MASTER

Assessing Bias and Fairness in Machine Learning through a Causal Lens

Allahabadi, Himanshi

Award date:
2020

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science
Data Mining Research Group

Assessing Bias and Fairness in Machine Learning through a Causal Lens

Master Thesis

Himanshi Allahabadi

Supervisor:
Dr. Mykola Pechenizkiy
Advisor:
Hilde Weerts

Eindhoven, October 2020

Abstract

Machine learning algorithms have garnered much attention and excitement in the recent past for solving problems in high dimensional spaces. As the number of real-world applications of machine learning grows, it is imperative to understand and curb the effects of existing biases in decision-making systems. Statistical fairness measures are not sufficient to interpret the fairness of algorithms, and quantitative fairness methods that depend on statistical correlations may not be appropriate to use when the discrimination arises from unfair data-generative processes. In this study, we discover biases in data, and explore fairness methods to mitigate their discriminatory effects. We draw conclusions about the aptness of the fairness methods for the biases identified in the data, and highlight their limitations. Finally, the need for counterfactual fairness is identified in lieu of the limitations.

Preface

The problem of fairness in machine learning is important and complex. There is no dearth of solutions in this problem space, and trying to wade through the vastness of fairness issues, fields of work, and algorithmic solutions has been a challenge. The controversial nature of this topic of fairness has often led to discussions with friends, which almost always end with ”it’s more complicated than that”. The area of causality piqued my interest a year ago, and as an overly eager, slightly reckless student, I dived headfirst into uncertainty and triggered what seemed like an infinite loop of exploration of solutions and their inherent limitations. In hindsight, it’s immensely rewarding to think about not just the outcome of this work, but the awareness and knowledge of the subject that I gained through this process. The lessons in independent research work are valuable in their own right, that had an impact on my values and growth.

I express my gratitude towards my supervisor, Mykola, for his feedback and ideas on my project. The questions raised by him have been extremely insightful for learning how to solve research problems. I’m also grateful to Hilde, for checking in on me regularly, for letting me talk all the ideas through, and for the conversations we had about fairness.

I’m incredibly lucky to have a family that believes in me and my dreams. Words fall flat in describing the strength I draw from my mother. This acknowledgement would be as incomplete without her mention as my life without her unending support. And last but far from the least, thank you Koen, for your faith in my capabilities. I couldn’t have done this without you.

The opportunity to engage your logical and creative faculties at the same time, to hurtle into unknowns and to come out with a different perspective is precious. In that sense, I hope I never stop being a student.

Learn why the world wags and what wags it. That is the only thing which the mind can never exhaust, never alienate, never be tortured by, never fear or distrust, and never dream of regretting. T. H. White, The Once and Future King

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the year 2016, a criminal case in the Supreme Court of Wisconsin led to a major discussion on the legality of using a risk assessment software called COMPAS (Har (2017)). The defendant of the case was African-American Eric Loomis, who was arrested for operating a vehicle that had been used in a shooting, and pleaded guilty for lesser charges of driving a stolen vehicle and dodging a police officer. He was sentenced to 6 years in prison, a decision made partly on the basis of his prior arrests, and partly based on the high-risk assessment made by COMPAS. In the same year, ProPublica analyzed the COMPAS assessments for >7000 cases, and claimed that the scores were racially biased, based on the difference of false positive and false negative rates across racial groups (Larson et al. (2016)). Developers of the algorithm at Equivalent refuted these claims, based on the high level of calibration between the racial groups (Brennan et al. (2009)). Additionally, it was provided that equalized false positive rates, equalized false negative rates and predictive parity are mutually incompatible, and thus nearly impossible to satisfy simultaneously (Angwin and Larson (2016)). Equivalent's refusal to share the actual algorithm prevents transparancy in their assessments of individuals, and presents a challenge in making assertions about how the algorithm works. Further, their rebuttal does not include arguments as to which fairness definition was chosen in the design of their algorithm, or why their predictions are beneficial for the justice system as a whole.

Several quantitative notions exist for algorithmic fairness based on metric similarity, including demographic parity (*statistical parity*), predictive parity (*calibration*), equalized odds and equality of opportunity. However, there is a disconnect between such group-based metrics and a definition of fairness centered on justice and equity. The word 'fairness' itself is contentious and abstract. Bias in algorithmic predictions can be due to a number of different sources. The COMPAS debate tend to focuses overtly on two different and incompatible statistical fairness measures, and ignores a discourse on the **nature** of bias within the system. For instance, in case the dataset is inadequate for representing different demographic groups, frequency-based correlations could be made by the algorithm. In contrast, for data whose generative processes involve more complex societal systems, inherent biases from the data can seep into model predictions. Examples of inherent bias in data generative processes have been elucidated under the concept of **redlining**, which is defined as the denial of services to residents of specific areas based on racial or ethnic factors. The intentional denial of loans to individuals based on their race is an example of disparate treatment, which can be represented as a direct effect between race and outcome. In contrast, disparate impact can be caused by an indirect effect of race on outcome through a proxy, such as average income within the area. Bias could also be added in the system due to the influence of any unobserved variables on the observed variables.

1.1 Research question

Methods in fair machine learning can define fundamentally different objectives to achieve algorithmic fairness. Further, the differences in nature of bias present in data influence its distribution, and consequently, the performance under different fairness objectives. For instance, a method that removes discrimination by compensating for class imbalance will be appropriate for preventing the representation bias from seeping into the algorithmic predictions. However, it might not appear to be as effective in cases where the generative mechanisms are responsible for creating algorithmic bias. Therefore, inquiry into the type of bias present could provide relevant information for the critique and comparison of fairness methods. Thus, we formulate our research question as follows: does the nature of bias present in data impact the relative statistical fairness performance of algorithmic fairness methods?

For this research, we focus on the two types of bias mentioned above, and two specific fairness methods from literature: re-weighing(Calders et al. (2009)) and optimized pre-processing(Calmon et al. (2017)). The question can be broken down into the following sub-goals:

- Investigation of the datasets for presence of bias, particularly arising due to unfairness in the underlying data-generative mechanisms
- Application of fairness preprocessing methods on the data and computation of statistical fairness metrics on predictions from a classifier trained with the preprocessed data

Ultimately, to answer the research question, results from the experiments are assimilated and fairness methods are evaluated in the view of biases found in the raw datasets.

1.2 Approach

Two real-world datasets were used for this research, namely, COMPAS recidivism and UCI Adult Income. Both datasets contain gender and race in their attributes. Representation bias is identified on the basis of class imbalance. Examining the datasets for inherent bias requires an experimental study. In Chapter 4, we describe a causal discovery framework based on Ke et al. (2019) to test for the presence of cause-effect dependencies between the protected attribute and the outcome.

For the second sub-problem, we chose two data pre-processing methods for fairness:

- Re-weighing attempts to reduce statistical dependence of outcome on the sensitive attribute by assigning weights to compensate for the bias
- Optimized pre-processing reduces the conditional dependence of outcome on sensitive attribute, subject to constraints on preservation of data utility and individual fairness.

The transformed datasets are used to train a logistic regression model. Comparisons are drawn between their performance for the two different sensitive attributes, race and gender. Further, an analysis of their trends is done to gauge the behaviour of the classifier with each set of processed data. This trend analysis aids the evaluation of the methods in the presence of the two different types of bias.

1.3 Results

The two approaches to fairness pre-processing show different trends in performance metrics under the presence of class imbalance and causal relationships between the specified sensitive attribute and the outcome. Specifically, re-weighing performs well for the class imbalance problem, while the probabilistic approach to discrimination control in optimized pre-processing performs well in the presence of a direct effect from the protected attribute. For the Income dataset, there is a clear tradeoff between accuracy and fairness for the two methods. However, for the indirect effect in COMPAS, the trends are not as clear, since neither of the two methods were designed for this specific problem. We identify the need for counterfactual reasoning to effectively mitigate unfairness arising through such indirect causal effects.

Chapter 2

Background

The successes of machine learning algorithms for solving problems in high dimensional spaces has garnered much attention and excitement in the recent past. As the number of real-world applications of machine learning grows, it is imperative to understand existing biases in formal decision-making systems, so that appropriate measures can be taken to constrain such biases from trickling into algorithms (Loftus et al. (2018)). Research into these problems is qualified under the umbrella term of algorithmic fairness. Algorithmic fairness encompasses the problems of mitigating different types of biases, as well as defining appropriate metrics of success for fairness. Bias in data can exist in a variety of ways, some of which can lead to discrimination in machine learning. In this chapter, we introduce the terms necessary to understand the context and goals of our research in algorithmic fairness. Datasets used and mentioned throughout the work are real-world observational datasets, namely the Adult Census Income and COMPAS recidivism datasets. Both datasets contain the sensitive attributes gender and race. Each case studied in this work is a combination of each dataset with each of the sensitive attributes. We first define two fundamentally different types of bias in data. Then, we branch into statistical fairness and its metrics. Fairness is also discussed from the viewpoint of causality, which provides perspective on the limitation of statistical fairness metrics for interpretation of fairness. A primer on causality is provided, as necessary for understanding the methodology used for discovery of bias.

2.1 Representation Bias

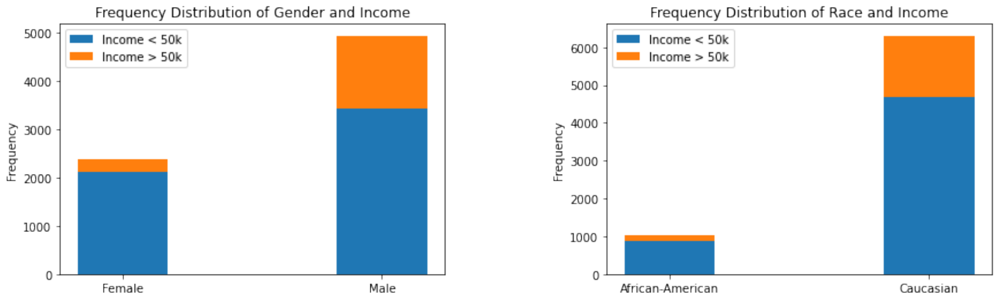
Representation bias arises from the way data is sampled from a population Schölkopf (2019). For instance, a lack of demographic diversity in datasets can lead to biases in favour of the better-represented demographic group. In essence, representation bias can be characterized by human prejudices seeping into dataset while defining the samples. As mentioned, an example of such a bias is imbalanced class representation. For a binary setting with sensitive attribute A, if the number of observations in class $A = a$ is disproportional to the number of observations in class $A = a'$, the dataset is said to have a class imbalance. A related metric is the the disparity between the probabilities of positive(negative) outcomes for the two groups. Due to error-prone sampling from a population, differences can arise in the distribution of outcomes for the two classes. These differences can be reinforced by algorithms as bias against the under-represented class. To infer whether a dataset is imbalanced, we computed the following metrics for each combination of dataset and protected attribute:

1. Imbalance in representation of groups, as the ratio of their frequencies of observations

2. Imbalance in outcomes, as the ratio of the probability of positive outcomes across the groups.

The next section presents these metrics for Adult Income and COMPAS datasets.

