Empirical Evaluation of Counterfactual Fairness in the Context of Statistical Fairness

Bachelor's Thesis for Artificial Intelligence

Richard Li

June 2019

Supervisor:
prof. T.M. Heskes
Radboud University Nijmegen
Institute for Computer Science and Information Science

Second Reader:
dr. W.F.G. Haselager
Radboud University Nijmegen
Donders Institute for Brain, Cognition and Behaviour

Abstract

Machine learning is on the rise and has been in the past decade. Only in the last few years, a distinct focus on fairness in machine learning has surfaced. As machine learning systems become more influential and widespread, the need arises to ensure that decisions that follow from these systems are fair. A variety of fairness definitions have already been proposed to serve that purpose. Counterfactual fairness is one such definition and is the focal point of this study. The abundant fairness definitions that sometimes clash can complicate the realization of fairness. In this study, counterfactual fairness is examined in conjunction with statistical fairness to investigate whether counterfactual fairness also allows for statistical fairness. A counterfactually fair predictor is constructed on a real-world data set about loan applicants, and this predictor is empirically evaluated on a set of statistical fairness definitions. The results suggest that counterfactual fairness can promote statistical fairness depending on which definitions are considered.

Contents

1 Introduction

Fair machine learning is an area of research that has become increasingly popular in the recent years. As machine learning continues to take a larger role in our lives, it becomes imperative to ensure that these machine learning algorithms do not carry over biases contained in the data.

Human bias is real and it has been reported that there sometimes is racial, social class, and gender bias in clinical judgement (Garb, 1997). A more pertinent example in the context of machine learning would be the COMPAS risk tool. The COMPAS risk tool is a system that assigns risk scores on how likely it is for a criminal defendant to re-offend. An analysis by ProPublica (Larson, Mattu, Kirchner, & Angwin, 2016) exposed that the COMPAS risk tool exhibits machine learning bias. The risk tool was biased against black defendants by being more likely to incorrectly assign a high risk score (higher chance of re-offending) to them, whereas white defendants were more often incorrectly assigned a low risk score. Several papers about fair machine learning discuss and evaluate this domain about recidivism (Kleinberg, Mullainathan, & Raghavan, 2016; Berk, Heidari, Jabbari, Kearns, & Roth, 2017).

There are various ways in which biases can creep into the data. Two common examples are sampling bias and human/historical bias. The former can be prevented by common sampling techniques such as (stratified) random sampling, however the latter is hard to avert before data collection because the data provided by the human might be inherently biased. Several formalizations of fairness have already been proposed in order to mitigate the inevitable biases in machine learning algorithms. Nevertheless, the issue of unfairness due to biases persists as there is no universal fairness definition that is considered applicable in all contexts.

1.1 Terminology and Notation

The terminology and notation can somewhat differ in the literature while referring to the same concepts. Therefore, the terminology and notation that will be used henceforth is explicated in this section, primarily in natural language but also formally if possible.

A	protected attributes
X	non-protected attributes
U	latent variables
Y	outcome variable
$\hat{Y}$	predictor

In fair machine learning, there is a subset of attributes that are called *protected* (sensitive) *attributes*, such as race and gender, which are viewed as the source of biases in the outcome. The subset of non-protected attributes will be denoted with X, and the subset of protected attributes with A. The subsets X and A both contain *observed* variables. The unobserved variables are denoted as U, otherwise referred to as *latent* variables. The actual outcome to predict is denoted as Y, which is observed. The prediction by the model for Y is denoted as $\hat{Y}$ and is called the *predictor*. The objective in fair machine learning is then to construct a fair predictor, i.e., a predictor that satisfies a fairness definition.

The *generative models* that will be presented later on are of the form:

Form:	$Variable \sim Distribution(Parameters)$
Example:	$Y \sim Bernoulli(\theta)$

The example means that $P(Y|\theta)$ is a Bernoulli distribution.

1.2 Formalizations of Fairness

The most prominent fairness formalizations in the literature include fairness through unawareness, demographic parity, equalized odds, individual fairness, and counterfactual fairness. Gajane and Pechenizkiy (2018) present an overview of this collection of fairness formalizations and succinctly discuss their properties.

Fairness through unawareness is perhaps the simplest formalization. It states that the predictor is fair if no protected attributes are included in its prediction. However, fairness through unawareness does not work when there are so-called proxies in the data, which are other non-protected attributes that hold information about the protected attributes (Gajane & Pechenizkiy, 2018).

Demographic parity (Statistical parity/Group fairness) aims to achieve equal probabilities for the outcome over groups defined by the protected attribute. The concern with demographic parity, however, is that some individuals who are statistically speaking simply more 'apt' that belong to certain subpopulations of the protected attribute might be disadvantaged in order to realize equal portions for each group.

Equalized odds entails that the true positive rate and false positive rate for every group defined by the protected attribute should be equal. A relaxed version of equalized odds is equality of opportunity, which only requires the true positive rate to be equal instead (Hardt, Price, & Srebro, 2016).

Individual fairness assumes that individuals who are similar should produce similar outcomes, even when the protected attribute varies. The absence of a generally agreed upon similarity metric for individuals makes individual fairness a tricky definition to apply in practice.

Counterfactual fairness regards the predictor as fair if the outcome remains unchanged when the value of the protected attribute is flipped to the counterfactual value while the other attributes that do not depend on the protected attribute are held constant. Assumptions are required for the causal models on which counterfactual fairness relies.

Within fairness research a distinction is made between statistical fairness and individual fairness. Statistical fairness concerns itself with achieving parity in some statistical measure for groups (e.g., demographic parity and equalized odds). Individual fairness, as the term may suggest, strives to attain equality between individuals rather than groups. Chouldechova and Roth (2018) discuss the strengths and weaknesses of both types of fairness.

The statistical definitions of fairness mentioned here make an appeal to intuition, however some combinations of these definitions are mathematically mutually exclusive (Friedler, Scheidegger, & Venkatasubramanian, 2016; Kleinberg et al., 2016; Berk et al., 2017).

1.3 Counterfactual Fairness

Counterfactual fairness (Kusner, Loftus, Russell, & Silva, 2017) relies on causal models to achieve fairness and as such needs to make assumptions to model the causal relationships between attributes. Counterfactual fairness falls within the more recent causal notions of fairness and operates more on an individual-level like individual fairness.

With counterfactual fairness questions such as '*Would this black defendant have been assigned the same risk score if he/she was white?*' can be asked. Intuitively, counterfactual fairness entails that the same individual should receive the same treatment had their protected attribute(s) been different. However, when changing the protected attribute(s), any other attributes that causally depend on the protected attribute will also change. Thus a causal model is necessary to capture these dependencies between various attributes.

Kusner et al. (2017) give the following definition of counterfactual fairness:

> **Counterfactual fairness:** A predictor $\hat{Y}$ is counterfactually fair if under any context $X = x$ and $A = a$,
>
> $$P(\hat{Y}_{A\leftarrow a}(U) = y | X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y | X = x, A = a)$$
>
> for all y and for any a' value attainable by A.

The idea behind counterfactual fairness is to model latent variables U in a causal model that allow Y to be predicted correctly, yet fairly as well. At prediction time, the protected attributes are omitted, and instead the learned latent variables together with the non-protected attributes are used to make a fair prediction. It should also be noted that any attributes in X that are descendants of A should not be used at prediction time. This follows from a lemma by Kusner et al. (2017, p. 5) that states that $\hat{Y}$ *is counterfactually fair if it is a function of the non-descendants of* A.

The *FairLearning* algorithm by Kusner et al. (2017, p. 6) allows for the construction of a counterfactually fair predictor once the causal model has already been established. In this algorithm the parameters of the causal model are firstly learned together with the latent variables U. When there is new data, the causal

model with learned parameters is used to learn the latent variables U for the new data. U together with any non-descendants of A in X are then used for the predictor $\hat{Y}$, which will then be counterfactually fair.

