International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-021-00259-z

REGULAR PAPER

Analyzing the impact of missing values and selection bias on fairness

Yanchen Wang[1] · **Lisa Singh**[1]

Received: 4 July 2020 / Accepted: 26 April 2021

Abstract
Algorithmic decision making is becoming more prevalent, increasingly impacting people's daily lives. Recently, discussions have been emerging about the fairness of decisions made by machines. Researchers have proposed different approaches for improving the fairness of these algorithms. While these approaches can help machines make fairer decisions, they have been developed and validated on fairly *clean* data sets. Unfortunately, most real-world data have complexities that make them more *dirty*. This work considers two of these complexities by analyzing the impact of two real-world data issues on fairness—missing values and selection bias—for categorical data. After formulating this problem and showing its existence, we propose fixing algorithms for data sets containing missing values and/or selection bias that use different forms of reweighting and resampling based upon the missing value generation process. We conduct an extensive empirical evaluation on both real-world and synthetic data using various fairness metrics, and demonstrate how different missing values generated from different mechanisms and selection bias impact prediction fairness, even when prediction accuracy remains fairly constant.

Keywords Machine learning fairness · Missing data · Data bias · Selection bias

1 Introduction

In today's big data world, algorithmic decision making is becoming more pervasive in areas that impact our everyday lives, including hiring, credit approval, and criminal justice. For example, in the USA, machine learning models are being used throughout the judicial system, e.g., at bail hearings to predict whether a defendant will flee, prior to trials to determine the likelihood of the defendant committing additional crimes, and at sentencing to predict the length of sentencing [27,37]. As more applications use algorithmic decision making, there are growing concerns about their transparency, accountability, and fairness [7,16,49].

In the USA, the Civil Rights Act of 1964 prohibits discrimination of people based on race, color, religion, sex, or national origin. These demographic traits are examples of *sensitive/protected attributes* or attributes that should not be dominant features used by machine learning algorithms to make predictions. Sensitive attributes are identified based on the task being conducted and the established legal framework. For example, age is a sensitive attribute under the Equal Credit Opportunity Act (ECOA), but not the Fair Housing Act (FHA). For the purposes of this paper, we consider a sensitive attribute to be a demographic feature that is considered protected by specific legislation, i.e., by the Civil Rights Act.

Much literature has focused on algorithms for fixing bias, where the algorithms are validated on fairly "clean" data sets [12], including the COMPAS recidivism data, the German credit data, and the adult income data [20,31]. These data sets either do not contain missing values or contain a very small fraction of missing values for the categorical features. Previous work has focused on data sets that have few missing values and consider selection bias of continuous variables as opposed to categorical ones. While this previous work is an important first step to building fundamental theoretical frameworks, real-world data sets containing these additional complexities need to be understood in the context of fairness.

Toward that end, this paper investigates the impact of two different *data issues* on fairness. We focus on understanding fairness in the context of a sensitive binary attribute and non-sensitive categorical attributes in the presence of missing values and/or selection bias. We show examples of these forms of problematic data using both real-world and synthetic data and evaluate the impact of missing values and selection bias using existing fairness measures. Finally, we

✉ Yanchen Wang
yw516@georgetown.edu

Lisa Singh
Lisa.Singh@georgetown.edu

[1] Georgetown University, Washington DC, USA

Published online: 31 May 2021

propose data fixing approaches that use different reweighting and resampling techniques to improve fairness and evaluate the effectiveness of our proposed methods. We show that our approach can improve fairness while maintaining levels of predictive accuracy similar to those that lead to unfair classifiers.

Our main contributions can be summarized as follows. (1) Because real-world data sets contain categorical data with missing values and selection bias, we formulate the problem of missing values and selection bias on prediction fairness for categorical data. (2) We propose data fixing approaches that employ reweighting and resampling techniques to improve fairness in data sets containing missing values and/or selection bias in categorical data. (3) We conduct an extensive empirical evaluation on real-world and synthetic data using two established measures of fairness to understand the types of scores that should be expected under different data distributions. (4) We demonstrate the effectiveness of our proposed fixing algorithms on data sets containing different distributions and proportions of missing values and different amounts of selection bias. (5) We release our code and our synthetic data set so other researchers can continue to make progress on this problem [1].