2.1.1 Data Imbalance in Income



(a) Distribution by Gender and Income

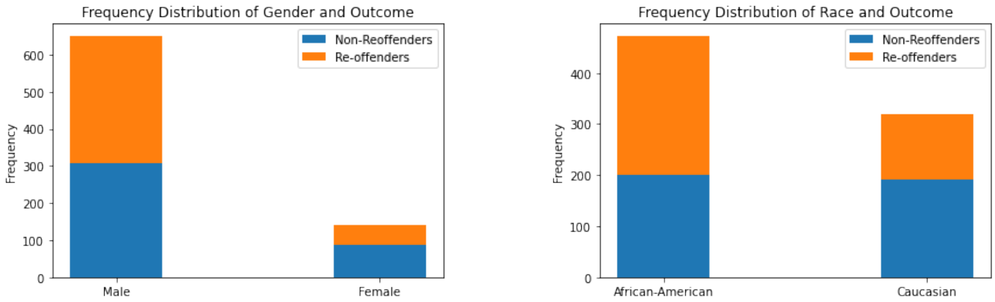
(b) Distribution by Race and Income

Figure 2.1: Bar plots depicting frequency distributions for group membership and outcomes in the Adult Income dataset. The gap in representation of race groups is bigger than the gap in representation of gender groups. However, the difference of positive outcome rates between the gender groups is larger than that between the race groups.

Figure 2.1 plots the frequency distributions of the raw dataset. Each case indicates the imbalance between protected and privileged group membership. The gap in representation is wider for race than for gender. Further, the rate of positive outcomes, i.e. the odds of earning more than 50k USD; was found to be higher for males than for females, and higher for Caucasians than for African-Americans, with the gap being wider for gender groups than for racial groups.

2.1.2 Data Imbalance in COMPAS

Figure 2.2 plots the frequency distributions of the raw dataset. Here, the privileged gender group is Female, as opposed to the setting in the income data. The gap in representation is wider for race than for gender. Further, the rate of positive outcomes, i.e. the odds of being a non-reoffender; were found to be higher for females than for males, and higher for Caucasians than for African-Americans, with the difference being almost the same for gender groups as for racial groups.



(a) Distribution by Gender and Recidivism

(b) Distribution by Race and Recidivism

Figure 2.2: Bar plots depicting frequency distributions for group membership and outcomes in the COMPAS dataset. The gap in representation of gender groups is wider than the gap in representation of racial groups. The difference between rates of positive outcome (non-reoffender)across the race groups is larger than the difference between success rates across racial groups.

2.2 Historical Bias

Bias in data can also exist in the form of unfair causal pathways in the generative mechanisms, between the protected attribute and the outcome. This can be termed as historical bias, owing to that fact that it is already existing within social and technical structures (Hoffmann (2019)), and can seep into from the data generation process despite perfectly randomized sampling methods (Suresh and Guttag (2019)). Thus, the essential characteristic of this bias, which also differentiates if from representation bias, is that it is a reflection of the society, and affects how the ground truth appears to be. An example of this bias is the illegal practice of redlining, in which mortgage lending or insurance services are restricted for communities from specific demographics. In this case, the area zip-code serves as a *proxy* for racial/ethnic groups. This creates disparate impact, which is different from a direct denial of services based on racial/ethnic features, known as disparate treatment.

In the next section, we discuss two different viewpoints of fairness which helps us better understand the two biases, for the ultimate goal of mitigation of discrimination arising from them.

2.3 Statistical Fairness Measures

Demographic parity or statistical parity suggests that a predictor is unbiased if the prediction $\hat{y}$ is independent of the protected attribute p so that:

$$P(\hat{y}|p) = P(\hat{y})$$

Here, the same proportion of each population are classified as positive. However, this may result in different false positive and true positive rates if the true outcome y does actually vary with the protected attribute p. Deviations from statistical parity can be measured by the statistical parity difference:

$$P(Y = 1|D = \text{unprivileged}) - P(Y = 1|D = \text{privileged})$$

It can also be measured by the disparate impact:

$$\frac{P(Y=1|D=\text{unprivileged})}{P(Y=1|D=\text{privileged})}$$

These metrics are commonly used to measure discrimination. Other measures of statistical fairness include average odds difference, which measures the differences of false positive rate and true positive rate between privileged and protected groups:

$$\tfrac{1}{2}\left[(FPR_{D=\text{unprivileged}} - FPR_{D=\text{privileged}}) + (TPR_{D=\text{unprivileged}} - TPR_{D=\text{privileged}}))\right]$$

Fairness has been increasingly studied as an essential component of machine learning systems (Oneto and Chiappa (2020), Courtland (2018), Mehrabi et al. (2019)). However, there is often a lack of consensus on the appropriate measure of fairness, due to conflicts in the structural nature of underlying data-generating mechanisms, as well as philosophical disagreements. According to Chen (2019), a predictor is said to achieve fairness through unawareness when sensitive attributes are not explicitly used for prediction. While the predictor aims for group-level fairness, it does not guarantee non-discrimination. Due to subtle correlations between protected attributes and a combination of other observed or unobserved features, discrimination by proxy may arise Datta (2017), Williams et al. (2018). For instance, hobbies might be an indicator of gender on resume screening systems, zip codes of current home or birthplace might reflect race, and so on.
In the presence of historic biases, it is possible to train a classifier to be fully accurate and also appear to be fair with respect to the statistical metrics, but in reality be far from fair, due to the causal dependency on the sensitive attribute, which may not be removed. This is explained theoretically in Chiappa and Isaac (2019) for COMPAS data. In the next section, we introduce the terminology needed to understand a causal view on fairness.

2.4 Causality

The pursuit to discover cause-effect relationships is posed with several challenges, one of the most common being observed or unobserved *confounding* (Dawid (2007)). Consider a set of variables $\{Z\}$ which have an impact on both- the dependent variable Y and the independent variable X under study. In the context of personalized medicine, the rate of success or failure of a treatment might be different for subgroups based on individual differences in:

- Genetics, affecting race/ethnicity
- Hormones and Adverse Drug Reactions, affecting gender/sex
- Social, economic and environmental factors.

Additionally, the outcome may be directly influenced by factors correlated to the above, such as access to healthcare and treatment. This led to the inception of the Randomized Control Trial RCT, wherein we control the value of the decision variable X, such as a treatment type, and study its effects on the probability distribution of Y.

2.4.1 Interventions

In the observational setting, we simply model the conditional distribution of y based on naturally occurring values of x in our data. The objective in traditional ML is to find the distribution $p(y|x)$ through optimization, and this suffices to give excellent results for applications related to diagnosis and forecasting. However, in some domains such as drug testing, online recommender systems and control problems, we wish to control the value of x to achieve a desirable effect on y (Dawid (2007)). In such a setting, the objective is to obtain $p(y|do(x))$, where $do(x)$ is the intervention through which the value of x is controlled. An intervention on a variable is different from filtering by or conditioning on that variable in that, all elements not caused by that variable should not be disturbed, i.e. have the same variation as before.

Consider a situation wherein there exists a statistical correlation between variable x and y, but the two are not causally related. The value of y can then be predicted by producing a mapping between x and y through traditional machine learning methods. However, the intervention on x has no bearing on the distribution of y, and consequently, this allows us to conclude that y is not an effect of x. In the observational setting, we cannot make such a statement. The use of an intervention thus ensures that the observed association between the intervened variable and the effect variable, can indeed be given causal interpretations.

Conducting a RCT to obtain interventional data may not always be possible, or even ethical. Several methods have been suggested to estimate the conditional distribution $p(y|do(x))$ based on observations that come from outside a controlled trial (Kusner et al. (2017), Kusner et al. (2018), Salimi et al. (2019), Chiappa (2019)).

2.5 Causal Bayesian Viewpoint on Fairness

The general objective of fairness is to take affirmative action to prevent discrimination, and to **justly** mitigate existing disparity or discrimination, which depends on the existing patterns of bias in the system. If algorithmic fairness is viewed with the single objective of boosting statistical fairness metrics, the perspective for interpretation of results gets limited. Explicit assumptions about cause-effect nature of data attributes helps to remove the ambiguity from statistical methods. Without prior knowledge of the pathways through which the sensitive attributes impact the outcome, it would be difficult to evaluate the behavior of algorithmic fairness methods beyond their performance on statistical fairness metrics. For instance, the unfairness caused by a direct effect of sensitive attribute on the outcome might be more easily mitigated than an indirect effect on the outcome through a non-sensitive attribute. For the latter, the effect of the non-sensitive attribute on the outcome would encode the unfair effect of the sensitive variable, along with information which is actually relevant for the predictor. Therefore, aptness of different fairness methods depends on their implications *for the given problem*, which requires some understanding of the patterns of bias within the datasets.

The inherent bias within data can be found through discovery of cause-effect mechanisms within the dataset. Graphical models can represent the distributions in a conveniently accessible manner. Causal models can be viewed as a special class of graphical models that not only represent the observational distribution, but also the distributions generated by intervening on the nodes. Thus, predictions for causal edges can be made through interventions.

2.6 Counterfactual Fairness

A variety of techniques have been availed to perform abduction-action-prediction(Shanmugam (2001)) in machine learning. A fair learning algorithm is proposed by Kusner et al. (2017), which utilises an assumed causal model to construct the conditional distribution of background variables given the training data. The predictor takes this distribution as input, along with all the non-descendants of the protected attribute. To approximate the said conditional distribution, they use the Markov Chain Monte Carlo sampling method. The requirement of the full causal model of underlying data generating processes poses practical challenges. For the task of fairness through causal reasoning, using such a method alone would require us to assume the exact causal model to use as an input. However, there can be disagreements, even within domain experts, on the exact causal relationships that presumably dictate the real world data-generating mechanisms (Russell et al. (2017)). Infact, one of the practical challenges in causal fairness is that it is difficult to reach a consensus in terms of what causal graph to use. An approach to mitigate this challenge is provided in Russell et al. (2017), where approximate counterfactual fairness is said to be achieved under causal uncertainty. Since different causal assumptions could hold true under different circumstances, they combine the outcomes given by multiple causal models. The work in Salimi et al. (2019) looks at the fairness problem as a task of database repair, i.e., given a fairness constraint and a database instance, perform a minimal set of updates such that the new database satisfies the constraint. This approach eliminates the need for access to the full underlying causal model.

In this research we limit our scope to causal viewpoint on fairness, which entails the problem of explanability of fairness solutions. In the next section, we explain the causal discovery method used.

2.7 Causal Model Discovery

Causal learning can help us exploit cause-effect information contained in joint probability distributions, which is largely ignored by statistical learning (Schölkopf (2019)). The most common types of causal models are Causal Bayesian networks and Structural Causal Models, but the two are closely linked: both are assumed to satisfy the causal Markov condition, which states that, conditioned on its parents, each variable is independent of its non-descendants. Further, both can be represented using Directed Acyclic Graphs, which simply make implicit causal assumptions explicit. This can be mathematically formalised as:

$$p(X_1, ...X_n) = \prod_{i=1}^{n} p(X_i|\mathbf{PA}_i) \tag{2.1}$$

Where $X_1, ..., X_n$ is a set of random variables and $\mathbf{PA}_i$ represents the parent nodes of X_i in the causal graph.

Many other factorizations exist to represent the joint distribution of observed variables. In those settings, changing one of the physical cause-effect mechanisms responsible for the statistical correlations in the data, can result in changes in multiple factors on the right hand side. Such a representation is thus entangled with respect to the statistical structure of the data and the factors of variation in it (Bengio (2009)). However, the aforementioned factorization is the only one that decomposes the joint distribution into causal conditional mechanisms which are responsible for all statistical dependence within the variables. For this reason, it is referred to as the disentangled factorization (Bengio (2009)). The assumption here is that the causal generative mechanisms within the observables of a system are composed of autonomous modules that do not inform or influence each other (Schölkopf (2019)). This has been referred to as the Independent Causal Mechanism Principle, and is popularly used to perform causality studies (Parascandolo et al. (2018), Lemeire and Dirkx (2006), Cai et al. (2019)). Intuitively, this principle is closely related to the Causal Markov assumption, where the independent modules are represented by the probability distribution of each observed variable conditioned on its parents. Based on these insights, a reasonable hypothesis is that changes in distribution are small if the data is represented

in the right space (Schölkopf (2019)).
The challenging task of learning causal models from data is the assumptions on their structure. For real-world datasets, this is especially a problem, since access to ground truth is uncertain. To overcome this issue, we adapt the causal discovery framework from Ke et al. (2019) so as to get some insight on the nature of the underlying generative mechanisms of dataset with an unknown ground truth causal model. The methodology of their research is described at length in the next section.