Suppressing the causal pathways in the data that lie at the very foundation of biases seems like a more sensible approach rather than choosing one of the many statistical constraints that do no inform us in any way about the origin of the biases. However as mentioned before, there is no agreement upon a fairness definition, as there are many, that appeals to everybody. Therefore there is a need to reconcile various fairness definitions.

1.4 Counterfactual Fairness and Statistical Fairness

The causal models that need to be specified for counterfactual fairness could be used to reason about statistical fairness. By analyzing the structure of the causal model, independence relations between variables can be deduced, and in combination with statistical fairness definitions that can be expressed in terms of relations between variables this allows us to reason about whether a causal model would achieve certain statistical fairness definitions. In the same vein as Hardt et al. (2016, p. 12-13), by looking at the relations between variables in the causal model (or dependency structure in their case) we can determine whether equalized odds would be satisfied when $\hat{Y}$ and A are independent conditional on Y ($\hat{Y} \perp\!\!\!\perp A \mid Y$).

The following two scenarios assume that we want to achieve counterfactual fairness, which means only non-descendants of A may be used for the predictor $\hat{Y}$. When doing so, we want to deduce whether the statistical fairness definitions of demographic parity and equalized odds would be satisfied by consulting the structure of the causal model.

Scenario 1: Figure 1 shows a simple causal model that corresponds to Figure 1a from Kusner et al. (2017, p. 4), but now additionally includes $\hat{Y}$ to reason about its relation to the other variables. Demographic parity would be achieved for this causal model when only U is used, because $\hat{Y}$ is independent of A as U is independent of A. Moreover, from the causal model it also follows that the predictor $\hat{Y}$ satisfies equalized odds when only using U. $\hat{Y}$ and A are independent conditional on Y in the model, which is what is required by equalized odds.

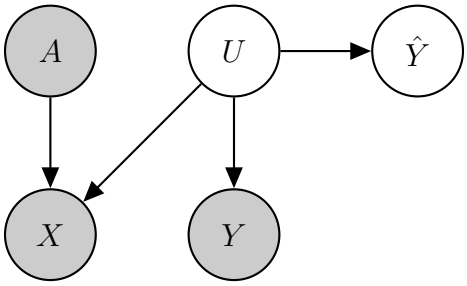

Figure 1: A simple causal model for Scenario 1 adapted from Kusner et al. (2017, p. 4) where an additional variable $\hat{Y}$ is added.

Scenario 2: The causal model in Figure 2 is similar to the one from Scenario 1, but it includes an extra edge from X to Y, which corresponds to Figure 1b from Kusner et al. (2017). Similarly, demographic parity is achieved in this model for the same reasons as in Scenario 1. However, equalized odds is not satisfied. The predictor $\hat{Y}$ and A are not independent anymore conditional on Y. There now is a path $A \to X \to Y \to U \to \hat{Y}$.

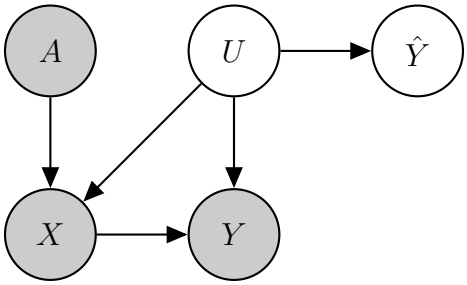

Figure 2: A simple causal model for Scenario 2 adapted from Kusner et al. (2017, p. 4) where an additional variable $\hat{Y}$ is added.

Thus depending on the structure of the causal model that is used for counterfactual fairness, it is sometimes possible to also achieve certain statistical fairness definitions. It is, however, not the objective to create a causal model that would satisfy statistical fairness, but one that would reflect the real world as accurate as possible.

1.5 Research Question

> *To what extent can a counterfactually fair predictor satisfy common statistical fairness definitions without explicitly imposing statistical constraints?*

The objective is to evaluate to what extent a predictor that is counterfactually fair can satisfy some common statistical fairness definitions. At the same time, no statistical constraints are imposed on this predictor, and the causal model is not tailored to achieve certain statistical fairness but is designed to reflect the real world as is required for counterfactual fairness. As mentioned before, there is a need to reconcile the many fairness definitions. This study tries to determine empirically if and to what extent counterfactual fairness and statistical fairness can be achieved alongside each other.

2 Method

2.1 Data Set

The data set that is used in the empirical evaluation of counterfactual fairness is the *German Credit Data* (GCD) data set[1]. This data set is chosen because it has been used before in fairness research (Chiappa & Gillam, 2018; Zemel, Wu, Swersky, Pitassi, & Dwork, 2013) and it has convenient properties[2] with respect to the objective of the study.

The data set has 20 attributes that are both numerical and categorical and contains 1000 records, which will be split 80/20 into a training and test set. The task associated with this data set is a binary classification of credit risk. A customer can either be good or bad in terms of credit risk, and as such this is encoded in a binary variable. Contextually a person with a good credit risk is likely to successfully repay the loan whereas a person with a bad credit risk is likely to default on a loan. Throughout this study, good (0) is considered the negative class and bad (1) is considered the positive class. The protected attribute in this data set would be the sex of the customer. Customers should not be discriminated against based on their sex.

		Predicted	
		good (0)	bad(1)
Observed	good (0)	0	1
	bad (1)	5	0

Table 1: Cost matrix for the GCD data set.

A cost matrix is supplied with this data set, which therefore should be taken into account in the evaluation of the model performance. False negatives are punished more severely than false positives according to the cost matrix. In other words, "*it is worse to class a customer as good when they are bad (5), than it is to class a customer as bad when they are good (1)*"[1].

[1]Source: `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`

[2]It has a binary protected attribute and a binary outcome that allow for a more straightforward computation of the statistical metrics.

A 'selection rate' of 30%[3] will be applied when classifying credit risk. This means that the top 30% of individuals assigned with the highest predicted credit risk probability will be classified as a bad credit risk. The selection rate can be interpreted as a way of screening individuals that are most likely to be a bad credit risk, i.e., individuals who have a high predicted credit risk probability.

From the 20 attributes, only a limited selection is used for the models. In total 7 attributes are chosen based on the causal model, which is explained in further detail in the next section. The 7 attributes are `amount`, `duration`, `housing`, `status`, `savings`, `age`, and `sex`, where `sex` is the protected attribute. The outcome variable to predict is then `credit_risk`.

$$A = \{\texttt{sex}\}$$

$$X = \{\texttt{amount, duration, housing, status, savings, age}\}$$

$$Y = \{\texttt{credit\_risk}\}$$

The attributes `amount`, `duration` and `age` are continuous variables. `amount` indicates the credit amount of the loan, `duration` is the duration of the loan in months and `age` is the client's age. The attributes `housing`, `savings` and `status` are categorical variables with varying levels. `housing` gives information about the type of housing of the client (e.g., rent or own), `savings` is the status of the savings account (e.g., $\geq$ 1000 DM[4]) and `status` is the status of the existing checking account (e.g., $\geq$ 200 DM). Lastly, `sex` and `credit_risk` are binary variables, where `sex` is the client's sex and `credit_risk` is an indicator if the client is considered a good or bad credit risk.

2.1.1 Pre-processing

The GCD data set contains no missing values. There are no attribute headers in the original data set but these have been added manually to capture what each attribute represents. The protected attribute sex is not explicitly one of the attributes but is contained in the attribute `personal_status_sex`, which combines

[3]The selection rate of 30% is chosen because that is the percentage of true bad credit risks in the GCD data set.

[4]DM = Deutsche Mark

the sex and marital status of a person. Sex is extracted from this compound attribute to use it to learn a counterfactually fair predictor with respect to sex. The binary attribute to predict `credit_risk` originally can take the values 1 (good) and 2 (bad). These values have been changed to 0 and 1 respectively to allow it to be incorporated into a causal model and to be used as the outcome in logistic regression.