The remainder of this paper is organized as follows. Related literature is presented in Sect. 2. In Sect. 3, we present our notation and some common real-world data issues. We then propose three fixing algorithms to mitigate the negative effects of these data issues in Sect. 4. In Sect. 5, we illustrate negative effects of data issues using synthetic data examples. We describe our data sets and evaluation method (Sect. 6) and empirically evaluate them in Sect. 7. Finally, Sect. 8 presents our conclusions and future work.

2 Related literature

In the statistical community, there exists a broad literature on missing values and selection bias. There is research describing different types of missing values [19,47], common causes of missing values [2,26,46] and problems that arise because of missing values [22,36,42]. While a number of negative impacts of missing values have been shown, the primary ones involve inference and interpretation issues with the models that are built [44,50]. There is also much work discussing data bias and selection bias, including the different causes and the negative effects of data bias [4,24,43,52]. Olteanu et al. consider this in the context of social media data and bias against race, gender, and age [43]. Baeza–Yates explores biases in digital content on the web, focusing on biases in news recommendation systems and tagging systems [4]. Wang and Wang discuss social influence bias in Amazon product reviews [52], while Kamishima and colleagues describe different types of computer system bias and how these biases arise as a result of historic selection bias [24]. In Sect. 3.3 and 3.4 , we will revisit the causes of some data issues and their impact.

Two main families of fairness definitions have been proposed by researchers: the statistical or group-based notions of fairness and the individual-based notions of fairness [18]. The statistical notion of fairness focuses on demographic parity and equalized odds [13,23,34]. The individual notion of fairness parallels the statistical notion in that we want similar individuals to be treated similarly [21]. We discuss the statistical notions of fairness in more detail in Sect. 3.2, where we formalize the definitions and measurements of fairness that we use.

Finally, a number of correcting algorithms have been proposed. These corrections can take place during pre-processing, in-processing, or post-processing. Pre-processing fixes attempt to modify the input data to make it less biased. Feldman et al. [23] propose methods to fix continuous data bias by changing the training data to remove dependencies between predictive features and sensitive attributes. Zemel and colleagues [54] and Calmon and colleagues [14] fix categorical data bias by changing conditional probability distributions of features based on the sensitive attribute and the outcome variable. In prediction tasks using external libraries such as pretrained embeddings in natural language processing (NLP), Bolukbasi and colleagues [8] find that word embeddings trained on Google News articles exhibit gender bias and they propose an algorithm that identifies words having a gender bias and then adjusts the word vectors to mitigate the bias. Brown and colleagues [10] find racial, gender and religious bias in their GPT-3 model trained on a dataset containing approximately one trillion words. The goal of in-processing fixes involves changing the constraints of the classifier to include fairness constraints. For example, some recent work modifies traditional classifiers, e.g., logistical regression and decision trees, to include a learning objective containing fairness constraints [33,35,53]. Post-processing algorithms change predicted labels after a classifier has been trained [28,45]. In other words, they do not change the input data or the constraints of the classifier. Instead, they modify the results of a learned classifier (the model) to guarantee fair prediction results on different groups. In this work, we propose corrections that will take place during pre-processing.

We also want to build connections across the work done in the statistics community and in the machine learning fairness community by mapping the existing fairness work to a more realistic data scenario. We frame fairness in the context of missing values and develop correction/fixing option to improve fairness while maintaining classifier accuracy. We focus on input data and pre-processing fixes. Current pre-processing fixing algorithms do not consider missing

[1] The code and the synthetic data can be found at https://github.com/GU-DataLab/fairness-and-missing-values.

values and/or selection bias. Our work fills this gap. Our approach is the first to characterize and quantify the impact of missing values on different data sets. We then extend the pre-processing fixing algorithm proposed by Calmon and colleagues [14] to handle both missing observations (selection bias) and missing values generated using different mechanisms. The original fixing algorithm learns probabilistic transformations to modify feature values and improve fairness. It was built and validated on data sets without missing values. We extend the algorithm by making missing values a new feature value in each non-sensitive feature, incorporating a new weighting scheme, and adjusting the selection process.