2.8 Meta-Transfer Objective for Causal Discovery

We begin by mentioning the assumptions required by the method used, and by related research which discover the correct underlying causal model amongst the possible choices, by measuring the speed of their adaptation to interventions (Bengio et al. (2020), Priol et al. (2020), Ke et al. (2019), Ke et al. (2020)).
The fundamental assumptions employed in the original meta-transfer learning framework for disentangling causal mechanisms are:

1. Data-generative processes consist of high-level mechanisms that are independent of each other (popularly known in the causality field as **Independent Causal Mechanisms principle**, Schölkopf (2019))

2. Very few high-level mechanisms need to change in order to adapt to perturbations in a transfer distribution caused by an intervention.

Meta learning is also known as *learning to learn*(Wang et al. (2017)), due to its capability of adapting to previously unseen tasks. This is utilised in this meta-transfer learning framework to learn the high-level representation of the attributes, as the factorization of their probabilities conditioned on their parents. The model consists of a learner and a ground-truth Structural Causal Model, each of which are parameterized by multi-level perceptrons (MLPs). Transfer distributions are generated by adding a small amount of noise in the probability distributions of a node (i.e. attribute), which is referred to as a soft intervention. Each transfer episode leads to an accumulation of the negative log likelihoods of variables under different different causal configurations, randomly sampled from the belief distribution, which is initialized such that it closely represents the input CPT.
Thus, the meta-transfer objective is to adapt faster to these interventional distributions, by learning causal pathways within the data-generative mechanisms. In each transfer episode, causal configurations in the form of adjacency matrices are sampled from a belief distribution. The belief is a representation of the probability of the relationship, representing confidence in the presence of edges in the graph. It is initialized so as to closely represent the input CPT. The adaptation trajectory consists of several such transfer episodes. Along each trajectory, the negative log-likelihood of the transfer distribution (under the sampled causal configurations in each episode) is accumulated, known as *transfer regret*. At the end of the trajectory, gradient of this accumulated negative log-likelihood is computed with respect to the edge belief parameters of the causal model. For a simple case for 2 variables, the transfer regret would be the following:

$$R = -log[\sigma(\gamma)L_{A\to B} + (1-\sigma(\gamma)L_{B\to A})] \tag{2.2}$$

where γ represents the structural parameters of the causal model, σ is the sigmoid function and $\sigma(\gamma)$ represents the belief in the edge $A \to B$. A larger gradient of the likelihood of $A \to B$ pushes weights towards $\sigma(\gamma)$, while a larger gradient of the likelihood of $B \to A$ pushes weights away from $\sigma(\gamma)$. Thus, we perform transfer learning with the objective to measure the speed of adaptation to interventional distributions after sparse changes to the model. The motivation for optimizing the adaptation rate to find the best causal structure comes from the mentioned assumption that the interventional changes in distributions are sparse when the knowledge about the distribution, in this case the cause and effect relations, have appropriately modularized representations.

2.9 Fairness Mitigation Methods

The methods chosen for this study are called **re-weighing** and **optimized pre-processing**. Both methods have a common general goal of controlling discrimination by reducing the dependency of the outcome on the sensitive attribute. The difference between them lies in their definition of and approach towards discrimination. Re-weighing views discrimination as the imbalance of the probabilities of positive outcomes between the protected and privileged groups, which can lead to a disparate impact on the outcome. The proposed solution is to assign weights to the tuples to compensate for the bias. In contrast, optimized pre-processing measures discrimination more generally, as the distance between the conditional probability distribution of the outcome given the sensitive attribute, from a target distribution of the outcome independent of the sensitive attribute. The proposed solution is to learn a randomized mapping from the original distribution, which is optimized for discrimination control (and a few other relevant constraints), and to apply it to the training dataset. The choice of fairness methods for this thesis was thus made with this difference in mind, which is in keeping with the distinction between biases in the dataset that we made in the previous chapter.

The remainder of this chapter provides general descriptions of the two fairness preprocessing methods, and presents the results obtained after applying the methods for controlling gender and racial biases for each dataset.

2.10 Reweighing

This approach introduced by Calders et al. (2009) reduces the dependence of the predictor variable on the sensitive attribute while maintaining the prediction accuracy. Here, discrimination is defined as:

$$P(Y = 1|A = 1) - P(Y = 1|A = 0), \tag{2.3}$$

where A represents the binary sensitive attribute and Y is the binary outcome. This discrimination measure is reminiscent of the demographic parity fairness definition, which posits that the predicted outcome should be independent of the sensitive attribute. The focus of this method is on achieving group fairness, by reweighing the observations. Higher weights are assigned to instances wherein $(Y = 1 \wedge D = 0)$ compared to $(Y = 0 \wedge D = 0)$. On the other hand, lower weights are assigned to $(Y = 1 \wedge D = 1)$ compared to $(Y = 0 \wedge D = 1)$. An advantage of this method is that it is non-invasive, since it does replace the outcome labels, rather, it multiplies the observation frequencies with different weights to learn an unbiased model.

2.11 Optimized Preprocessing

In this probabilistic method proposed in Calmon et al. (2017), along with discrimination control, attention is also given towards utility and individual-level sample distortion. They combine the three into a single optimization objective. The problem formulation is as follows: given a dataset of n independently and identically distributed samples $\{A_i, X_i, Y_i\}_{i=1}^{n}$ drawn from the joint distribution $p_{A,X,Y}$ where A is the sensitive variable, Y is the outcome and X are the remaining attributes, find a randomized mapping $p_{\hat{X},\hat{Y}|X,Y,A}$ such that the loss in utility, $\Delta(p_{\hat{X},\hat{Y}}, p_{X,Y})$ is minimized, subject to the following constraints:

1. **Discrimination control:** conditional probability $p_{\hat{Y}|A}$ is similar for any two values of A
2. **Individual distortion control:** the mapping is penalized for certain "large" changes to an individual's attributes.

The first constraint is applied to promote group fairness, while the second is for individual fairness.

2.12 Conclusion

Class imbalance can be caused by a non-randomized data sampling from populations, which can lead to representation bias in the sampled groups (Schölkopf et al. (2012)). However, algorithmic bias may not be always be due to unrepresentative data samples. The pathways of causal dependence between a protected attribute and outcome can signify unfairness patterns within underlying data-generative mechanisms, leading to historic bias. Learning from such an unrepresentative dataset can have a **disparate impact** on predictions. On the other hand, *unjustifiable* effects between sensitive attributes and outcomes lead to differences in false positive and negative rates in predictions. Therefore, the bias added to the algorithm through such pathways might require different methods to achieve algorithmic fairness than the ones that are useful for imbalanced datasets.

In this study, we focus on the relative performance of fairness mitigation methods, in the presence of the effects of class imbalance and unfair causal paths. Our hypothesis is that the approach of reweighing is more appropriate for curbing discrimination caused by representation bias, while optimized pre-processing would fare better when there is a direct causal relationship between sensitive attribute and outcome.

Chapter 3

Methodology

In Section 2.1.1, we assessed the datasets for class imbalance, caused by a non-randomized data sampling from populations. However, algorithmic bias may not be always be caused by prejudices seeping through the data sampling processes. The pathways of causal dependence between a protected attribute and outcome can signify unfairness patterns within underlying data-generative mechanisms, which can be termed as historical bias in the data. Such type of bias can lead to differences in base rates of outcomes for the two groups identified as privileged and protected. For developing non-discriminatory machine learning methods, a serious challenge is thus posed by the bias that seeps in through unfair dependencies in data-generative processes. In this chapter, we describe the methodology used for discovery of causal relationships within the attributes of the datasets, with the goal of finding pathways through which sensitive attributes influence the outcomes. For the sake of this work, the causal effects deemed to be unfair are characterized as:

1. Direct effect of the sensitive attribute on the outcome, which can be termed as disparate treatment

2. Indirect effects of sensitive attributes on the outcome through non-sensitive attributes, known as disparate impact.

In COMPAS dataset, the attribute COMPAS score is a proxy of the outcome, i.e. the risk of **recidivism**. Based on the logic of temporal causality, the phenomenon which is first to occur is identified as cause, while the phenomenon that follows is its effect. Therefore, if there exists a causal dependency between recidivism and scores, the edge would be logically directed towards the score ($Recidivism \rightarrow Score$). Since the focus in this study is on effects *on* the outcome, an edge discovered from Recidivism to Score may not be useful for understanding the unfairness patterns for recidivism prediction. However, in a supervised learning setting where we deal with fixed batches of data while ignoring temporal correlations between the data attributes, predictions are made by implicitly giving causal interpretations to statistical correlations. Therefore, for COMPAS dataset, we additionally learn the effects of attributes on the COMPAS Score, and interpret those as an indirect biasing effect on the outcome in the broader context of machine learning.
The rest of this chapter outlines the methodology used, which is already discussed at length in Chapter 2.

3.1 Assumptions

To reiterate, the meta-transfer causal induction framework requires assumptions of faithfulness to the ground truth causal model, and the principle of Independent Causal Mechanisms. Further, it measures the speed of adaptation to interventions as a training signal to discover cause-effect relationships.

It is interesting to note that these assumptions are based on observations about properties of the world, rather than the data distribution itself. This contrasts with several causal induction methods used in literature. However, the experiments done in the original research are limited to synthetic and semi-synthetic datasets from the BN repository. Owing to differences and/or biases in real-world data collection and aggregation processes, and given that the ground truth is often difficult to come by, recovery of the true, correctly parametrized causal structure for such data is a challenge. As discussed in the literature review of this work, many methods for causal induction and inference rely on structural assumptions on the ground truth causal model. Since our first goal is to get an understanding of the data distribution itself (in terms of pathways between sensitive variables and outcomes), and since the ground truth is unknown in our case, it makes little sense to use those methods. Instead, we test each dataset using multiple hypothetical causal factorizations under the assumptions mentioned above. The adapted approach is detailed in the next sections.

3.2 Obtaining Conditional Probabilities from Observations

To prepare the observational data for causal discovery, we first create Bayesian networks using the dataset. We do this for each hypothetical causal model of interest, so for each dataset, we end up with several candidate causal models.
For the original method, the required input is a Bayesian Interchange Format file(bif), a standardized format for representing graphical belief networks for discrete variables. The bif file contains the probability distribution of each variable conditioned on its parents. For each graph to be tested for causality, the dependency structure of the variables must be represented as a joint factorization that encodes the structural assumptions for that candidate. Therefore, the raw data is first used for fitting several Bayesian networks, each representing a candidate dependency graph. We used an open-source framework called CausalNex, a Python library that uses Bayesian Networks to combine machine learning with domain knowledge for causal reasoning. A succinct description of the method is as follows:

1. Encoding non-numeric variables
2. Structure learning
3. Obtaining candidate structures
4. Learning the CPD of each node factorized with respect to the candidate structure.

A more detailed account of the steps can be found in Appendix A. The learned CPDs are used to create a .bif file to provide as input to the causal discovery algorithm.

