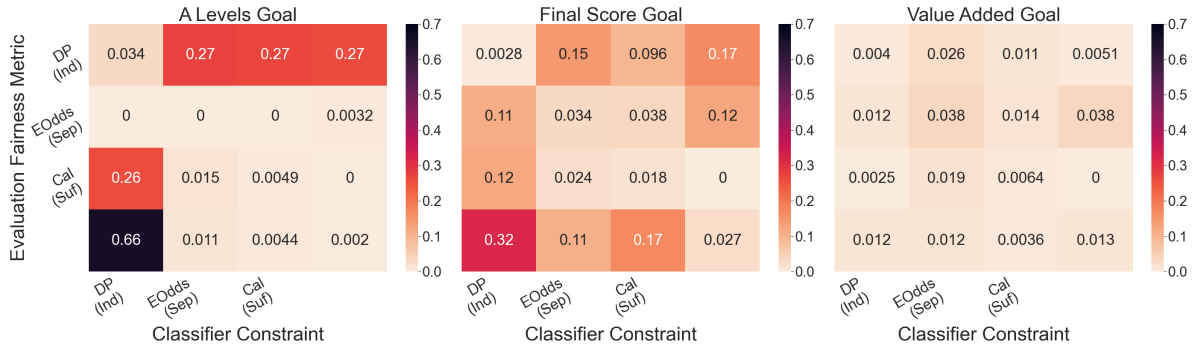

Figure C.7: Cost grid of All Fairness Constrained Classifiers Trained w.r.t Cost of Three Fairness Metrics Including Relaxations

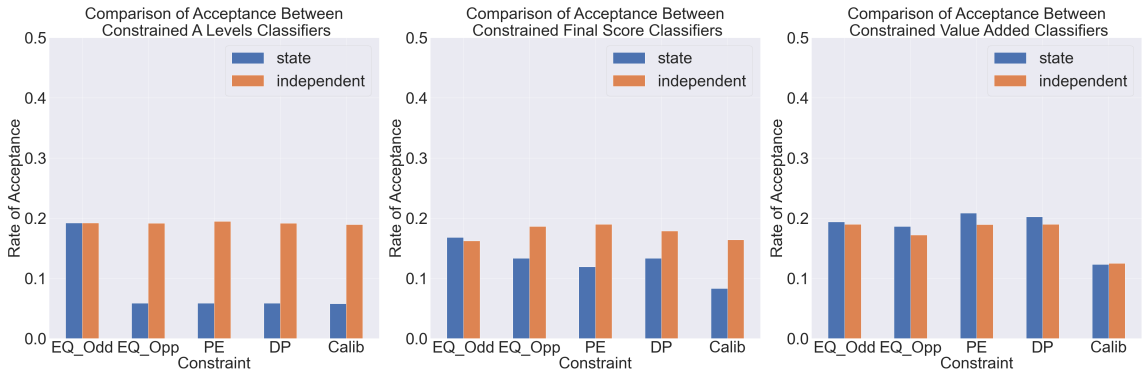

Figure C.8: Rate of Acceptance Between Groups for Each Goal w.r.t Each Fairness Constraint

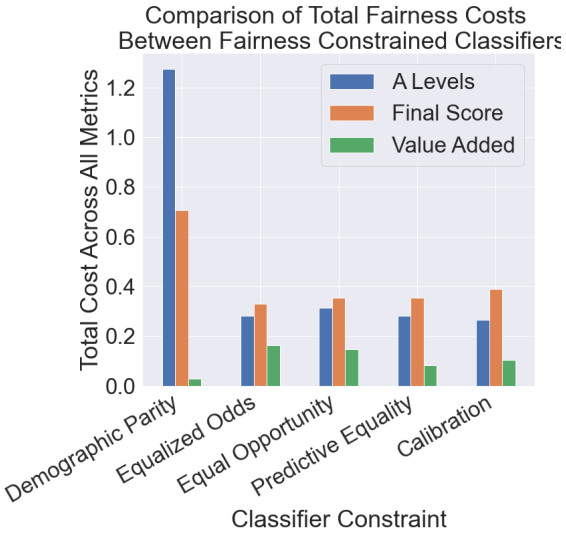

Figure C.9: Comparison of Total Cost of Fairness Across All Three Goals Across All Fairness Constraints