3 Preliminaries

This section begins with definitions and notation. We then present two real-world data issues: missing values and selection bias, explaining how they arise and defining them using our notation.

3.1 Definitions and notation

Let $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$ be the training set containing n individual observations and $\mathbf{r}_i$ represents the ith observation in the sample. Each observation contains one binary outcome, one categorical sensitive attribute and multiple categorical non-sensitive features. Let $\mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^p\}$ be the set of p non-sensitive categorical features such as education level and state of residence that are used as features for prediction, and let $\mathbf{x}^j = \{x_1^j, x_2^j \ldots x_n^j\}$ be a set of categorical feature values for the jth non-sensitive feature in $\mathbf{X}$. It can be the case that $\mathbf{X}$ contains missing values. We define missing values as an additional feature value, where $x_i^j = \emptyset$ if the feature value for the jth non-sensitive feature at the ith observation is missing. Let $\mathbf{M} = \{\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^p\}$ be a set of binary missing value indicators for the non-sensitive features $\mathbf{X}$ and $\mathbf{m}^j = \{m_1^j, m_2^j \ldots m_n^j\}$ be the set of binary missing value indicators for the jth non-sensitive feature in $\mathbf{X}$. We say that $m_i^j = 1$ if x_i^j is missing, and $m_i^j = 0$ if x_i^j is not missing for observations $i \in \{1, 2, \ldots, n\}$ and non-sensitive features $j \in \{1, 2, \ldots, p\}$.

Let $\mathrm{S} = \{s_1, s_2, \ldots, s_n\}$ be a binary sensitive attribute that is constructed from either a binary variable, e.g., gender, or a categorical variable, e.g., race, where s_i is the sensitive attribute value for the ith observation and $\mathrm{S} \notin \mathrm{X}$. If the sensitive attribute is categorical, it is converted to a binary attribute with two values, privileged group and unprivileged group. For example, our sample may have race as the sensitive attribute and $\{\texttt{White}, \texttt{Black}, \texttt{Hispanic}, \texttt{Asian}\} \in \mathrm{S}$. We convert these values into a binary attribute. We treat one or more races as the privileged group, and all the other races as the unprivileged group (1 vs. the rest). Hypothetically, we may treat `White and Asian` as the privileged group and `Black and Hispanic` as the unprivileged group in a classifier deciding on loan approvals. There is no universal rule on categorization—it is task-specific. [2] For the ith observation, we say $s_i = 0$ if the observation is in the unprivileged group and $s_i = 1$ if the observation is in the privileged group. Finally, in this work we make the assumption that a sensitive attribute does not contain any missing values. By doing this, we can better understand the impact of missing values on outcomes as it relates to the sensitive attribute without adding a confounding factor.

In this work, our classification task is binary. Our goal is to predict labels in Y using a set of non-sensitive features. Let $\mathrm{Y} = \{y_1, y_2, \ldots, y_n\}$ and for the ith observation, $y_i \in \{+, -\}$ with $+$ be a favorable outcome such as getting approved for a loan and $-$ be an unfavorable outcome. Similarly, let $\hat{Y}$ represent the predicted outcome with $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ and $\hat{y}_i \in \{+, -\}$ is the predicted label of the ith observation.

3.2 Fairness background

Bias is prejudice in favor of or against something or someone, usually in a way that is unfair and unfairness is different treatment of people based on a sensitive attribute.

Formal definition of fairness From a legal perspective, many anti-discrimination laws define unfairness using disparate treatment and disparate impact to show unfair treatment of people based on a sensitive attribute. Disparate treatment is intentional discrimination based on a sensitive attribute, whereas disparate impact is unintentional discrimination [3].

Researchers define fairness using two concepts, demographic parity and equalized odds [51]. Demographic parity requires that the predicted label be independent of the sensitive attribute. More formally, $P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1)$. Equalized odds requires that the prediction label and the sensitive attribute are independent conditional on the true class. That is, $P(\hat{Y} = 1|S = 0, Y = y) = P(\hat{Y} = 1|S = 1, Y = y), \forall y \in \{-, +\}$ [28]. It has been shown that except in trivial cases, any practically useful classifier satisfying equalized odds cannot be discriminatory [25].