3.3 Causal Discovery Framework

Discovery of causal relationships is done through a series of experiments with conditional probabilities of observational datasets, factorized according to hypothetical Bayesian models. The proposed method is an adaptation from existing literature and experiments in causality in ML, which are based on measuring the speed of adaptation to interventions to learn the causal graph (Ke et al. (2019), Bengio et al. (2020), Priol et al. (2020)). Each experiment entails assumptions on the ground truth dependencies within the dataset, tested by a meta-transfer learning agent for the presence or absence of the cause-effect edges as defined by the input factorization. Under each hypothetical causal model, the agent yields different true positive and false negative edges. Further, for specific settings wherein true dependencies within the dataset deviate starkly from the hypothesis such that the faithfulness condition is not satisfied, learning from interventional distributions can open up spurious pathways between nodes, introducing false positives in the output.

The inputs used in the experiments of the original work are conditional probability distributions of Bayesian datasets from BNLearn repository. In contrast, the inputs used in our study are conditional probabilities obtained through real-world data collection and sampling processes, under assumptions on their true causal structure. If the structure in input is wrongly estimated, it would follow that the data observations are not well-represented by the assumed causal model, which could lead to opening of spurious paths between attributes. Within the causal meta-learner, this leads to violation of the causal Markov assumption in the input, which requires that the probability distribution over variables of the causal graph are factorized according to the causal graph (Equation 2.1).

The applied approach of trying different Bayesian factorizations for the given data is trial-and-error based. For a given dataset D with protected attributes attribute S, non-protected attributes $< X >$, and outcome Y, a variety of Bayesian datasets were produced with S, Y and a subset of $< X >$, each representing a specific configuration of cause-effect relationships in the given subset of data. Each meta-transfer learning experiment requires an input conditional probability table of a candidate Bayesian model. In the original research, the datasets used were generated from Bayesian processes, and the meta-transfer agent learns the adjacency matrix that represents the ground truth input. In each of each of our experiments, the input is a CPT obtained from a Bayesian model which may or may not represent the unknown ground truth. Thus, each experiment tests a hypothetical Bayesian model for the dataset, by testing the assumptions made on the probabilistic dependencies from the observational data. Naturally, some of the assumed dependencies in the inputs would not hold true for data, leading to false negative edges in the output adjacency matrix. Further, for an input CPT which deviates more from the ground truth of the data, the gradient accumulated along transfer learning episodes on interventional distributions can lead to opening of backdoor paths for certain edges. This could result in bias and yields erroneous results for those edges.

Analysis of results from all experiments for a given dataset helps to infer the presence and absence of direct causal effects of protected attributes on the outcome, which ultimately aids the evaluation of behaviours of fairness pre-processing methods in the presence of such a direct effect. In the next chapter, experimental setup and results are presented for causal discovery, as well as for the fairness mitigation methods used, which were described in Chapter 2. In Chapter 5, we provide a discussion of the results.

Chapter 4

Experiments and Results

In this chapter, we describe the experiment setup and results for the two sub-goals of the research question, causal discovery and fairness mitigation.

4.1 Causal Discovery

The current section is structured as follows: the Experiments and Results subsection provides details for the experimental setup for each dataset, and presents the results. This is accompanies with interpretation of each result. We summarize our findings in the conclusions subsection.

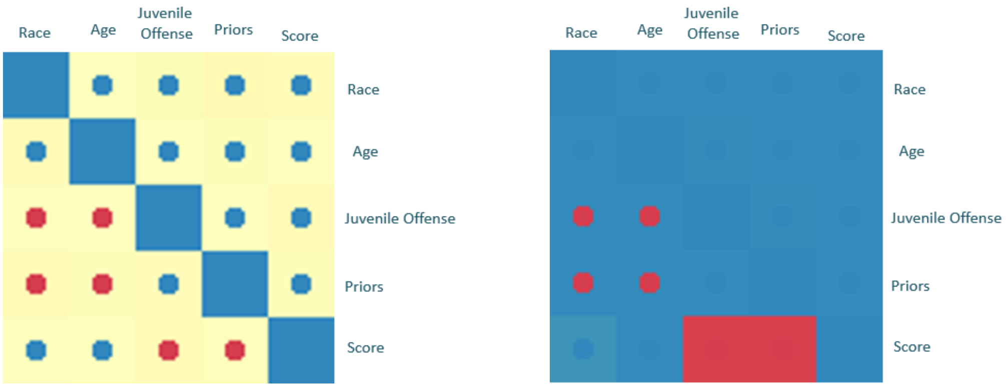
4.1.1 Experiments and Results

Here, we provide implementation details and results of the causal discovery experiments for each dataset.

COMPAS

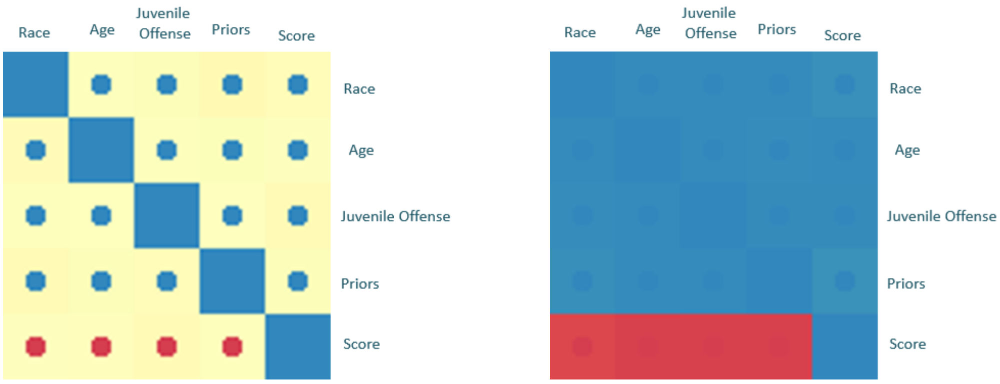
Hypothetical ground truth CPDs were created for testing for presence of a causal path between Race and COMPAS Score as well as Race and Recidivism. The variables Score and Recidivism were not used together. This is because our goal is to estimate effects on Recidivism, and causal effect of Score on Recidivism is makes little sense in the real world, since the scores were computed through the COMPAS algorithm to predict risk of recidivism. However, in machine learning settings, prediction of Recidivism has been done with Scores included the input, which is a proxy of Recidivism. If a logistic regression model is trained to predict recidivism using a dataset with Score, effect of Race on Score would inadvertedly seep into the conditional probability of Recidivism given the Score. Therefore, we perform experiments to test the effects of attributes on the Score, and perceive such effects as indirect effects on Recidivism within the broader context of machine learning predictions.
Figures 4.1 and 4.2 represent the results from experiments with Race, Age, Juvenile and Prior offences, and Scores. It was found that Race, Age and Priors all have causal effects on Scores.

Figures 4.3 and 4.4 show results for testing the presence of direct and indirect effects of Gender on Score respectively, both of which give false negatives. Thus, no unfair causal effect of gender on the outcome is observed.



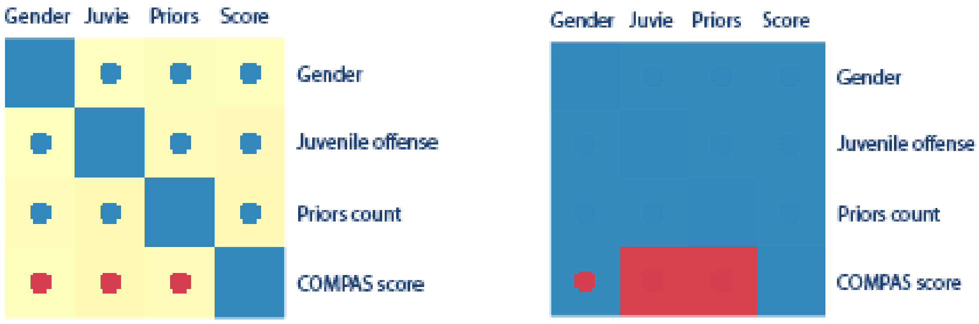
(a) Hypothetical ground truth (b) Output

Figure 4.1: Causal structure discovery, assuming that Race and Age affect Scores **indirectly**.



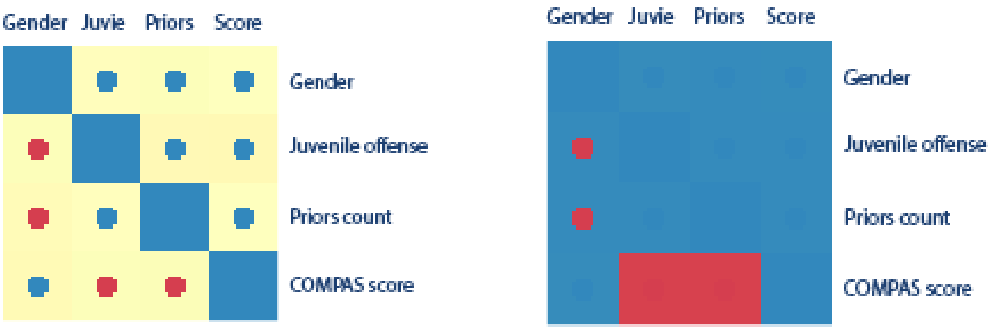
(a) Hypothetical ground truth (b) Output

Figure 4.2: Causal structure discovery, assuming that Race and Age affect Scores **directly**.



(a) Hypothetical ground truth (b) Output

Figure 4.3: Causal structure discovery, assuming that Gender affect Scores **directly**.

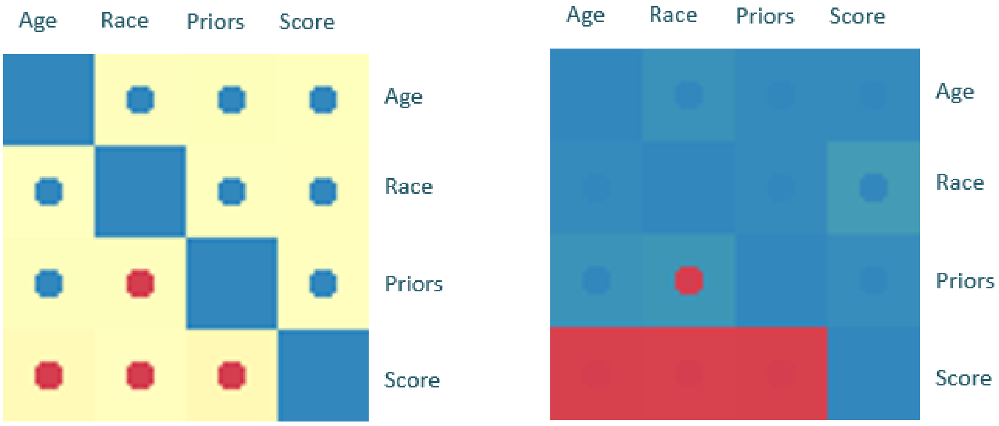


(a) Hypothetical ground truth

(b) Output

Figure 4.4: Causal structure discovery, assuming that Gender affect Scores **indirectly**.

Figure 4.5 shows the results for direct and indirect edges between race and score. Again, a true positive was found for *Race* → *Score*, but none for *Race* → *Priors*.