Figure C.10: Fairness Constraint Cost vs Mean Squared Error for Different Goals Trained Under the Same Constraint

Appendix D

Table of Results for Each Goal Under Each Fairness Metric

D.1 Experiment 1

D.1.1 A Levels

Table D.1: Fairness Costs of All Fairness Constrained Classifiers Using A Level Goal

Constraint	Equalized Odds				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.003	1.0	0.117	0.979	1.0
Independent	0.012	1.0	0.389	0.981	1.0
Disparity	0.009	0.0	0.271	0.00168	0.0

Constraint	Equal Opportunity				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0027	1.0	0.117	0.979	1.0
Independent	0.0118	1.0	0.389	0.981	1.0
Disparity	0.009	0.0	0.271	0.0017	0.0

Constraint	Predictive Equality				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.003	1.0	0.117	0.979	1.0
Independent	0.012	1.0	0.389	0.981	1.0
Disparity	0.009	0.0	0.271	0.0017	0.0

Constraint	Demographic Parity				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.302	1.0	0.382	0.301	1.0
Independent	0.012	1.0	0.389	0.981	1.0
Disparity	0.29	0.0	0.0063	0.68	0.0

Constraint	Calibration				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0	1.0	0.115	1.0	1.0
Independent	0.0	0.997	0.380	1.0	0.998
Disparity	0.0	0.0032	0.265	0.0	0.002

D.1.2 Final Score

Table D.2: Fairness Costs of All Fairness Constrained Classifiers Using Final Score Goal

Constraint	Equalized Odds				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.122	0.905	0.266	0.627	0.976
Independent	0.108	0.921	0.377	0.807	0.958
Disparity	0.013	0.015	0.111	0.181	0.018

Constraint	Equal Opportunity				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.092	0.889	0.239	0.685	0.973
Independent	0.109	0.926	0.379	0.807	0.961
Disparity	0.017	0.037	0.140	0.122	0.012

Constraint	Predictive Equality				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.127	0.944	0.278	0.627	0.986
Independent	0.107	0.919	0.375	0.810	0.957
Disparity	0.020	0.026	0.098	0.183	0.029

Constraint	Demographic Parity				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.192	0.974	0.336	0.534	0.993
Independent	0.070	0.862	0.332	0.858	0.932
Disparity	0.122	0.112	0.004	0.324	0.061

Constraint	Calibration				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0	0.882	0.163	1.0	0.974
Independent	0.0	1.000	0.331	1.0	1.000
Disparity	0.0	0.118	0.168	0.0	0.026

D.1.3 Value Added

Table D.3: Fairness Costs of All Fairness Constrained Classifiers Using Value Added Goal

Constraint	Equalized Odds				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.273	0.894	0.432	0.529	0.952
Independent	0.270	0.871	0.421	0.520	0.944
Disparity	0.004	0.023	0.011	0.009	0.008

Constraint	Equal Opportunity				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.249	0.873	0.409	0.546	0.945
Independent	0.228	0.835	0.381	0.551	0.933
Disparity	0.021	0.038	0.028	0.005	0.012

Constraint	Predictive Equality				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.233	0.854	0.392	0.558	0.939
Independent	0.228	0.835	0.381	0.551	0.933
Disparity	0.005	0.019	0.011	0.006	0.006

Constraint	Demographic Parity				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.227	0.842	0.384	0.561	0.935
Independent	0.228	0.835	0.381	0.551	0.933
Disparity	0.002	0.007	0.003	0.010	0.002

Constraint	Calibration				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0	0.955	0.244	1.0	0.985
Independent	0.0	1.000	0.251	1.0	1.000
Disparity	0.0	0.045	0.007	0.0	0.150

D.2 Changing Distribution of School Systems

D.2.1 A Levels

Table D.4: Fairness Costs of All Fairness Constrained Classifiers Using A Level Goal