Measurements of fairness A number of researchers have proposed different metrics for quantifying fairness. Feldman et al. and Zafar et al. [23,53] propose disparate impact ("p%-rule") to measure fairness. This metric is closest to the legal definition of fairness and is often used in anti-discrimination law to quantify fairness and discrimination. Chouldechova et al. [17] propose group conditioned fairness measures, including grouped false positive rate

[2] We understand that keeping each minority class separate is also advantageous for measuring fairness. We will explore larger domain categorical sensitive values in future work.

(s-FPR) and grouped false negative rate (s-FNR). These metrics are closely connected to the notion of equalized odds.

In this paper, we use $p\%$-rule and error rate balance to quantify fairness of classifiers. The $p\%$-rule is defined as:

$$\min\left(\frac{P(\hat{Y}=+|S=1)}{P(\hat{Y}=+|S=0)}, \frac{P(\hat{Y}=+|S=0)}{P(\hat{Y}=+|S=1)}\right)$$

The higher the value, the fairer the classifier is. Generally, if the $p\%$-rule is greater than 80%, or 0.8, the classifier is considered to be non-discriminatory [6]. Consider the example of a loan approval. Suppose gender is the sensitive attribute and getting approved for a loan is the positive outcome. The $p\%$-rule measures the ratio between the probability of getting approved if the customer is female versus male. If the approval rate for male applicants is significantly higher than female applications, we can say that this classifier is discriminatory based on gender.

Error rate balance is defined as balancing the false positive rate and false negative rate across all sensitive groups. In particular, the goal is to achieve: $P(\hat{Y}=-|S=0,Y=y) = P(\hat{Y}=-|S=1,Y=y), \forall y \in \{-,+\}$. When $y=+$, the constraint equalizes the false negative rate (FNR) across two sensitive groups. When $y=-$, the constraint equalizes the false positive rate (FPR). For a fair classifier, the error rate difference should be small across all sensitive groups. Using our loan approval example, the false negative rate (FNR) represents the how often an applicant is qualified for the loan, but the classifier decides that applicant is not qualified. The false positive rate (FPR) is the rate represents the rate at which an applicant is not qualified for a loan, but the classifier decides the applicant is qualified. The goal of error rate balancing is to make sure that the FPR and the FNR are similar and small.

3.3 Data issue: missing values

Missing values are very common in quantitative research especially in survey research [1]. King and colleagues show that approximately half of the respondents of political science surveys do not answer all of the questions [38]. Missing values occur at two different levels, the unit level and the item level [19]. A unit level missing value occurs when there is no information collected from a respondent and information about the respondent does not appear in the training set. If the non-response rate is different across sensitive groups, then unit level missing values can be viewed as a type of selection bias. An item level missing value occurs when a respondent does not answer all of the survey questions, and the incomplete information is represented as missing values in the training set. In this paper, when we discuss missing values, we are discussing item level missing values. When we discuss selection bias, we are considering unit level missing values. The missing value problem can be described based on different characteristics. We consider the following characteristics since they are prevalent in the literature: the proportion of missing values, the missing value mechanisms, and the patterns of missing values [19].

Proportion of missing values The proportion of missing values can affect the quality of the statistical analysis and prediction output. Schafer and colleagues [48] and Bennett and colleagues [5] suggest that statistical analysis is likely to be biased if the percentage of missing values is more than 5% to 10%. Let $U = \{u_1, u_2, \ldots, u_n\}$ be a binary variable indicating whether an observation contains one or more missing values. Recall $\mathbf{M}$ is the set of binary missing value indicators in $\mathbf{X}$ and $\mathbf{m}^j = \{m_1^j, m_2^j \ldots m_n^j\}$ is the set of binary missing value indicators for the jth non-sensitive feature in $\mathbf{X}$. We say that $m_i^j = 1$ if x_i^j is missing, and $m_i^j = 0$ if x_i^j is not missing. For the ith observation, $u_i = 0$ if there are no missing values for any of the variables in this observation, i.e., $m_i^1 = m_i^2 = \cdots = m_i^p = 0$, and $u_i = 1$ if the equality does not hold. We define the proportion of missing values in a data set as follows: $\lambda = \frac{\sum_{i=1}^{n} u_i}{n}$, where n is the number of observations.