A log transformation is applied to the continuous variables for normalization and these are also transformed to have zero mean ($\mu = 0$) and unit variance ($\sigma = 1$). The categorical variables are transformed to binary variables in order to be used in the implementation of the causal models, which is described later on. To transform categorical variables to binary variables, multiple levels are grouped together in such a way that there are two groups, which were then given a value of 0 and 1. The binary split was implemented by first ordering the levels in the categorical variables and then trivially splitting the levels somewhere in the middle to group them.

The categorical variables are not used as is because it is quite complicated to incorporate categorical variables into the causal model as rather strong prior knowledge is required about the probability distribution of the various levels from the categorical variable. Converting the categorical variables to ordinal variables was considered, but a similar problem arises as cut points have to be defined for which the values are not known.

2.2 Causal Model

A causal model is required to learn latent variables from the data. These latent variables should be predictive as well as fair because they will replace all the protected attributes and their descendants at prediction time. Chiappa and Gillam (2018) also work with the GCD data set and provide a causal model for path-specific counterfactual fairness in their paper. This causal model will be used as a basis because causal modeling is not the objective of this study nor do we have sufficient domain knowledge to construct such a causal model.

Kusner et al. (2017) introduce three levels for modeling the causal model with increasingly stronger assumptions. The second level is chosen where latent vari-

ables are postulated that are the '*non-deterministic causes of observable variables*' because in terms of strength of assumptions it is in the middle ground and the first level is not viable as there are too few non-descendants of A in the causal model as adapted from Chiappa and Gillam (2018). We do not want to make too strong assumptions as we lack expert input or domain knowledge and therefore level 2 is chosen over level 3.

2.2.1 Graphical Model

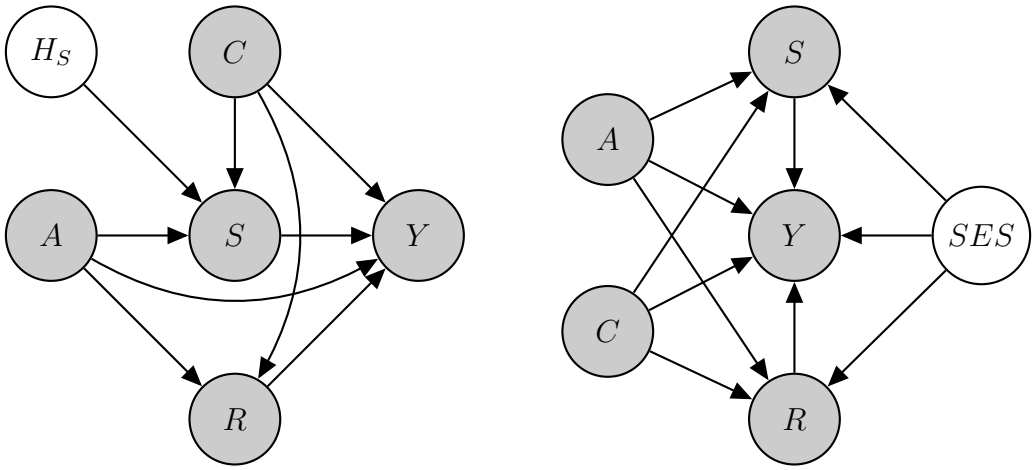

Figure 3: Causal models for GCD data set. (*Left*): adapted from Chiappa and Gillam (2018). (*Right*): modified version of the left causal model for the present study. The original model is rearranged and H_s is removed while SES is added.

Figure 3 shows the causal model as adapted from Chiappa and Gillam (2018) and the modified version of it for this study. For consistency the naming scheme of the nodes in the model is carried over. Additionally, SES is added to the set of latent variables U.

$$
\begin{aligned}
A &= \{\texttt{sex}\} \\
C &= \{\texttt{age}\} \\
R &= \{\texttt{amount, duration}\} \\
S &= \{\texttt{housing, savings, status}\} \\
U &= \{\texttt{SES}\} \\
Y &= \{\texttt{credit\_risk}\}
\end{aligned}
$$

The causal model remains largely the same but it is rearranged slightly. The latent variable H_s[5] is removed and instead the latent variable SES (socio-economic status) is introduced as a parent to S, R and Y. The assumption is that the latent variable SES influences S, R and Y, and that it propagates information about S and R (descendants of protected attribute A) to the predictor $\hat{Y}$.

2.2.2 Generative Model

The distributions of the variables in the causal model are listed below. The variables that encompass multiple attributes, i.e., R and S, use a compressed notation. The notation for these two variables should be interpreted in the way that all the attributes that are encompassed by that variable share the same type of distribution.

$$
\begin{aligned}
\tau_R &\sim Gamma(1,1) \\
R &\sim \mathcal{N}(r_0 + r_U \cdot U + r_A \cdot A + r_C \cdot C,\ \tau_R) \\
S &\sim Bernoulli\Big(\frac{1}{1+\exp\left(-(s_0 + s_U \cdot U + s_A \cdot A + s_C \cdot C)\right)}\Big) \\
U &\sim \mathcal{N}(0,1) \\
\theta &= \frac{1}{1+\exp\left(-(y_0 + y_U \cdot U + y_A \cdot A + y_C \cdot C + y_R \cdot R + y_S \cdot S)\right)} \\
Y &\sim Bernoulli(\theta)
\end{aligned}
$$

The parameters of the distributions are mostly computed using regression. The regression functions all have an intercept and are parameterized by their parents as defined in the causal model. The outcome variable Y is binary and thus a Bernoulli distribution is suitable. Y is essentially estimated with logistic regression by applying the logit function to θ, which results in a range of probabilities between 0 and 1. Similarly, the binary attributes in S are also estimated in this way. The continuous variables in R take on a Gaussian distribution with a mean that is estimated with linear regression. The precision τ_R[6] of this Gaussian distribution is defined with a Gamma distribution, which is a conjugate prior for precision. The latent variable U also has a Gaussian distribution with a set mean of 0 and

[5] Chiappa and Gillam (2018) introduced H_s for path-specific counterfactual fairness but did not explain what H_s represents.

[6] JAGS uses precision τ to parameterize the Gaussian distribution as opposed to standard deviation (Stan) or variance.

precision of 1. The values of U are unknown and as such this is expressed by using a weakly informative prior for U.

2.3 Implementation

The implementation of counterfactual fairness follows the code from Kusner et al. (2017) closely, which is provided on Github[7]. The causal modeling language of choice is JAGS and is used in conjunction with the statistical language R.

```
model {
  ...
  for (i in 1:N) {
    u[i] ~ dnorm(0, 1)

    amt[i] ~ dnorm(amt0 + amt_u * u[i] + amt_a * a[i] + amt_c * age[i],
      amt_tau)
    dur[i] ~ dnorm(dur0 + dur_u * u[i] + dur_a * a[i] + dur_c * age[i],
      dur_tau)

    hous[i] ~ dbern( 1 / (1 + exp(-(hous0 + hous_u * u[i] +
      hous_a * a[i] + hous_c * age[i]))) )
    sav[i] ~ dbern( 1 / (1 + exp(-(sav0 + sav_u * u[i] +
      sav_a * a[i] + sav_c * age[i]))) )
    stat[i] ~ dbern( 1 / (1 + exp(-(stat0 + stat_u * u[i] +
      stat_a * a[i] + stat_c * age[i]))) )

    logit(theta[i]) <- y0 + y_u * u[i] + y_a * a[i] +
      y_amt * amt[i] + y_dur * dur[i] + y_c * age[i] +
      y_hous * hous[i] + y_sav * sav[i] + y_stat * stat[i]

    y[i] ~ dbern(theta[i])
  }
}
```

Figure 4: JAGS code snippet for the causal model.