Constraint	Equalized Odds				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0 0.994 0.211 1.0 0.998				
Independent	0.0 1.000 0.601 1.0 1.000				
Disparity	0.0	0.006	0.390	0.0	0.002

Constraint	Equal Opportunity				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0	0.994	0.211	1.0	0.9
Independent	0.0	1.000	0.601	1.0	1.0
Disparity	0.0	0.006	0.390	0.0	0.00

Constraint	Predictive Equality				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0 0.994 0.211 1.0 0.998				
Independent	0.0 1.000 0.601 1.0 1.000				
Disparity	0.0	0.006	0.390	0.0	0.002

Constraint	Demographic Parity				
Metric:	FPR	TPR	Selection Rate	PPV	NP
State	0.0	0.998	0.213	1.0	1.
Independent	0.0	0.393	0.236	1.0	0.
Disparity	0.0	0.606	0.024	0.0	0.4

Constraint	Calibration				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0	0.979	0.208	1.0	0.994
Independent	0.0	1.000	0.601	1.0	1.000
Disparity	0.0	0.021	0.393	0.0	0.006

D.2.2 Final Score

Table D.5: Fairness Costs of All Fairness Constrained Classifiers Using Final Score Goal

Constraint	Equalized Odds				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.110 0.930	0.30 0.719	0.977		
Independent	0.103 0.874	0.46 0.880	0.892		
Disparity	0.007	0.056	0.16	0.161	0.085

Constraint	Equal Opportunity				
Metric:	FPR	TPR	Sele		
State	0.113 0.942	0.305 0.716	0.981		
Independent	0.114 0.894	0.475 0.871	0.906		
Disparity	0.001	0.048			

Constraint	Predictive Equality				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.114 0.943	0.307 0.714	0.981		
Independent	0.109 0.887	0.469 0.876	0.902		
Disparity	0.006	0.056	0.163	0.162	0.079

Constraint	Demographic Parity				
Metric:	FPR	TPR	Sel		
State	0.115 0.942	0.307 0.712	0.981		
Independent	0.017 0.715	0.340 0.973	0.800		
Disparity	0.098	0.227			

Constraint	Calibration				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0 0.841	0.195 1.0	0.954		
Independent	0.0 1.000	0.463 1.0	1.000		
Disparity	0.0	0.159	0.268	0.0	0.046

D.2.3 Value Added

Table D.6: Fairness Costs of All Fairness Constrained Classifiers Using Value Added Goal

Constraint	Equalized Odds				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.156	0.790	0.313	0.624	0.925
Independent	0.152	0.738	0.337	0.691	0.875
Disparity	0.004	0.053	0.024	0.067	0.050

Constraint	Equal Opportunity			
Metric:	FPR	TPR	Selection Rate	PPV
State	0.179 0.826	0.339 0.603	0.935	
Independent	0.197 0.767	0.377 0.642	0.882	
Disparity	0.019	0.059	0.039	0.039

Constraint	Predictive Equality				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.179 0.829	0.339 0.604	0.936		
Independent	0.170 0.767	0.359 0.675	0.885		
Disparity	0.008	0.062	0.020	0.072	0.051

Constraint	Demographic Parity			
Metric:	FPR	TPR	Selection Rate	PPV
State	0.178	0.824	0.338	0.604
Independent	0.184	0.777	0.371	0.661
Disparity	0.006	0.048	0.034	0.058

Constraint	Calibration				
Metric:	FPR	TPR	Selection Rate	PPV	NPV
State	0.0	0.839	0.207	1.0	0.95
Independent	0.0	1.000	0.316	1.0	1.00
Disparity	0.0	0.161	0.108	0.0	0.05