(a) Hypothetical ground truth

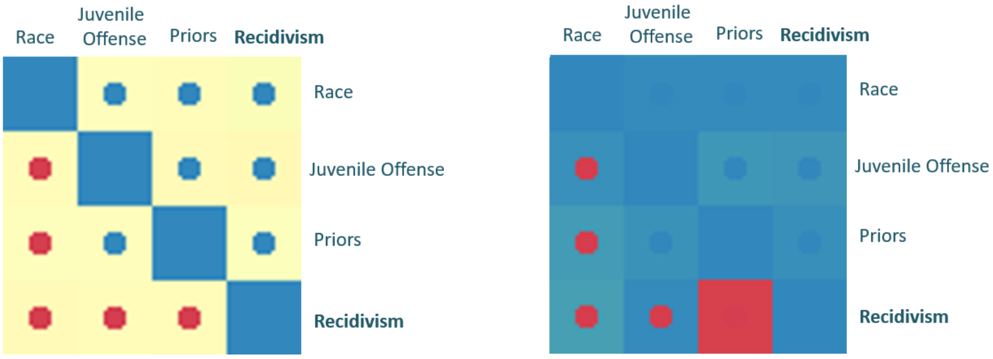
(b) Output

Figure 4.5: Causal structure discovery, assuming that Race and Age affect Scores **directly** and **indirectly**.

In the experiments with Recidivism shown in Figure 4.6, Priors was learnt to have a causal effect on Recidivism, but not Race. Therefore, no direct effect is found between Race and Recidivism. However, we perceive the recovered edge between Race and COMPAS score as a source of indirect bias for Recidivism.

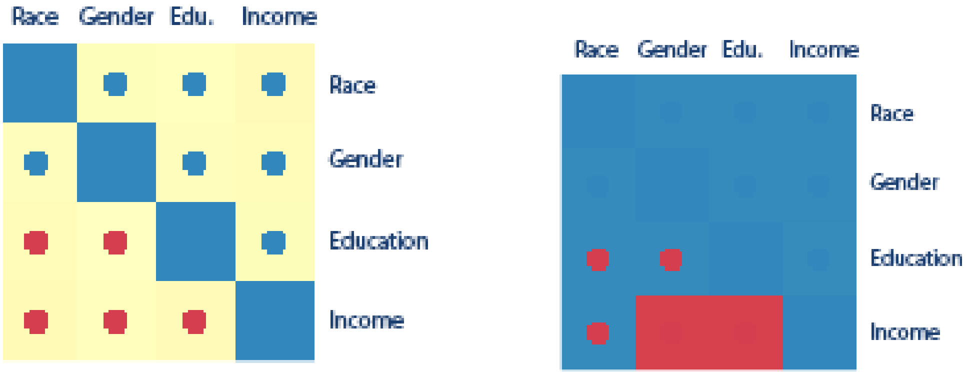
4.1.2 Income

We tested for the presence of causal effects between gender, education and income. Within the chosen set attributes, direct causal effects on income from gender and education were learned, shown in Figures 4.7 and 4.8. As seen in Figure 4.7, no direct or indirect effects from Race were observed on the Income.



(a) Hypothetical ground truth (b) Output

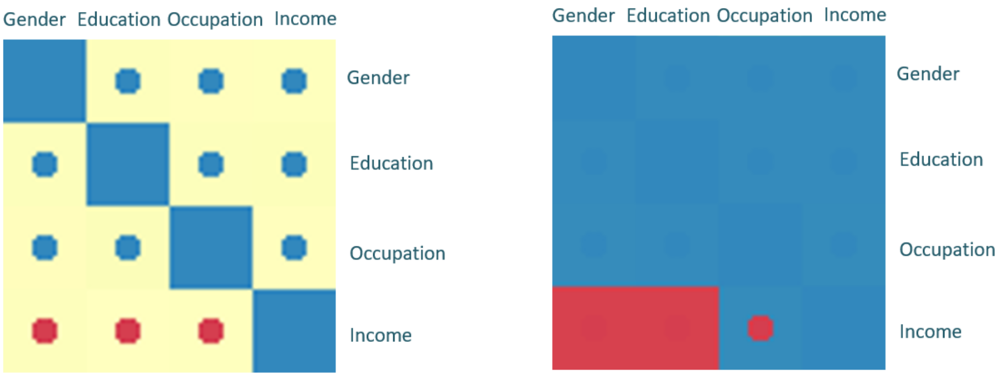
Figure 4.6: Causal structure discovery, assuming that Race and affects Recidivism **directly**.



(a) Hypothetical ground truth (b) Output

Figure 4.7: Causal structure discovery, assuming that Gender and Race affect Income **indirectly** and *directly*.

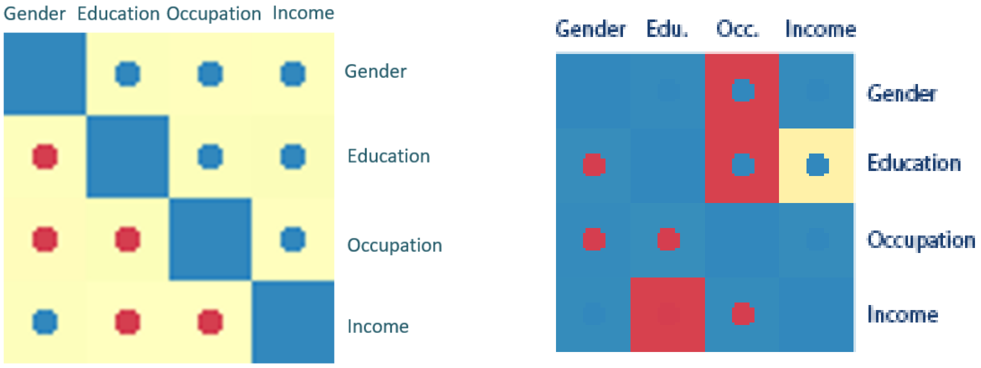
The last experiment shown in Figure 4.9 tests indirect effects from Gender to Income, and yields false positive edges as shown in Figure 3.9. For two of the input edges $Gender \rightarrow Occupation$ and $Education \rightarrow Occupation$, the output shows false negatives. This could be due to a difference between trends in marginal associations between the variables and their conditional probabilities, leading to a reversal of the edge, a phenomenon known as Simpson's paradox. Lastly, we find a slight causal edge from Income to Education, which is erroneous. This could be caused due to the transfer episodes wherein the intervention is performed on the collider node Occupation, which opens a spurious pathway: $Gender \rightarrow Education \rightarrow Income \leftarrow Occupation$. The opening of this path creates bias in the interventional distribution, leading to the partial false positive edge $Income \rightarrow Education$.



(a) Hypothetical ground truth

(b) Output

Figure 4.8: Causal structure discovery, assuming that Gender affects Income **directly**.



(a) Hypothetical ground truth

(b) Output

Figure 4.9: Causal structure discovery, assuming that Gender affects Income **indirectly**.

4.1.3 Conclusion

We discover unfairness patterns in the data by learning its causal structure, under the assumption that the data-generative processes are composed of independent mechanisms. The causal disentanglement method described above requires inputs in the form of conditional probability distributions(CPDs) of Bayesian network datasets. In the experiments done by Ke et al. (2019), the datasets used are semi-synthetic, where the ground truth causal models are known. The results of their experiments prove that their method successfully recovers the ground truth causal models, under the stated assumptions. However, for real-world datasets, ground truth causal models are difficult to come by. Therefore, we used their methods as a hypothesis testing framework. We produce several joint factorizations for the same dataset, under different assumptions on the ground truth causal model. The dataset was fitted to the candidate causal models (represented as Bayesian networks) to obtain CPDs for each. Subsequently, multiple experiments were performed per dataset, using the candidate CPDs obtained in the previous step. For each experiment, the candidate CPD was provided to the meta-transfer learner, which uses nonstationarities in transfer distributions as a training signal to learn the independent cause-effect mechanisms underlying the original dataset. The output is an adjacency matrix representing the causal relationships between variables. The results from all experiments for the dataset were analyzed, to infer the pathways between sensitive attributes and outcome. It was found that Race has a causal effect on COMPAS score. In the Income dataset, Gender is seen to have an effect on Income. In contrast, Gender does not have any causal effects on the COMPAS score. Neither does Race have a causal effect on Income. Thus, the datasets have different historical biases. In this way, we gain an understanding of algorithmic bias through a causal lens.

4.2 Fairness Mitigation

4.2.1 Experiments

We used the AI Fairness 360 Toolkit to for mitigating and evaluating unfairness. For simplicity of our setup, the protected attributes and outcomes are all binary variables. For the Income dataset, the attributes used were education, age and income, besides the protected attributes. For the COMPAS dataset, we used race, gender, prior counts, score, age, and recidivism. As distortion control in optimized preprocessing for the Income dataset, reductions in income, small changes in age and small increases in education are allowed with small probabilities. Large changes are not allowed for age and education. In the COMPAS dataset, increase in recidivism and large increases in age and prior counts are heavily penalized. Decrease in recidivism and smaller changes in age and prior counts are given smaller penalties.

4.2.2 Results

Table 4.1 presents the results obtained for combination of dataset, protected attribute, and preprocessing method. In each case, a logistic regression classifier is trained the following performance metrics are computed:

- **Balanced Accuracy:** average of the accuracies in predicting positive and negative outcomes
- **Statistical Parity Difference:** difference of the rate of favorable outcomes received by the unprivileged group to the privileged group. The ideal value is 0, and the acceptable fairness range is between -0.1 and 0.1
- **Disparate Impact:** ratio of rate of favorable outcome for the unprivileged group to that of the privileged group. A value < 1 implies higher benefit for the privileged group and a value >1 implies higher benefit for the unprivileged group.
- **Average Odds Difference:** average difference of false positive rate (false positives / negatives) and true positive rate (true positives / positives) between unprivileged and privileged groups. The ideal value of this metric is 0. A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group.
- **Equal Opportunity Difference:** This metric is computed as the difference of true positive rates between the unprivileged and the privileged groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group. The ideal value is 0. A value of < 0 implies higher benefit for the privileged group and a value > 0 implies higher benefit for the unprivileged group. Fairness for this metric is between -0.1 and 0.1.

Dataset	Protected Attribute	Processing Method	Balanced Accuracy	Statistical Parity Diff.	Disparate Impact	Average Odds Diff.	Equal Opportunity Diff.
COMPAS	Race	None	0.6774	-0.2494	0.66	-0.1927	-0.1877
COMPAS	Race	Optimized	0.6752	-0.097	0.8534	-0.0678	-0.0691
COMPAS	Race	Reweighting	0.6342	0.0546	1.1062	0.1042	0.1215
COMPAS	Gender	None	0.6774	-0.2724	0.6631	-0.2439	-0.1392
COMPAS	Gender	Optimized	0.6714	-0.095	0.8619	-0.0725	-0.0523
COMPAS	Gender	Reweighting	0.6562	-0.1188	0.8342	-0.0946	0.0111
Adult	Gender	None	0.7437	-0.358	0.7205	-0.3181	-0.3769
Adult	Gender	Optimized	0.7013	-0.0722	0.7895	-0.0487	-0.0429
Adult	Gender	Reweighting	0.7134	-0.0705	0.7785	0.0188	0.0293
Adult	Race	None	0.7437	-0.2435	0.5878	-0.1966	-0.202
Adult	Race	Optimized	0.7443	-0.0853	0.7804	-0.0619	-0.062
Adult	Race	Reweighting	0.7311	-0.0523	0.8508	0.0419	0.1083

Table 4.1: Metrics of predictions from a logistic regression predictor, with and without applying preprocessing methods to the datasets.

Discussion of the results is provided in the next chapter, with respect to our findings from causal discovery.