[7]Github repository: `https://github.com/mkusner/counterfactual-fairness`

The implementation consists of mainly two parts. The first part includes the two causal models. One of the causal models is used to learn parameters of the model with training data. The other causal model is used to learn the latent variables for the test data while using the learned parameters from the previously trained causal model. In this first part, the *FairLearning Algorithm* (Kusner et al., 2017, p. 6) is performed. Figure 4 shows the (partial) JAGS code for the training causal model. The part where the priors are specified for variables such as `amt0` and `dur_u` is not displayed as all the priors are just weakly informative priors with a Gaussian distribution ($\mu = 0$, $\tau = 1$) with the exception of `amt_tau` and `dur_tau`, which have a Gamma distribution as described in the *Generative Model*. The causal model for the test data is similar, but it does not need priors and Y (and thus also θ) is left out because Y is supposedly unknown for the test data.

The second part of the implementation are the actual classifiers. For the classifiers a generalized linear model is used for logistic regression. These classifiers only use U (SES) and C (age) as independent variables to predict Y (credit risk) fairly as both U and C are non-descendants of A.

2.4 Analysis

2.4.1 Counterfactual Fairness

The counterfactually fair predictor (Fair model) will be compared to a Full model and Unaware model, which serve as baselines, in terms of performance. The Full model contains all observed attributes[8] including the protected attribute and its descendants. The Unaware model uses all observed attributes[8] except for the protected attribute, which follows the fairness through unawareness definition.

Additionally, the Full and Unaware model are empirically tested for counterfactual fairness. We would like to find out whether these two models are counterfactually fair or not in order to properly analyze the results. To empirically assess counterfactual fairness, samples have to be drawn from the causal model. These samples either have the original sex from the original data (original samples) or

[8]All attributes as in the ones included in the causal model except U, not all 20 original attributes.

the opposite sex (counterfactual samples). The parameters of the causal model are learned by fitting it to all of the the original data before sampling takes place. For the original samples, all attributes (R, S, Y, U) are sampled except A (sex) as that is set to the original value of A.

For the counterfactual samples, we make an intervention on A and set it to the counterfactual value. The non-descendants of A are held constant[9], which are C (age) and U (socio-economic status) in this model, while the descendants of A are sampled (R, S, Y). This is in line with the counterfactual fairness definition: '*In other words, changing A while holding things which are not causally dependent on A constant will not change the distribution of $\hat{Y}$*' (Kusner et al., 2017, p. 3).

The distributions of predicted credit risk of the original and counterfactual samples are plotted to evaluate whether the predictor is counterfactually fair. The predictor can be considered counterfactually fair in the case that the distributions lie exactly on top of each other.

The Fair model itself is not tested for counterfactual fairness because it is counterfactually fair by design. Only the attributes that remained constant across the original and counterfactual samples are used by the Fair model, which means that the data is identical and will result in identical predictions for both samples.

2.4.2 Statistical Fairness Evaluation

The three statistical fairness notions mentioned earlier, which are demographic parity, equalized odds and equality of opportunity, are used to evaluate the counterfactually fair predictor in the context of statistical fairness as they are most commonly used.

The predictions by the Fair model are used to compute the necessary statistical metrics for the corresponding statistical fairness notions. The following definitions are conditional probabilities for the three statistical fairness notions with a binary predictor in the context of the GCD data set.

[9]Being held constant here means that the values of the non-descendants of A are set to the same values as in the samples with original sex

Demographic parity

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = a')$$

To achieve demographic parity the probabilities for being a predicted good credit risk or predicted bad credit risk should be equal for male and female applicants. Because we are dealing with a binary outcome, equal probabilities for bad credit risk given sex implies equal probabilities for good credit risk given sex, and the same holds the other way around.

Equalized odds

$$P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = a') \wedge$$

$$P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = a')$$

For equalized odds the true positive rate (TPR) and false positive rate (FPR) of credit risk should be equal for male and female applicants. This means that the probability that a male is predicted to be a bad credit risk when he is indeed a bad credit risk should be equal to the probability that a female is predicted to be a bad credit risk when she is indeed a bad credit risk (TPR). Also it should hold that the probability of a male predicted to be a bad credit risk when he is in fact a good credit risk should be equal to the probability that a female is predicted to be a bad credit risk when she is in fact a good credit risk (FPR).

Equality of opportunity

$$P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = a')$$

Equality of opportunity is a relaxed version of equalized odds where only the true positive rate should be equal for male and female applicants. Satisfying equalized odds thus automatically means that equality of opportunity is also satisfied.

To get a more complete overview of the statistical properties of the counterfactually fair predictor, some other (less common) statistical measures are also included. Verma and Rubin (2018) provide an extensive overview of fairness definitions and incidentally also evaluated a logistic regression classifier on the GCD data set using some of those fairness definitions.

Predictive parity

$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = a')$$

For predictive parity to hold, the Positive Predictive Value (PPV) of the credit risk should be equal for male and female applicants. The PPV is the fraction of true positives out of all predicted positives. In other words, it is the probability of actually having a bad credit risk when predicted to have a bad credit risk.

Conditional use accuracy equality

$$P(Y = 1|\hat{Y} = 1, A = a) = P(Y = 1|\hat{Y} = 1, A = a') \wedge$$

$$P(Y = 0|\hat{Y} = 0, A = a) = P(Y = 0|\hat{Y} = 0, A = a')$$

Conditional use accuracy equality requires both the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV) of credit risk to be equal for male and female applicants. NPV is the counterpart of PPV that was described before. The NPV is the fraction of true negatives out of all predicted negatives. In other words, it is the probability of actually having a good credit risk when predicted to have a good credit risk. Satisfying conditional use accuracy equality then also means that predictive parity is satisfied.

Overall accuracy equality

$$P(\hat{Y} = Y, A = a) = P(\hat{Y} = Y, A = a')$$

Overall accuracy equality means that the accuracy within subgroups of the protected attribute should be equal. In this case, the accuracy for male and female applicants should be the same. Accuracy would be the fraction of correctly identified good and bad credit risks. Intuitively this means that predictions for the male subpopulation should not be more accurate than the predictions for the female subpopulation and vice versa.

Balance for positive class

$$E(\hat{Y}_P|Y = 1, A = a) = E(\hat{Y}_P|Y = 1, A = a')$$

Balance for positive class is somewhat different than the aforementioned fairness definitions as those deal with predicted labels. Balance for positive class needs

the expectation of the predicted probability of the credit risk for male and female applicants with actual bad credit risks to be equal. $\hat{Y}_P$ represents this predicted probability of the credit risk as opposed to $\hat{Y}$ which represents the predicted binary labels of credit risk. Males and females with actual bad credit risk should on average be assigned the same predicted probability of credit risk.

Balance for negative class

$$E(\hat{Y}_P|Y = 0, A = a) = E(\hat{Y}_P|Y = 0, A = a')$$

Balance for negative class is the counterpart of balance for positive class. Balance for negative class needs the expectation of the predicted probability of the credit risk for male and female applicants with actual good credit risks to be equal. Males and females with actual good credit risk should on average be assigned the same predicted probability of credit risk.

The Full model will also be evaluated on statistical fairness in the same way to help interpret the results of the Fair model. The Full model would be used if we would not care about fairness and just wanted to maximize accuracy, and in this sense the Full model serves as the baseline for statistical fairness. The aim in this comparison is then to find out how the counterfactually fair model, which does not impose statistical constraints, performs in statistical fairness compared to a presumably unfair model, which also does not impose statistical constraints. Implicitly the question here is if counterfactual fairness benefits statistical fairness as well.

3 Results

3.1 Baseline Comparison

The Full, Unaware, and Fair model are subjected to a performance analysis in order to deduce the performance decline by the Fair model. Generalized linear models are trained on a training set ($N_{train} = 800$) for all three models. Performance is expressed as accuracy on the test set ($N_{test} = 200$), but as there is also a cost matrix associated with this data set, a simple cost measure will be employed as well. The costs that are used are the ones as displayed in Table 1. As such, the cost measure is as follows:

$$cost = 5 \cdot \text{FN} + \text{FP}$$

where FN are the False Negatives and FP are the False Positives.