Appendix E

Fairness Literature

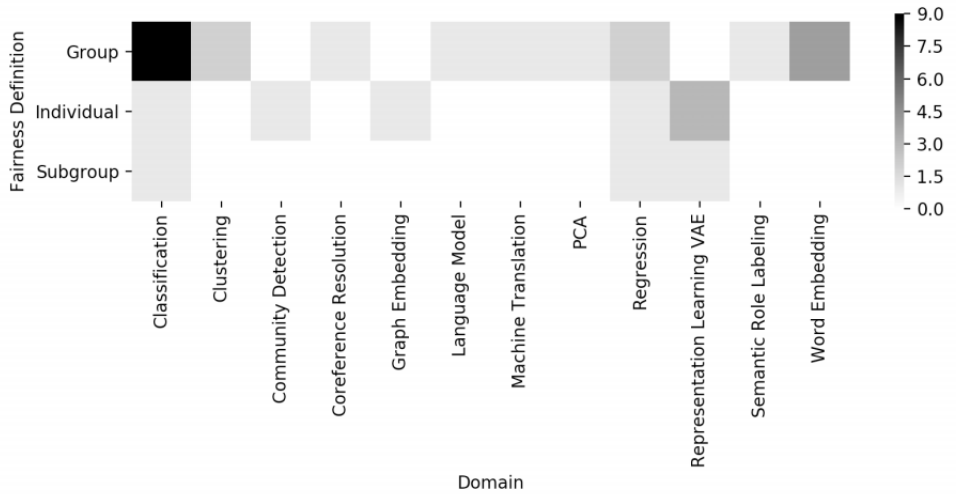

Figure E.1: Previous contributions in the field of fairness and machine learning processes, grouped by domain and fairness definition. [26]

Appendix F

Avenues for Discrimination

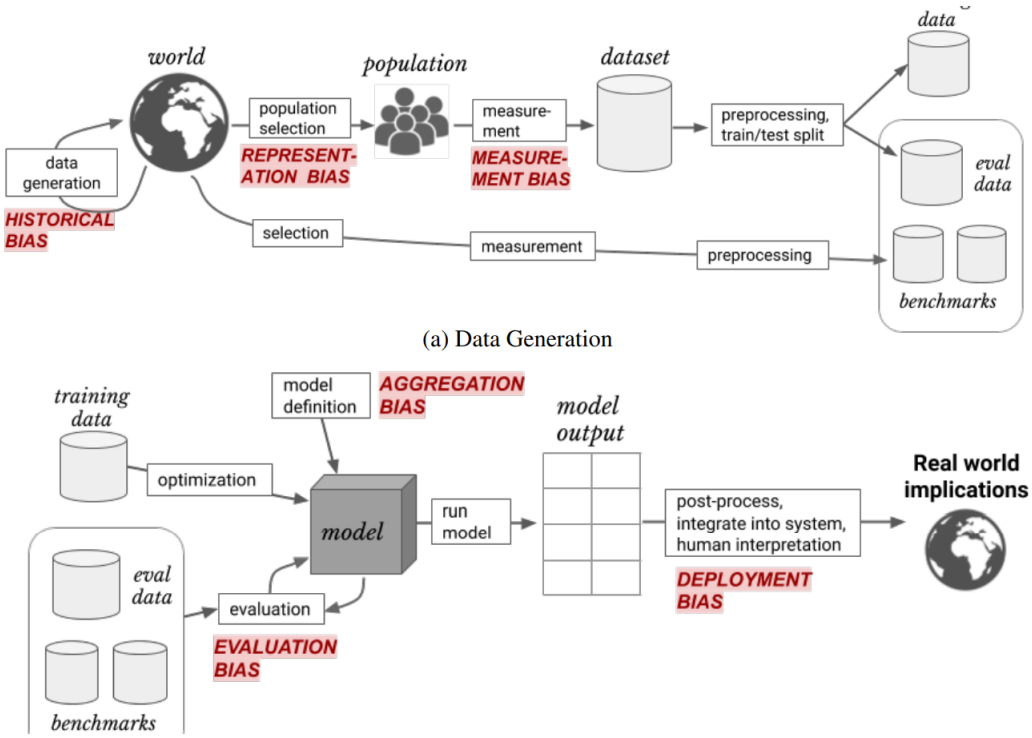

Figure F.1: Avenues of Discrimination within the Machine Learning Process [34]