Chapter 5

Discussion

In the previous chapters, we described the methodologies and experimental results for each sub-problem of the research question: (a) identification of potential sources of bias for the algorithm, and (b)statistical fairness mitigation. In this Chapter, we assimilate our findings from the experiments in bias discovery and fairness mitigation. Based on the representation bias measured through dataset imbalance in Chapter 2, along with causal-effect dependencies learned in Chapters 3 and 4, a distinction is made for each combination of dataset and protected attribute. Then, we interpret the the two fairness preprocessing methods for each dataset and protected attribute, with respect to the sources of bias from the data or algorithm in each case.

The rest of this chapter is structured as follows: Section 5.1 outlines the types of biases found for the two datasets. In Section 5.2, statistical fairness results from Chapter 5 are highlighted for each dataset and protected attribute. This leads to interpretation of the two fairness preprocessing methods, vis-à-vis the data distribution and the nature of its attributes, which answers the second sub-problem of our research: relative performance of fairness methods in the presence of different sources of bias in the system.

5.1 Bias in Datasets

We highlight and discuss the following characteristics found in the datasets:

- group imbalance and proportions of true outcomes within each group,
- presence of causal pathways of dependence between protected attributes and outcomes.

Dataset	**Protected Attribute**	**Group Imbalance**	**Positive Outcomes Imbalance**
COMPAS	Race	1.48	1.42
COMPAS	Gender	4.62	1.29
Adult Income	Race	6.19	1.71
Adult Income	Gender	2.07	2.67

Table 5.1: Representation bias in the two datasets, measured for each protected attribute.

5.1.1 Adult Income

Imbalance in dataset : as shown in Table 5.1, both cases of the Income dataset indicate the imbalance between protected and privileged group membership, but the gap in representation is much wider for race than for gender. Further, the rate of positive outcomes, i.e. the odds of earning more than 50k USD; was found to be higher for males than for females, and higher for Caucasians than for African-Americans, with the gap being wider for gender groups than for racial groups.
Causal Dependence : the results from Chapter 4 reveal a cause-effect relationship between gender and income, but not between race and income. This is depicted in 5.1. The only other edge discovered is between education and income. Further, no indirect effect of gender on income was discovered through other attributes, such as education or occupation.

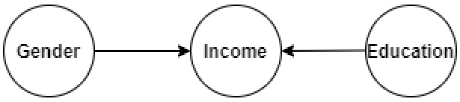
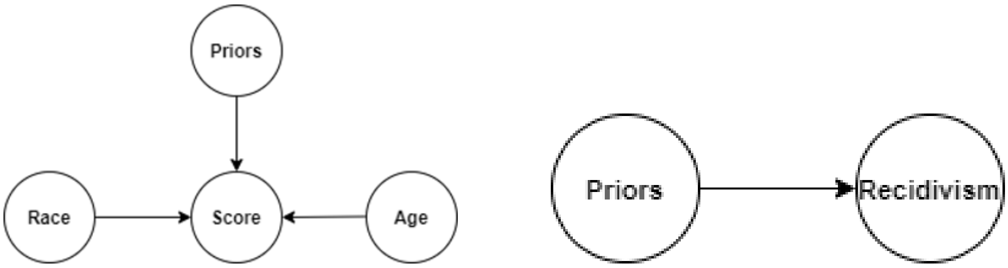


Figure 5.1: Graphical representation of cause-effect relationships learned for attributes of the Adult Income dataset. Race is excluded from the figure, since it was not found to have any cause-effect relationship with the remaining attributes.

5.1.2 COMPAS

Imbalance in dataset : here, the privileged gender group is Female, as opposed to the setting in the income data. As seen in table 5.1, he gap in representation is wider for race than for gender. Further, the rate of positive outcomes - in this case- the odds of being a non-reoffender; was found to be higher for females than for males, and higher for Caucasians than for African-Americans, with the difference being almost the same for gender groups as for racial groups.
Causal Dependence : the results from Chapter 4 reveal a cause-effect relationship from race to COMPAS scores, but not to the actual recidivism outcomes. Attributes such as priors count and juvenile priors count were also found to have causal effects on the scores. A causal effect of race on scores indicates racial discrimination, which can be qualified as disparate treatment. This is in line with the phenomenon of redlining, wherein a disproportionate amount of police attention is given to members of a specific race over others, leading to a difference in base rates for the protected and privileged groups. Details of the workings of COMPAS algorithm is not publicly known, and it is difficult to exactly assess its behaviour. However, if the scores are to be included in the dataset for training a classifier, they would encode fair information from prior arrests, but would also indirectly add racial bias in the algorithm. Experiments done with true recidivism reveal a causal edge with priors count, but not with race, or even juvenile count. This is depicted in 5.2.



(a) Cause-effect relationships learned for COMPAS scores

(b) Cause-effect for true recidivism

Figure 5.2: Graphical representation of cause-effect relationships learned for COMPAS dataset.

5.2 Statistical Fairness Metric Predictions

In this section, metrics of classification and statistical fairness are presented for unprocessed, reweighed, and optimized data. Namely, we compare balanced accuracy, statistical parity difference, disparate impact, along with trends in false positive and true negative rates. Together with the quantitative and qualitative results presented in the previous section, this section facilitates discussion on efficacy of different fairness methods for different sources of bias.
Figures 5.3 and 5.4 depict the performance of the processing methods in the Income dataset, for gender and race bias respectively. Figures 5.5 and 5.6 depict the performance of the processing methods in the COMPAS dataset, for gender and race bias respectively.

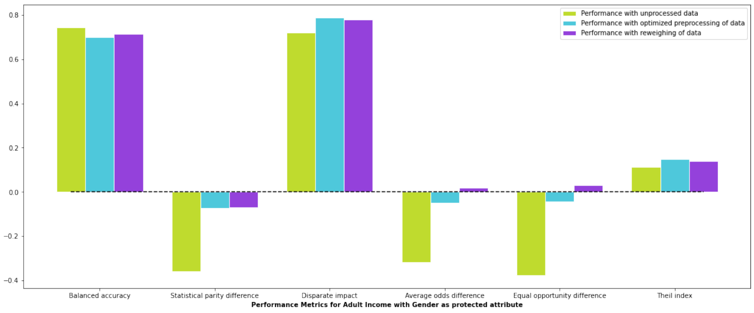


Figure 5.3: Performance metrics obtained from LR predictions of the unprocessed and processed Adult Income dataset. Fair preprocessing was applied for controlling gender bias.

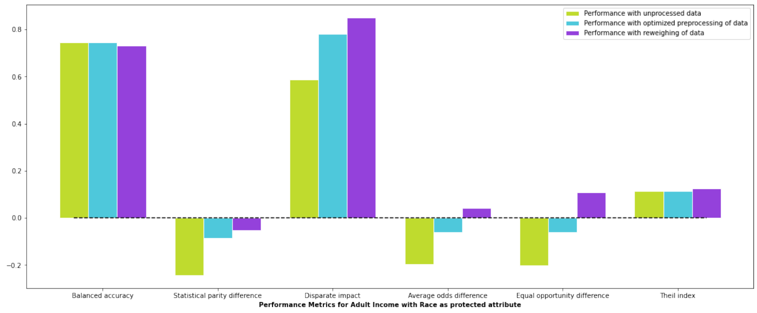


Figure 5.4: Performance metrics obtained from LR predictions of the unprocessed and processed Adult Income dataset. Fair preprocessing was applied for controlling racial bias.

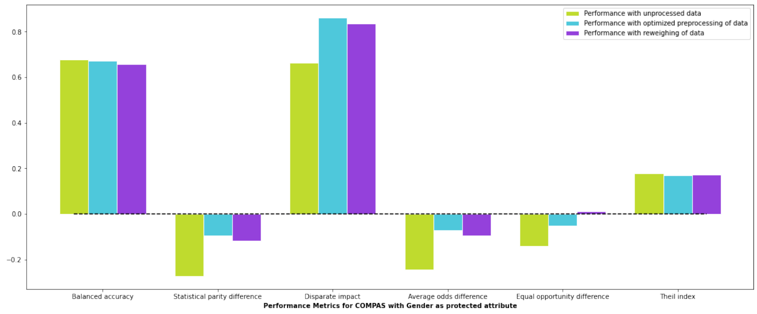


Figure 5.5: Performance metrics obtained from LR predictions of the unprocessed and processed COMPAS dataset. Fair preprocessing was applied for controlling gender bias.

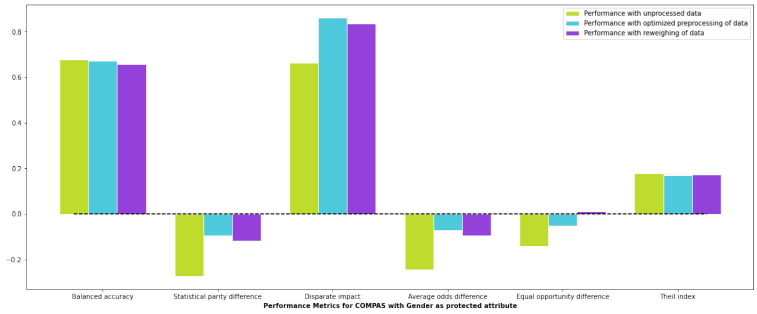


Figure 5.6: Performance metrics obtained from LR predictions of the unprocessed and processed COMPAS dataset. Fair preprocessing was applied for controlling racial bias.

5.3 Assimilation of Results

In this section, we interpret the results presented in section 5.2, by incorporating the observations on qualitative and quantitative characteristics of the data distribution, highlighted under 5.1. For each dataset in the subsequent sections, the followinf structure is followed:

1. Evaluation of metric predictions from unprocessed data

2. Evaluation of metric predictions from processed data

3. Observations and conclusions, based on all the results

5.3.1 Adult Income

Metric predictions from unprocessed data

For this subsection, we refer to an overview of types of bias in the dataset and fairness metrics of raw predictions, provided in Table 5.2.

Dataset	Protected Attribute	Causal dependence	Representation Imbalance	Positives Imbalance	Disparate Impact	EOD	AOD
COMPAS	Race	Indirect effect	1.48	1.42	0.66	-0.1927	-0.1877
COMPAS	Gender	None	4.62	1.29	0.6631	-0.2439	-0.1392
Adult	Gender	Direct effect	2.07	2.67	0.7205	-0.3181	-0.3769
Adult	Race	None	6.19	1.71	0.5878	-0.1966	-0.202

Table 5.2: Qualitative and quantitative characteristics of the data distributions, along with statistical fairness metrics of predictions on raw(unprocessed) data.

Disparate impact is a measure of statistical parity, which relates to the independence of outcome and sensitive attribute. Under the given implementation, disparate impact would be close to 1 for a predictor which is independent of gender, with values < 1 signifying bias against protected group; and values > 1 signifying bias against privileged group. In the case of gender bias in the Income dataset, predictions from unprocessed data have a disparate impact of 0.72; while for race, the value is 0.58. The disparity for racial groups can be attributed to a significantly high imbalance in their representation in data. In contrast, the imbalance in representation of the genders is not too high, but the imbalance in positive outcomes of the two gender groups is notably higher in this case.

Although the income attribute has a causal dependency on gender but not on race, the disparate impact would suggest a lower dependence of income on gender than on race. However, the **average odds difference** and **equal opportunity difference** are both higher in the case of gender. If the true outcome truly has a dependency on the protected attribute, a high disparate impact could lead to different false positive and true positive rates (Prince (2019)). This would lead to a higher difference of average odds, as well as a higher difference of equal opportunity. This is apparent from our results for controlling the effect of gender on the outcome.

Metric predictions from preprocessed data

The same performance metrics are compared for predictions on the pre-processing methods, for gender and racial bias control, wherein the protected groups are female and African-American, respectively.