		PRED	
		good(0)	bad(1)
OBS	good(0)	TN	FP
	bad(1)	FN	TP

The table above describes in which position the TN, TP, FN and FP reside. In the current context, good (0) is regarded as the negative class and bad (1) is regarded as the positive class. *OBS* means the observed credit risk in the data set and *PRED* means the predicted credit risk of the classifier.

	Accuracy	**Cost**
Full	68.5%	187
Unaware	67.5%	193
Fair	64.5%	211

Table 2: Accuracy and cost on the test set ($N_{test} = 200$).

As expected, the Full model achieves the highest accuracy as it is allowed to use all available attributes to predict the outcome. The Unaware model performs slightly worse in accuracy compared to the Full model as it forgoes the use of the protected attribute as in fairness through unawareness. The Fair model uses only U (SES) and C (age) to fairly predict the credit risk but it has the lowest accuracy. Presumably this loss in accuracy is because of the trade-off between fairness and accuracy (Zafar, Valera, Rodriguez, & Gummadi, 2017).

Full			
		PRED	
		good (0)	bad(1)
OBS	good (0)	109	32
	bad (1)	31	28

Unaware			
		PRED	
		good (0)	bad(1)
OBS	good (0)	108	33
	bad (1)	32	27

Fair			
		PRED	
		good (0)	bad(1)
OBS	good (0)	105	36
	bad (1)	35	24

Table 3: Confusion matrices for the Full, Unaware, and Fair model.

Taking the cost matrix into consideration, the same trend in performance remains as the Full model performs the best in terms of cost, the Unaware model follows closely, and the Fair model places last.

Fair vs Full			
		Full	
		correct	wrong
Fair	correct	116	13
	wrong	21	50
p-value = 0.2299			

Fair vs Unaware			
		Unaware	
		correct	wrong
Fair	correct	114	15
	wrong	21	50
p-value = 0.4047			

Full vs Unaware			
		Unaware	
		correct	wrong
Full	correct	129	8
	wrong	6	57
p-value = 0.7893			

Table 4: Contingency tables of all pairwise combinations of the Full, Unaware, and Fair model. The listed p-values are obtained from the McNemar's test.

In order to determine whether the loss in accuracy is significant, the McNemar's test is performed for the three models in a pairwise fashion. The test returned insignificant ($\alpha = 0.05$) for all three pairwise comparisons, and as such we can assume that the performance loss of the Fair model is negligible compared to the Full and Unaware model.

3.2 Counterfactual Fairness

In this part the distributions of predicted credit risks of the Full and Unaware model are plotted to empirically evaluate whether they are counterfactually fair. The trained models from before were used to make predictions on the original sampled data and counterfactual sampled data. 'Raw' predictions, i.e., predicted credit risk probability, were used to be able to plot the density plots. The Kolmogorov-Smirnov test is performed to find out whether the densities for the predicted credit risk probabilities of the original and counterfactual sampled data differ significantly.

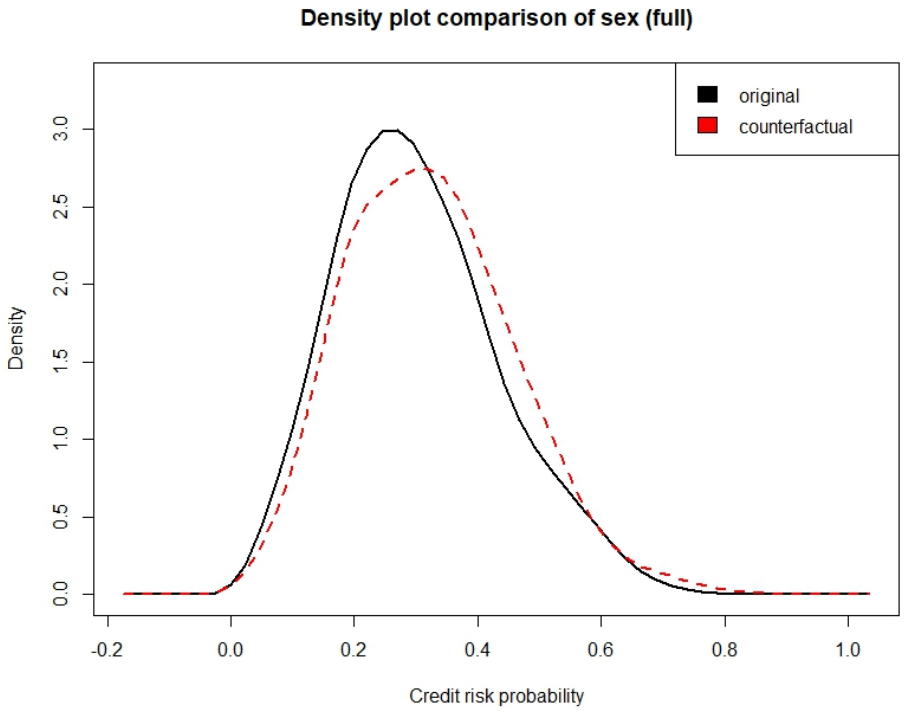

Figure 5: Density plot of predicted credit risk probability for the Full model.

The Kolmogorov-Smirnov test for the Full model reported a p-value of 0.0007261, which means that there is a significant difference in densities ($p < \alpha = 0.05$). These distributions do not lie on top of each other and together with the outcome of the statistical test this leads to the conclusion that the Full model is not counterfactually fair with respect to sex.

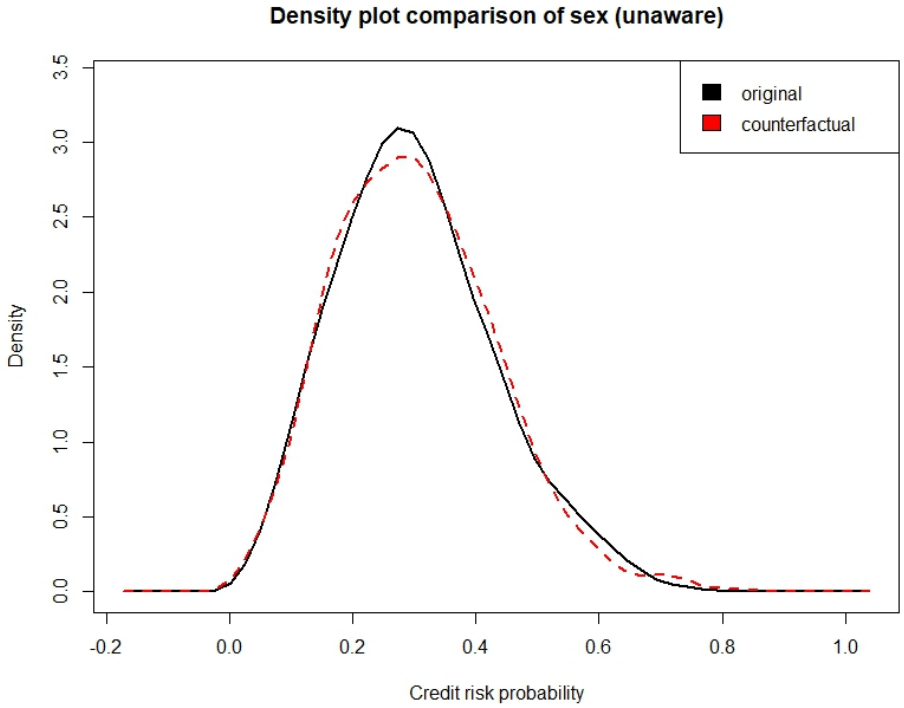

Figure 6: Density plot of predicted credit risk probability for the Unaware model.

The Kolmogorov-Smirnov test for the Unaware model reported a p-value of 0.9689, which means that there is no significant difference in densities ($p > \alpha = 0.05$). The distributions almost lie on top each other and together with the outcome of the statistical test this leads to the conclusion that the Unaware model is counterfactually fair with respect to sex.

As explained in the *Method*, the Fair model is counterfactually fair by design. The distributions would lie exactly on top of each other because only the non-descendants of A, which remained constant across the original and counterfactual sampled data, would be used for the prediction, resulting in identical predictions.