Missing value mechanisms To understand different missing value mechanisms, we partition the training set $\mathbf{R}$ into two subsets: the observed data, $\mathbf{R}_{obs}$, and the missing data, $\mathbf{R}_{mis}$, where $\mathbf{R} = (\mathbf{R}_{obs}, \mathbf{R}_{mis})$. Previous work has defined three classes of generative mechanisms for missing values: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR) [47]. With MAR, the probabilities of missing values given the data set $\mathbf{R}$ depend only on observed data ($\mathbf{R}_{obs}$) [19]:

$$P(\mathbf{M}|R) = P(\mathbf{M}|\mathbf{R}_{obs}, \mathbf{R}_{mis}) = P(\mathbf{M}|\mathbf{R}_{obs})$$

For example, suppose we are given a training data set with one binary outcome, one binary sensitive attribute and three categorical non-sensitive attributes. Suppose education level is a categorical non-sensitive feature containing missing values. If the probabilities of missing values in the education level depend only on observed data, i.e., the sensitive attribute and the other two non-sensitive features, then those missing values are under MAR. With MAR, the proportion of missing values can vary across different categories of sensitive values as long as the probability of the missing values in the non-sensitive categorical feature (education level) depends only on observed data.

MCAR is a special case of MAR in which the probabilities of missing values depend neither on the observed data ($\mathbf{R}_{obs}$) nor the missing data ($\mathbf{R}_{mis}$) [19]. With MCAR, missing values are completely randomly distributed:

$$P(\mathbf{M}|\mathbf{R}) = P(\mathbf{M}|\mathbf{R}_{obs}, \mathbf{R}_{mis}) = P(\mathbf{M})$$

The last missing value mechanism is MNAR. It occurs when the probability of a missing value depends on both the observed values and the missing values themselves [19]:

$$P(\mathbf{M}|R) = P(\mathbf{M}|\mathbf{R}_{obs}, \mathbf{R}_{mis}) \neq P(\mathbf{M}|\mathbf{R}_{obs})$$

Using our education example, suppose the missing values in education level depend on the education level of the respondents. For example, respondents with a lower education level are less likely to disclose their information than respondents with a higher education level. We would say that those missing values are under MNAR.

With MAR and MNAR, there are multiple methods for filling missing values, the most popular of which is imputation [40]. There are no established methods for filling in missing data under MNAR, but Jakobsen and colleagues show that it is possible to fill missing data under MNAR with sensitivity analyses [30]. It can also be difficult to determine the missing value mechanism for a data set. It is possible to use Little's multivariate test [41] to determine the MCAR mechanism. However, the MAR and MNAR mechanisms cannot be distinguished using observed data [15,29].

Patterns of missing values Missing values can also be characterized based on the types of patterns they form. There are three patterns of missing values identified in the literature: univariate (multivariate), monotone, and arbitrary [48]. The univariate missing value pattern occurs if only one feature in the data contains missing values.[3] The monotone missing value pattern occurs when $\mathbf{x}^j$ is missing, and $\mathbf{x}^{j+1}$, $\mathbf{x}^{j+2}, \ldots \mathbf{x}^p$ are also missing. It typically arises during longitudinal studies if respondents drop out during the study period. Finally, the arbitrary missing value pattern is the most general pattern in which any set of variables may be missing from any observation in the data. For example, there may exist missing values in the outcome variable Y and/or in the sensitive attribute S. In this paper, we focus on the impact of missing values within non-sensitive features. Therefore, we do not consider missing values in Y and S, but instead only consider missing values in non-sensitive features $\mathbf{X}$ that have a univariate missing value pattern.

3.4 Data issue: selection bias

Selection bias occurs if observations from some groups in the sample are oversampled and others are undersampled. In this case, some groups are over represented and others are under represented. If we build a classifier using the biased data, the model would be biased toward some groups of users. Kamiran et al. explain that if a data set is unbiased, i.e., sensitive attribute S and outcome Y are statically independent, we have [32]:

$$P_{exp}(S = s, Y = y) = P(S = s) \times P(Y = y),$$
$$s \in \{0, 1\}, y \in \{-, +\}$$

$P_{exp}(s, y)$ is the probability of the sample distribution given statistical independence.