Gender: while controlling for gender bias in the Income dataset, both methods lead to a significant loss in accuracy, relative to the remaining three cases. Optimized pre-processing fared slightly

better on fairness metrics, but at a larger cost on accuracy compared to reweighing.
Race: although no causal edge was discovered to indicate a probabilistic dependence of income on race, the imbalance in group representation is notably high. The loss in accuracy after reweighing is small but present, while optimized processing yields a higher accuracy than raw predictions. Both methods boost the disparate impact significantly, with re-weighing outperforming optimized pre-processing.

5.3.2 COMPAS

Metric predictions from unprocessed data

The accuracy on the COMPAS dataset is generally lower, and **disparate impact** values for gender and race are both close to 0.66. However, the average odds difference is higher for gender bias, and a wider difference is observed in the true negative and false positive rates between the gender groups. Further, the dataset is much more imbalanced for gender groups than for racial groups.

Metric predictions from preprocessed data

The performance metrics are compared for predictions on the pre-processing methods, for gender and racial bias control, wherein the protected groups are male and African-American, respectively.
Gender: both methods perform similarly in accuracy and disparate impact, with optimized pre-processing faring slightly better in both.
Race: optimized preprocessing yields a disparate impact of 0.85, while the value for reweighing is 1.1. This, along with the positive sign of average odds difference and equal opportunity difference, indicates that the direction of bias is reversed after reweighing. Optimized preprocessing manages to maintain almost the same accuracy as the raw predictions, while reweighing leads to a relatively larger loss in accuracy.

5.4 Conclusion

Given the results from experiments on causality, and the trend revealed by the statistical fairness metrics, we observe that the biases added through the two protected attributes, gender and race, manifest differently in the algorithm.

5.4.1 Adult Income

For controlling gender discrimination, optimized preprocessing fares better overall on the statistical fairness metrics than re-weighing. However, the trend in their accuracies is the opposite. Thus, a tradeoff is observed between accuracy and fairness for the two methods in this case. For racial discrimination, both methods lead to similar accuracies, but re-weighing outperforms optimized-preprocessing in fairness.

5.4.2 COMPAS

It was found that the balanced accuracy of the raw COMPAS predictions is lower than that achieved in the Income dataset. For controlling race discrimination, optimized preprocessing leads to improvements in statistical fairness metrics, while re-weighing learns a representation which favours the protected group, at a higher cost to accuracy. This cost can be perceived as a tradeoff for algorithmic equity. For controlling gender discrimination, both methods perform fairly similar, but optimized pre-processing leads to slightly better accuracy as well as disparate impact

Chapter 6

Conclusions

The pathways of causal dependence between a protected attribute and outcome can signify unfairness patterns within underlying data-generative mechanisms. Through a causal Bayesian view on fairness(Chiappa and Isaac (2019)), we know that bias added through such pathways manifests discrimination in the algorithm which cannot be sufficiently explained through statistical fairness metrics. In order to meaningfully interpret the efficacy of algorithmic fairness methods, a view of the data-generative mechanisms within the datasets is a critical asset.
The following conclusions were made for the Income dataset from the previous chapter:

- In terms of fairness metrics, optimized preprocessing performed slightly better for controlling gender bias, which is a direct causal effect, while reweighing outperforms for racial bias introduced due to imbalanced group representation.
- For gender and racial bias in income dataset, a tradeoff between accuracy and fairness of the two preprocessing methods is maintained.

Further, given the causal dependency of the outcome on gender learned within the true distribution, the loss in accuracy can be interpreted as the cost of reducing the probabilistic dependence of the predicted outcome on gender. Therefore, we argue that this loss is justifiable for the goal of group-level fairness.
For COMPAS dataset, the following conclusions were made:

- With race as protected attribute, reweighing appears to bias the results towards the privileged group
- Reweighing appears to perform more reliably for gender bias, which is caused by group imbalance, than for racial bias, which is an indirect causal effect through score
- Tradeoff between accuracy and fairness of the two preprocessing methods is maintained for racial bias, but not for gender bias.

For controlling racial bias in COMPAS, re-weighing leads to a significant loss in accuracy, along with a sign reversal in its statistical fairness metrics. Since there was no direct causal relationship discovered between race and recidivism, we assume that the distribution of recidivism in the dataset represents the ground truth. Therefore, the loss in accuracy in this case indicates to us that using reweighing in such a context can lead to unreliable predictions. By definition, reweighing makes the group membership of individuals compete against the weight assignment.

In this case, there is a competing third factor. In essence, reweighing assigns a higher weight to individuals of the protected group with negative outcomes than to those in the protected group with positive outcomes, and vice-versa for the privileged group. However, race has an causal effect on COMPAS scores. Therefore, COMPAS scores in the training dataset pass on biased information to recidivism predictions. We conclude that the indirect effect of race on the outcome through COMPAS scores presents a more complex pattern of unfairness, which would require specialised methods for mitigation.

Thus, we find that datasets with inherent bias require different mitigation methods than those without. Such datasets also present an additional challenge: due to bias in the true distribution of the outcomes, mitigation methods can result in more significant losses to accuracy. On one hand, it can be argued that this loss is justifiable, since it is exchanged for more equitable decision-making. However, whether this is a correct conclusion to make in a given situation, is difficult to establish without any semblance of a ground truth on fairness together with accuracy. Another consideration is the reversal of direction of disparity, i.e. apparent reverse bias in the fairness results, which is observed in the case of reweighing for racial bias in COMPAS. Again, it may be argued that this tradeoff is justifiable under the goal of algorithmic equity, and again, unless the ground truth $fair$ outcomes are known, it is difficult to establish whether this is indeed justifiable. However, this consideration relates to the goal of algorithmic justice, for which fairness can be considered a proxy. While this is out of scope for the research question we address, this is a noteworthy limitation of our methods. The limitation is explained in the following section, through the example of our results.

6.1 Future Work

A relevant direction for resolving the issues mentioned above is *counterfactual reasoning.* "Would the outcome be different if the individual belonged to a different demographic group?" Such questions occur very naturally to humans. Equipping algorithms to do counterfactual reasoning could help to provide data-driven answers to such questions. Therefore, a future direction for this work would be to develop counterfactual explanations to the outcomes, which provides a useful and relevant metric of fairness through a causal lens. Counterfactual fairness in algorithms is a vastly growing field with several important contributions. Just like any other machine learning algorithm, the goodness of such a method would depend on quality of the observed dataset. Yet, counterfactual reasoning would bring us objectively closer to a unified definition of fairness, and establish a fairer ground truth to perform further studies.

Bibliography

State v. loomis: Wisconsin supreme court requires warning before use of algorithmic risk assessments in sentencing. *Harvard Law Review*, 881 N.W.2d 749 (Wis. 2016), March 2017. 1

Julia Angwin and Jeff Larson. Bias in criminal risk scores is mathematically inevitable. *ProPublica*, December 2016. URL `www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say`. 1

Yoshua Bengio. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, 2009. 10

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher J. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *ICLR*. OpenReview.net, 2020. 11, 17

T. Brennan, W. Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36:21 – 40, 2009. 1

Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery with cascade nonlinear additive noise models. *CoRR*, abs/1905.09442, 2019. 10

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *ICDM Workshops*, pages 13–18. IEEE Computer Society, 2009. 2, 12

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized pre-processing for discrimination prevention. In *NIPS*, pages 3992–4001, 2017. 2, 12

Jiahao Chen. Fairness under unawareness: Assessing disparity when protected class is unobserved. *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, January 2019. 8

Silvia Chiappa. Path-specific counterfactual fairness. In *AAAI*, pages 7801–7808. AAAI Press, 2019. 9

Silvia Chiappa and William S. Isaac. A causal bayesian networks viewpoint on fairness. *CoRR*, abs/1907.06430, 2019. 8, 35

R. Courtland. Bias detectives: the researchers striving to make algorithms fair. *Nature*, 558: 357–360, June 2018. 8

Anupam Datta. How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *arXiv*, July 2017. 8

A. Dawid. Fundamentals of statistical causality. 2007. 8, 9

A. Hoffmann. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication Society*, 22:900 – 915, 2019. 7

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *CoRR*, abs/1910.01075, 2019. 2, 11, 17, 23

Nan Rosemary Ke, Jane X. Wang, Jovana Mitrovic, Martin Szummer, and Danilo J. Rezende. Amortized learning of neural causal representations. *CoRR*, abs/2008.09301, 2020. 11

Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS*, pages 4066–4076, 2017. 9

Matt J. Kusner, Chris Russell, Joshua R. Loftus, and Ricardo Silva. Causal interventions for fairness. *CoRR*, abs/1806.02380, 2018. 9

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica*, May 2016. URL `www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`. 1

J. Lemeire and E. Dirkx. Causal models as minimal descriptions of multivariate systems. 2006. 10

Joshua R. Loftus, Chris Russell, Matt J. Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *CoRR*, abs/1805.05859, 2018. 5

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. 8

L. Oneto and S. Chiappa. Fairness in machine learning. 2020. 8

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 4033–4041. PMLR, 2018. 10

S. Prince. Bias and fairness in AI. August 2019. URL `www.borealisai.com/en/blog/tutorial1-bias-and-fairness-ai/`. 31

Rémi Le Priol, Reza Babanezhad Harikandeh, Yoshua Bengio, and Simon Lacoste-Julien. An analysis of the adaptation speed of causal models. *CoRR*, abs/2005.09136, 2020. 11, 17

Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. When worlds collide: Integrating different counterfactual assumptions in fairness. In *NIPS*, 2017. 10

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *SIGMOD Conference*, pages 793–810. ACM, 2019. 9, 10

Bernhard Schölkopf. Causality for machine learning. *CoRR*, abs/1911.10500, 2019. 5, 10, 11

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *ICML*. icml.cc / Omnipress, 2012. 13

Ram Shanmugam. Causality: Models, reasoning, and inference : Judea Pearl; cambridge university press, cambridge, uk, 2000, pp 384, ISBN 0-521-77362-8. *Neurocomputing*, 41(1-4): 189–190, 2001. 9

Harini Suresh and John V. Guttag. A framework for understanding unintended consequences of machine learning. *CoRR*, abs/1901.10002, 2019. 7

J. X. Wang, Zeb Kurth-Nelson, Hubert Soyer, Joel Z. Leibo, Dhruva Tirumala, Rémi Munos, Charles Blundell, D. Kumaran, and Matt M. Botvinick. Learning to reinforcement learn. *ArXiv*, abs/1611.05763, 2017. 11

B. Williams, C. Brooks, and Yotam Shmargad. How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy*, 8:78–115, 2018. 8

Appendix A

Learning Bayesian Networks from Data

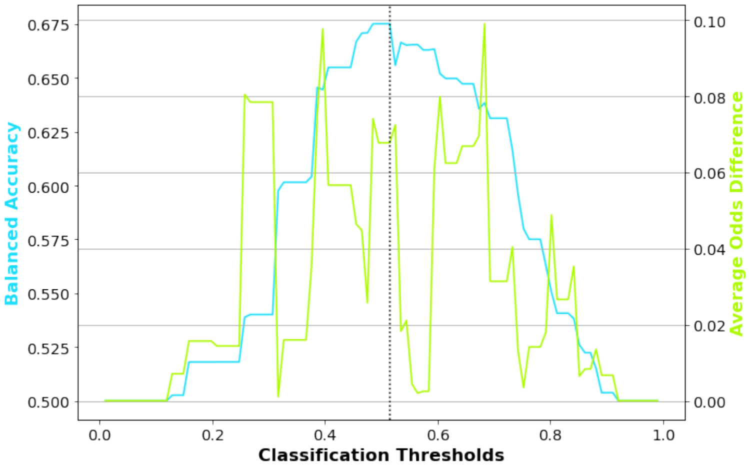
The following steps were used to obtain CPDs required as input for the meta-transfer learner.