3.3 Statistical Fairness

The statistical metrics were computed post hoc using the predictions on the test set and the actual outcomes for the Full and Fair model. Predictions for the male and female subpopulations were separated and those were then used to compute the corresponding statistic with respect to sex. A test set of 200 records in combination with a selection rate of 30% is rather small to compute the statistical metrics[10]. To obtain more reliable results, the test set was artificially enlarged.

The test set was augmented by using five different 80/20 splits on the original data set, i.e., different seeds, for the training and test set. In this way there are five training and test set pairs. For each training and test set pair, the Full and Fair model would be trained separately on the training set and then predictions are made on the corresponding test set. The predictions on each test set are gathered and accumulated. Each of the five test sets has 200 records and therefore the effective test set size becomes $N_{test} = 1000$.

Table 5 and 6 show the computed statistical metrics for the Full and Fair model respectively. To be considered statistically fair, the odds for **Male** and **Female** should be equal or at least close together. The odds ratio is another way to conveniently observe whether the models are statistically fair. Preferably the odds ratio for each measure is close to 1, which simply entails that the odds do not really differ for males and females.

The p-values are also listed in these tables to test for significant differences between the statistics for males and females. All the statistical metrics except balance for positive or negative class use the Fisher's Exact probability test. As balance for positive and balance for negative class are expectations, a Welch Two Sample t-test is performed instead.

[10]With $N_{test} = 200$ and a selection rate of 30%, only 60 individuals will be classified as bad credit risks, which will be even less when having to divide it based on sex.

Full				
	Male	**Female**	**Odds ratio*$_{(M/F)}$**	***p*-value**
Demographic parity	0.2199	0.4717	0.47	$2 \cdot 10^{-15}$
True Positive Rate (TPR)	0.3651	0.6364	0.57	$7.633 \cdot 10^{-6}$
False Positive Rate (FPR)	0.1643	0.3846	0.43	$1.085 \cdot 10^{-9}$
Positive Predictive Value (PPV)	0.46	0.4667	0.99	1
Negative Predictive Value (NPV)	0.7744	0.7619	1.02	0.7523
Overall accuracy equality	0.7053	0.6226	1.13	0.01085
Balance for positive class	0.3354	0.4092	-	0.0006902
Balance for negative class	0.2573	0.3242	-	0.0006488

* Balance for positive and negative class do no have an odds ratio because these are not odds but expectations.

Table 5: Statistical fairness metrics expressed as odds and/or probabilities for male and female in the Full model. All the p-values are computed using the Fisher's Exact probability test except for balance in positive or negative class, for which a Welch Two Sample t-test is performed.

Fair				
	Male	**Female**	**Odds ratio*$_{(M/F)}$**	**p**-value
Demographic parity	0.2786	0.3459	0.81	0.0319
True Positive Rate (TPR)	0.4021	0.4182	0.96	0.808
False Positive Rate (FPR)	0.2312	0.3077	0.75	0.03683
Positive Predictive Value (PPV)	0.4	0.4182	0.96	0.8076
Negative Predictive Value (NPV)	0.7703	0.6923	1.11	0.03614
Overall accuracy equality	0.6672	0.5975	1.12	0.03349
Balance for positive class	0.3408	0.3573	-	0.5007
Balance for negative class	0.2962	0.2902	-	0.6332

* Balance for positive and negative class do no have an odds ratio because these are not odds but expectations.

Table 6: Statistical fairness metrics expressed as odds and/or probabilities for male and female in the Fair model. All the p-values are computed using the Fisher's Exact probability test except for balance in positive or negative class, for which a Welch Two Sample t-test is performed.

4 Discussion

4.1 Evaluation

Performance

The performance of the Fair model is the worst out of the three models, which include the Full and Unaware model. Both in terms of accuracy and cost, the Fair model is outperformed by the Full and Unaware model. However, this is expected as there would be a trade-off between fairness and accuracy. Furthermore, the accuracy drop does not seem to be significant compared to the two other models.

Counterfactual Fairness

The Full model is shown to be not counterfactually fair as the distributions do not lie on top of each other and the Kolmogorov-Smirnov returned significant. This is a somewhat important outcome when comparing the Full and Fair model in statistical fairness in the next section as we now know that we have a non-counterfactually fair model and a counterfactually fair model respectively, which allows us to evaluate whether counterfactual fairness can benefit statistical fairness. The Unaware model actually appears to be counterfactually fair. This result might indicate that there are no real proxies in the data. Or in the context of a causal model, there might be no descendants for A (sex).

Statistical Fairness

To evaluate to what extent the counterfactually fair model satisfies statistical fairness, we will look at the odds for male and female, the odds ratio, and the p-value of the statistical test in Table 5 and 6.

Demographic parity: the Full model does not achieve demographic parity and the Fair model does not either according to statistical test as both differences in odds are significant. From the odds and odds ratio we can also deduce that males and females do not have equal odds. Both models are more likely to classify females as bad credit risks. Interestingly is that the Fair model seems to come closer to demographic parity compared to the Full model when looking at the odds and odds ratio.

Equality of opportunity and equalized odds: the Full model does not satisfy equality of opportunity as the TPR for males and females differ significantly. Looking at the actual odds and odds ratio, the disparity is clearly noticeable. Naturally equalized odds is also not satisfied by the Full model as the TPR for males and females is unequal, and the FPR is unequal anyway. The Fair model does satisfy equality of opportunity but not equalized odds. The TPR for males and females is very similar reflected by an odds ratio close to 1. Although the FPR for the Fair model is not equal, it is at least closer to equal when looking at the odds (ratio).

Predictive parity and conditional use accuracy equality: the Full model is able to achieve predictive parity as well as conditional use accuracy equality. The Fair model only manages to satisfy predictive parity, but not conditional use accuracy equality. The PPV for males and females is almost the same with an odds ratio close to 1 for both the Full and Fair model. The Full model additionally has a NPV with an odds ratio close to 1. In the Fair model, the NPV for males is higher than for females, which can be interpreted as male applicants predicted to have good credit risk are more likely to have actual good credit risk compared to females.

Overall accuracy equality: neither the Full or Fair model managed to achieve overall accuracy equality. The accuracy for males is significantly higher than the accuracy for females in both models. The accuracy for both sexes in the Full model is higher than in the Fair model, which is consistent with the *Performance* evaluation. The models being more accurate for males might be caused by the fact that in original GCD data set the number of male (690) and female (310) records is unbalanced.

Balance for positive class or negative class: the Full model does not have balance for positive or negative class as the expectation of the predicted credit risk probability varies significantly for males and females. In the Full model, females are on average predicted to have a higher credit risk probability (i.e., more likely to be a bad credit risk) than males regardless of whether they are an actual good or bad credit risk. The Fair model, however, does have balance for positive and negative class, meaning that the model on average does not assign higher credit risk probabilities to females compared to males or the other way around.

To summarize, the Full model achieves predictive parity and conditional use accuracy equality. The Fair model achieves equality of opportunity, predictive parity, balance for positive class, and balance for negative class. Even though the Fair model does not satisfy demographic parity and does not have equal FPR with respect to sex, the Fair model does reduce the disparity in these measures compared to the Full model.

It should be noted that demographic (statistical) parity cannot be satisfied at the same time as the two balance conditions for positive and negative class provided that the base rates for sex are unequal as shown by Kleinberg et al. (2016). The Fair model results are consistent with this finding. Another incompatibility between statistical fairness definitions for which the results are consistent, is the incompatibility between conditional use accuracy equality and equality in the false negative rates (FNR) and false positive rates (FPR) as outlined by Berk et al. (2017). The FNR is not listed in Table 5 and 6 but can be easily computed and is shown in the table below.

	Male	**Female**	**Odds ratio** (M/F)	***p*-value**
FNR_{Full}	0.6349	0.3636	1.75	$7.633 \cdot 10^{-6}$
FNR_{Fair}	0.5979	0.5818	1.03	0.808

The FNR for the Full model is clearly not equal, whereas the FNR of the Fair model is close to equal. The Full model satisfies conditional use accuracy equality but does not manage to achieve equality in FNR and FPR, which is consistent with the previous statement by Berk et al. (2017).