Selection bias can be caused by many types of sampling bias such as representation bias, population bias, and nonresponse bias. Table 1 summarizes each type of selection bias.

To quantify the severity of selection bias, we define *bias*, β as follows:

$$\beta = \sum_{s \in S} \sum_{y \in Y} \frac{|\,|\mathbf{R}(S = s, Y = y)| - |E(\mathbf{R}(S = s, Y = y))|\,|}{n}$$
$$\forall S = \{0, 1\}, Y = \{-, +\}$$

where $|\mathbf{R}(s, y)|$ is actual number of observations with sensitive attribute $S = s$ and outcome $Y = y$ and $|E(\mathbf{R}(s, y))|$ is the expected number of observations given statistical independence, $|E(\mathbf{R}(s, y))| = P_{exp}(S = s, Y = y) * n$. We take the absolute value of the difference between the expected number of observations and the actual number of observations in each group. With this metric, $\beta = 0$ if there is no selection bias and 1 if the data are extremely biased, i.e., the dataset contains only observations from one sensitive group and the individuals in the group have the same outcome.

4 Proposed fixing method

In this section, we present algorithms to mitigate the negative effects of missing values and selection bias on fairness. Our approach is to adapt strategies that have been used in other contexts to our scenario. We use the fairness fixing algorithm by Calmon et al. [14] as a starting point. This algorithm learns a probabilistic transformation to change feature values labels in the data to control discrimination. While this algorithm was developed with categorical data in mind, it does not account for missing values and selection bias. We adjust the transform to incorporate new reweighting and resampling schemes depending on the specific data characteristics.

4.1 Fixing biased missing values through reweighting

Our approach for addressing biases caused by missing values is to weight observations based on whether or not they contain missing values. When training our classifier, we assign a class weight that is determined by the imbalance in the

[3] Some researchers define the univariate missing value pattern slight differently, allowing for one or more features in the data to contain missing values [19].

Table 1 Definition and examples of different types of selection bias

Sampling bias	Definition	Example
Representation bias	The sample is collected from a small group and people in this group may share certain characteristics that are not representative of the population of interest	Can result from convenience sampling: a group that is easier to get information from
Population bias	Sample lacks diversity such as demographic diversity and geographical diversity, and therefore, the sample does not represent the overall population	The demographics of a social media sample does not generalize to the US population
Non-response bias	It occurs when a group of respondents has a significantly lower response rate than other groups	Optional surveys to employees may be completed by those who have more time

data. Class weights of most classifiers are based on the frequencies of feature values. We use this idea to address the missing value problem. We want the classifier to learn more information from observations that do not contain missing values than observations containing missing values. Thus, we assign a higher weight to observations without missing values and a lower weight to observations with missing values so as to oversample higher quality observations. Let W = $\{w_1, w_2, \ldots, w_n\}$ represent weights for observations, where w_i is the weight for the ith observation. We assign weights as follows:

$$w_i = \begin{cases} \frac{P(S=s_i, Y=y_i, U=0)}{P(S=s_i, Y=y_i)} & \text{if } u_i = 1 \\ \\ \frac{P(S=s_i, Y=y_i)}{P(S=s_i, Y=y_i, U=0)} & \text{if } u_i = 0 \end{cases}$$

Recall that S represents the sensitive attribute, Y represents the outcome variable, and U represents a missing value indicator. We see that the weight difference between missing and non-missing observations depends on the proportion of missing values in each sensitive group and on the outcome. In groups with a larger fraction of missing values, the weight for observations containing missing values is much smaller than the weight for observations with no missing values. In groups with a small proportion of missing values, weights are similar for observations with and without missing values.