1. Created a dataframe with only the relevant columns and dropped rows with any empty values
2. Encoded non-numeric columns using LabelEncoder from sklearn's preprocessing module
3. Learned the structure using NOTEARS algorithm
4. Removed all weak edges(< 0.8) by thresholding
5. Modified the structure further to obtain a candidate graph
6. Discretized and labelled the numerical features to make them categorical
7. Fit the probability distribution of the BN model using the discretized data
8. Learn CPD of each node in the BN using fit_cpds method from the BayesianNetwork class

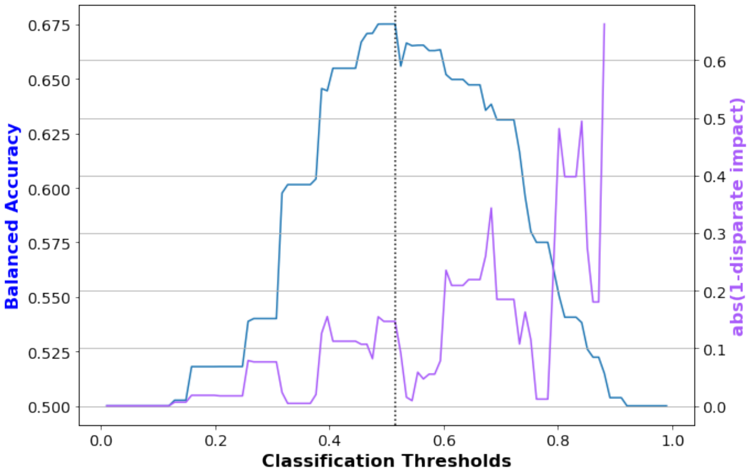
Appendix B

Tradeoff Between Balanced Accuracy and Fairness

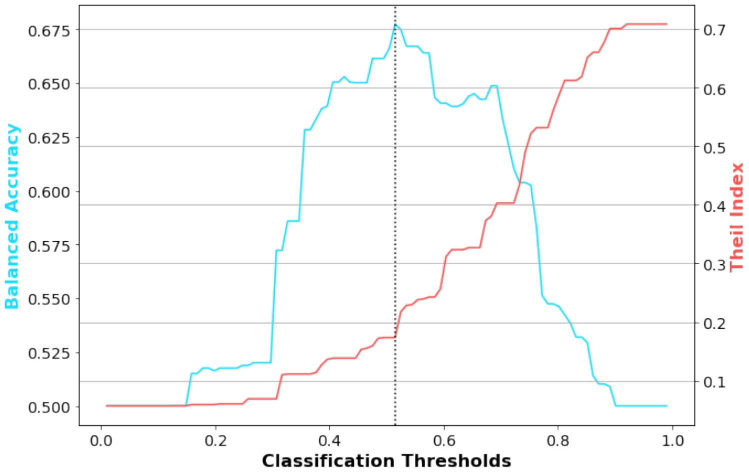
B.1 Optimized Prepocessing for Race in COMPAS



(a) Average Odds Difference

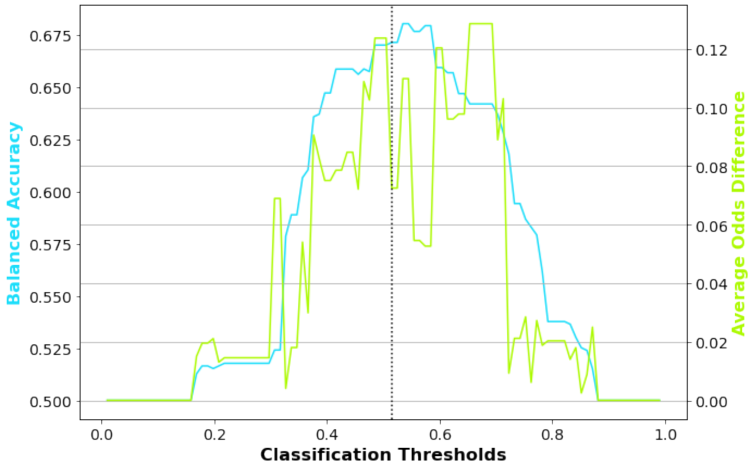


(b) Disparate Impact

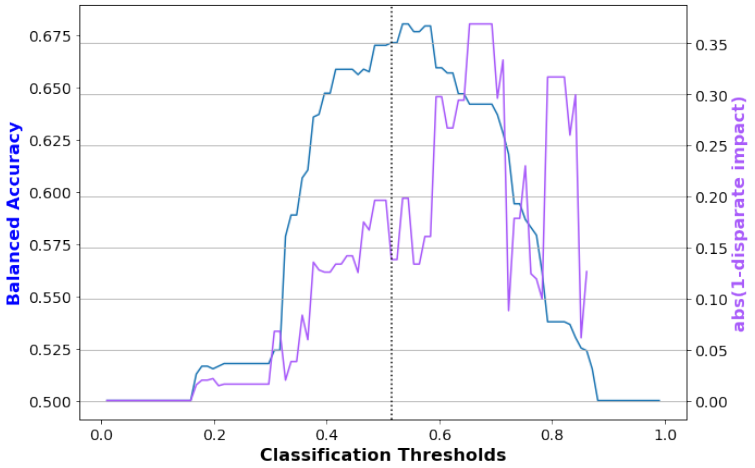


(c) Theil Index

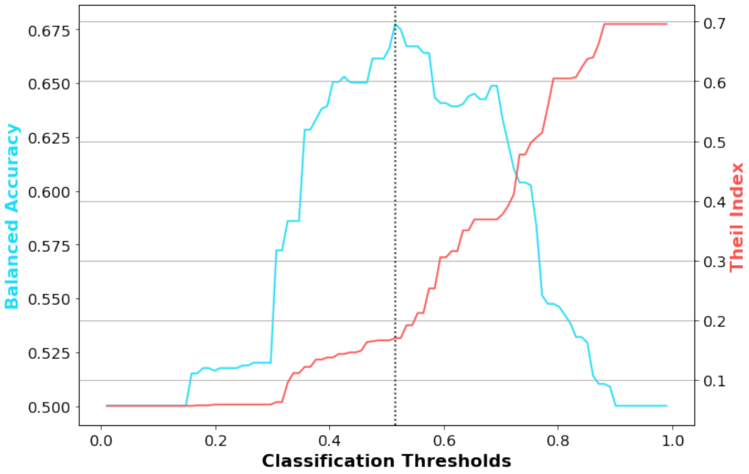
B.2 Optimized Prepocessing for Gender in COMPAS



(d) Average Odds Difference

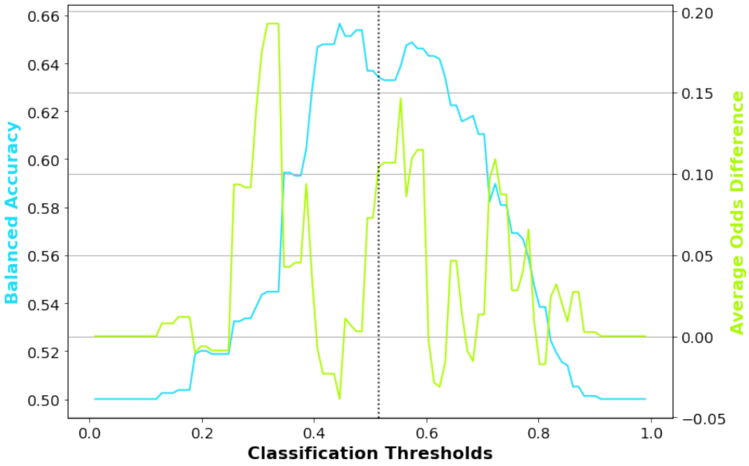


(e) Disparate Impact

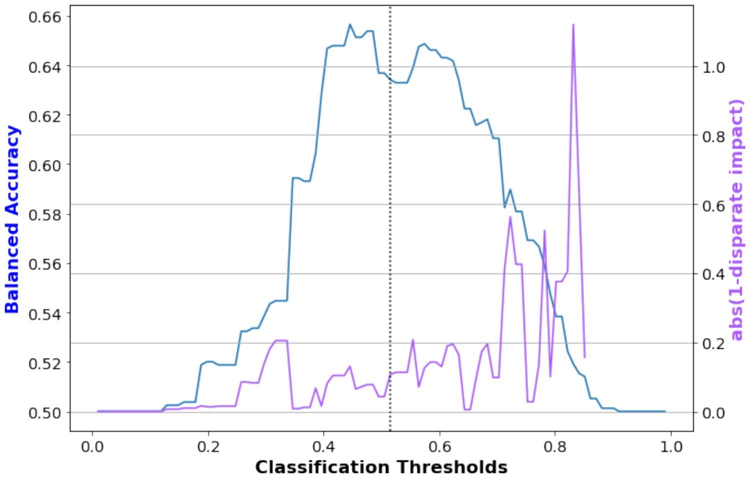


(f) Theil Index

B.3 Re-weighing for Race in COMPAS

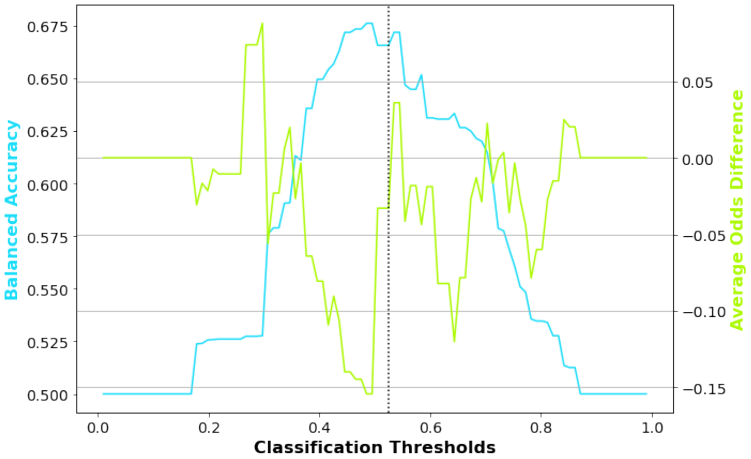


(g) Average Odds Difference

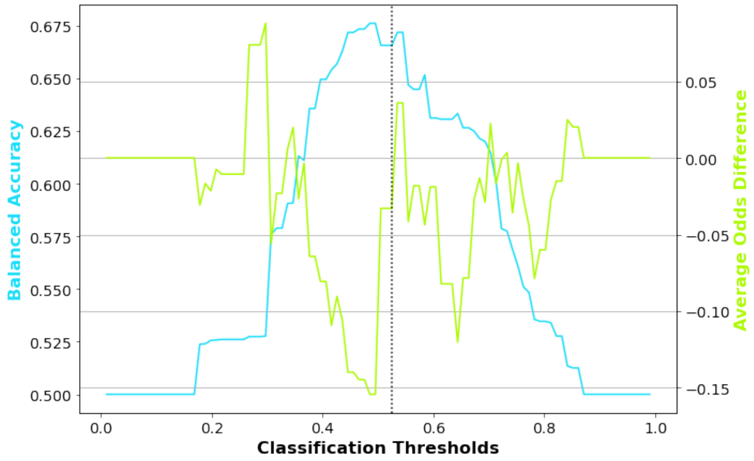


(h) Disparate Impact

B.4 Re-weighing for Gender in COMPAS



(i) Average Odds Difference



(j) Disparate Impact