Depending on which statistical fairness definitions are considered, the Fair model is more statistically fair than the Full model, but the other way around holds as well. However, overall the Fair model is able to satisfy more statistical fairness definitions than the Full model and also seems to come closer to fairness in measures that are not satisfied by either model.

4.2 Limitations

The accuracy of the Full, Unaware, and Fair model is not that impressive when comparing it to the accuracy of 76.0% achieved by Chiappa and Gillam (2018)

on the same data set for example. The categorical to binary conversion of the attributes might have played a part in the lower accuracy. Naturally information is lost when squashing together multiple levels into only two levels. The binary split itself might also not have been optimal as it was rather trivial resulting in lower performance overall. When running the Full model with normalized data that retained the categorical variables[11], the accuracy that is achieved is 74.5%. This supports the suspicion that the conversion from categorical to binary could have induced the deterioration in performance.

A crucial aspect of counterfactual fairness is the causal model. As Kusner et al. (2017) emphasize, it is important to consult experts and use domain knowledge to construct a causal model that reflects the real world. However, in this study the causal model is adapted from another paper and modified using assumptions that appeal more to intuition rather than domain knowledge. As such, the causal model in this study might not reflect reality accurately because of possibly flawed assumptions. In the case that this causal model does correspond to the real world, then there would still be the possibility of other causal models fitting the data just as well or better. One way to alleviate the problem when there are multiple plausible causal models is *Multi-World Fairness* (Russell, Kusner, Loftus, & Silva, 2017). In the approach of *Multi-World Fairness*, approximate counterfactual fairness is achieved in multiple causal models, and thus this method can be used whenever there is uncertainty about the proper causal model.

Regarding the implementation of counterfactual fairness, at first an attempt was made to run multiple chains to assess convergence when training the causal model. Soon, however, it became apparent that the chains would likely never converge by plotting the chains and after consulting the Gelman and Rubin's convergence diagnostic. This means that there is no unique configuration of parameters that fit the causal model and the data. Thus the decision was made to only use one chain in the implementation. A probable explanation is that both U itself and its coefficients (e.g., `y_u, amt_u, dur_u`) have to be learned as specified in the generative distributions of the causal model. The product of U and its coefficients can be factorized in many ways and thus the learned parameters in

[11]Exactly the same data and 80/20 split that was used in the *Performance* evaluation, but now without converting the categorical variables to binary variables.

different chains can diverge when the factorization differs. It is unclear whether certain chains with their parameters would yield better or worse performance than other chains, but this could be worth investigating.

4.3 Future Work

In fairness research it is often assumed that the protected attributes are given but the actual process of determining what attributes are considered sensitive is overlooked. Only sex is regarded as a protected attribute in the causal model used in this study. A question that is not exclusively related to counterfactual fairness but for fairness in general is that of what attributes should be considered sensitive. Evidently this is a context-dependent issue, but most likely also one of personal opinion. To illustrate in the case of the GCD data set, age might have been used as the protected attribute because of age discrimination. One could argue that good or bad credit risk should also not be biased towards young or old individuals alike. As a matter of fact, Zemel et al. (2013) use age as the protected attribute in their research for the GCD data set. On the other hand, age can be viewed as a predictive independent variables as it could be possible for older people to be more financially stable and as a result more likely to repay the loan, i.e., a good credit risk. However, this is mere conjecture to accentuate the difficulty in selecting the protected attributes.

Another aspect in fairness research that has not been explored much is the impact that fair machine learning may have in the long term. Liu, Dean, Rolf, Simchowitz, and Hardt (2018) studied the delayed impact of fair machine learning using two statistical fairness definitions, namely demographic parity and equality of opportunity. Their findings showed that in the long run, these fairness definitions may not improve fairness and may even be harmful. As counterfactual fairness is not a statistical fairness definition, it might be worthwhile to explore what kind of delayed impact counterfactual fairness can have.

5 Conclusion

This study set out to empirically evaluate counterfactual fairness in the context of statistical fairness without imposing statistical constraints. To this end, a counterfactually fair predictor for the GCD data set was constructed by following the *FairLearning* algorithm by Kusner et al. (2017). The counterfactually fair predictor is able to satisfy some of the statistical fairness definitions that were considered in this study, namely equality of opportunity, predictive parity, balance for positive class, and balance for negative class. All of the aforementioned definitions except predictive parity are not satisfied by the Full (non-counterfactually fair) model, but the Full model does manage to achieve conditional use accuracy equality for which the Fair model failed to do so, however that could be due to the incompatibility between statistical fairness definitions. Overall counterfactual fairness, at least in the GCD data set, can promote statistical fairness depending on which definitions are considered. Ultimately counterfactual fairness itself is just one of many fairness definitions. There will be fairness definitions that are regarded as fair by some and not by others. Therefore as long as there is no universal fairness definition that we can all agree upon, compromises will have to be made and the best we can do is to try to reconcile the multitude of fairness definitions.

6 References

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments:the state of the art. *arXiv:1703.09207*.

Chiappa, S., & Gillam, T. P. S. (2018). Path-specific counterfactual fairness. *arXiv:1802.08139*.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv:1810.08810*.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv:1609.07236*.

Gajane, P., & Pechenizkiy, M. (2018). On formalizing fairness in prediction with machine learning. *arXiv:1710.03184*.

Garb, H. N. (1997). Race bias, social class bias, and gender bias in clinical judgment. *Clinical Psychology: Science and Practice*, *4*(2), 99–120.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv:1610.02413*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv:1609.05807*.

Kusner, M., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How we analyzed the compas recidivism algorithm.* Retrieved from `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. *arXiv:1803.04383*.

Russell, C., Kusner, M. J., Loftus, J. R., & Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness - FairWare '18*.

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning*.

A Appendix

Github Repository

`https://github.com/rusane/counterfactual-fairness`

All the code and files from the thesis can be found on the Github repository. Some code files require `.Rdata` files, which were obtained by simply running the code and saving parts of it. The `.Rdata` files, however, are not on the repository as the files exceed the file size allowed by Github.

Categorical to Binary Conversion of Variables

housing
A151: rent A152: own A153: free

$$\{\text{A153}\} \to 0$$
$$\{\text{A151}, \text{A152}\} \to 1$$

savings
A61: $<$ 100 DM A62: $100 \leq ... < 500$ DM A63: $500 \leq ... < 1000$ DM
A64: $\geq$ 1000 DM A65: unknown/no savings account

$$\{\text{A61}, \text{A62}, \text{A65}\} \to 0$$
$$\{\text{A63}, \text{A64}\} \to 1$$

status
A11: $<$ 0 DM A12: $0 \leq ... < 200$ DM A13: $\geq$ 200 DM
A14: no checking account

$$\{\text{A11}, \text{A14}\} \to 0$$
$$\{\text{A12}, \text{A13}\} \to 1$$

Formulae for Statistical Fairness

All the formulae below are applied to the predictions on the subpopulations within a protected attribute.

TP	True Positives
FP	False Positives
FN	False Negatives
TN	True Negatives
$\hat{Y}_P$	predicted credit risk probability

$$\text{Demographic parity} = \frac{TP + FP}{TP + FP + FN + TN}$$

$$\text{TPR} = \frac{TP}{TP + FN} \qquad \text{FPR} = \frac{FP}{FP + TN} \qquad \text{FNR} = \frac{FN}{TP + FN}$$

$$\text{PPV} = \frac{TP}{TP + FP} \qquad \text{NPV} = \frac{TN}{TN + FN}$$

$$\text{Overall accuracy equality} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Balance for positive class} = \frac{1}{TP + FN} \sum_{i \in \{TP\} \cup \{FN\}}^{TP+FN} \hat{Y}_P^i$$

$$\text{Balance for negative class} = \frac{1}{TN + FP} \sum_{i \in \{TN\} \cup \{FP\}}^{TN+FP} \hat{Y}_P^i$$

Raw Data for Statistical Fairness

Full		
	Male	**Female**
TP	69	70
FP	81	80
FN	120	40
TN	412	128
N	682	318

Fair		
	Male	**Female**
TP	76	46
FP	114	64
FN	113	64
TN	379	144
N	682	318

Table 7: The predictions expressed as TP, FP, FN, and TN for male and female on the test set that were used in the evaluation on statistical fairness for the Full and Fair model.

Convergence Diagnostics

Trace plots of a small subset of parameters in the causal model are displayed in Figure 7 to assess convergence. The trace plots show that not all four chains converge. Convergence is implied in the trace plots when all chains are mixed together well, which is not the case in this example. The red and black chains diverge from the blue and green chains and do not seem to start converging in later iterations. The density plots also clearly show that two distinct values for each parameter are learned.

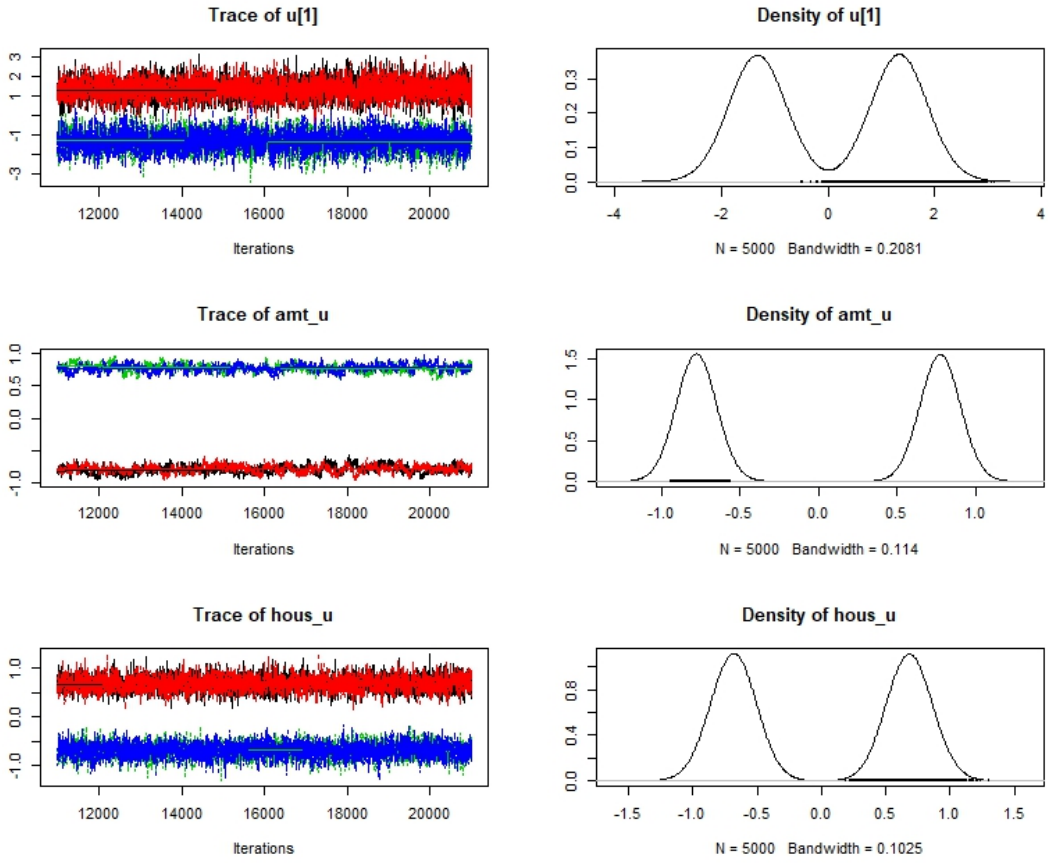

Figure 7: Trace plot of four chains from MCMC sampling for the parameters `u_1`, `amt_u`, and `hous_u`. The chains can be distinguished by the different colours (black, red, green, blue). The four chains do no converge for any of the three parameters and different values for the parameters are learned in the diverging chains.

Figure 8 shows the Gelman Rubin's Convergence Diagnostic plot for the same three parameters as before. The shrink factor[12] should be approximately 1.0 for all parameters to have reached convergence. Clearly the shrink factor is not anywhere close to 1.0 and the shrink factor also does not seem to move towards 1.0. Both the trace plot and the Gelman and Rubin's convergence diagnostic plot indicate that there is no convergence and that convergence will also probably not happen in even later iterations.

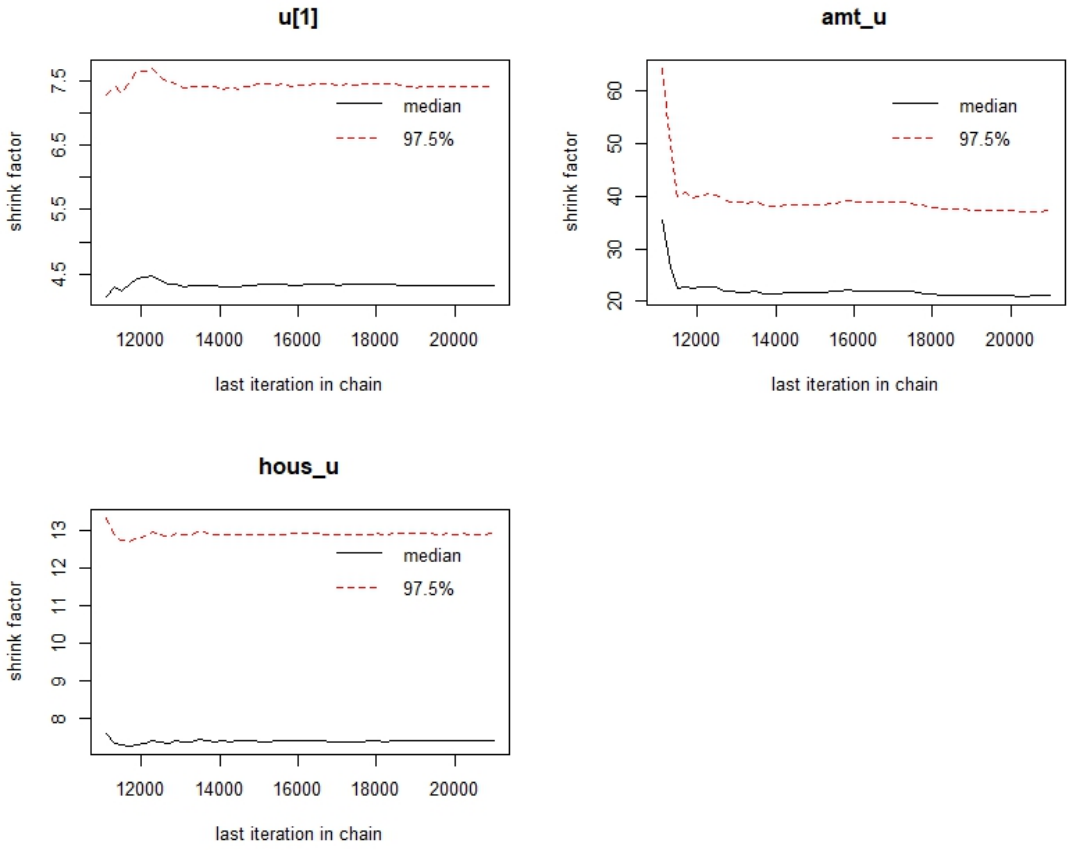

Figure 8: Gelman and Rubin's convergence diagnostic plot for the parameters `u_1`, `amt_u`, and `hous_u`. The shrink factor is not close to 1.0 at all and does not seem to move towards 1.0 in later iterations, which suggests that convergence is improbable.

[12]Also called the potential scale reduction factor (PSRF), which computes the ratio of the parameter variance within and between chains.