Table 2 shows examples of the distribution of missing values and the weights for each observation with and without missing values in each group. Looking at the unprivileged group with a positive outcome, we see that the total number of observations in this group is 1500 and number of observations with missing values in this group is 600. With our weight formula, the weight for observations with missing values is $\frac{1500-600}{1500} = 0.6$ and the weight for observations without missing values is $\frac{1500}{1500-600} = 1.67$. The low weight for the observations with missing values means that these observations have a lower quality. Looking at the other rows in the table, we see that those observations with missing values all have a higher quality than the first row. We will evaluate the merits of this weighting scheme in Sect. 7.

4.2 Fixing selection bias using resampling

As previously mentioned, selection bias exists when some groups are oversampled and others are undersampled. When we have oversampled groups, we want to remove some observations. When we have undersampled groups, we want to add some observations. Researchers have shown that uniform resampling is a simple method to solve selection bias [32,39].[4] Algorithm 1 shows our approach for fixing selection bias with uniform resampling. The input to the algorithm is the set of observations $\mathbf{R}$, the sensitive attributes S, and the output labels Y. The output of the algorithm is $\mathbf{R}'$, our transformed data set after resampling. For each sensitive group and outcome, $|\mathbf{R}(s, y)|$ represents the actual number of observations in the training data and $|E(\mathbf{R}(s, y))|$ is the expected number of observations given statistical independence. In the oversampled groups, we want to uniformly resample $|\mathbf{R}(s, y)|$, and randomly drop these observations to reach the expected number of observations in that group (lines 3–5). In particular, the number of observations that needs to be dropped is equal to $|\mathbf{R}(s, y)| - |E(\mathbf{R}(s, y))|$. In the case of undersampled groups, we use bootstrapping, i.e., sampling with replacement, to increase the number of observations by randomly adding some repeated observations (lines 5–7). We bootstrap $|E(\mathbf{R}(s, y))| - |\mathbf{R}(s, y)|$ number of observations from $\mathbf{R}(s, y)$ and append them to the original training data. We pause to mention that our resampling algorithm can only fix selection bias that violates the statistical independence between the sensitive attribute and the outcome mentioned in Sect. 3.4. This algorithm cannot fix selection bias that requires external knowledge to identify. For exam-

[4] Instead of uniform resampling, we could extract contextual information from data and perform data augmentation based on the contextual information. Because of the potential bias introduced when doing this with categorical data, we leave that for future work.

Table 2 Distribution of missing values and weights

Sensitive group	Outcome	Total number of observations	Number of missing observations	Weight on missing observations	Weight on non-missing observations
Unprivileged	Positive	1500	600	0.6	1.67
Unprivileged	Negative	2500	250	0.9	1.11
Privileged	Positive	3000	500	0.83	1.2
Privileged	Negative	2000	400	0.8	1.25

ple, implicit bias due to policing interventions [9] may not be noticeable statistically if it is not more systemic.

Algorithm 1: Fixing selection bias

```
Input: R, S, Y
Output: R'
initialize R' = ∅ for s ∈ S do
    for y ∈ Y do
        if | E(R(s, y)) |<| R(s, y) | then
            R = random_drop(R(s, y)) append R to R'
        else
            R = bootstrap(R(s, y)) append R to R'
        end
    end
end
return R'
```

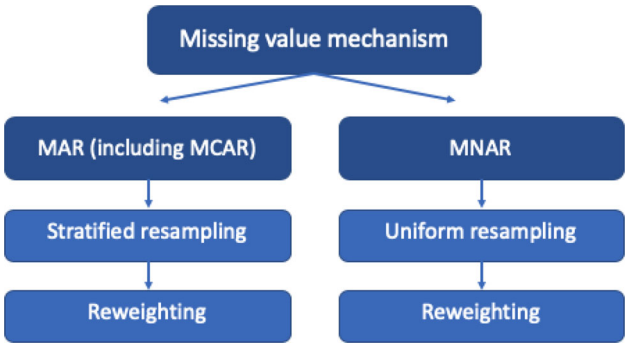

Fig. 1 Flowchart of our fixing algorithm with both selection bias and missing values

4.3 Fixing both selection bias and missing values

When both missing values and selection bias occur in the training data, we want to fix both problems by combining the fix algorithms proposed for missing values and selection bias. Because the impact of missing values on fairness differs depending upon the mechanism, we propose two different fixing algorithms, each specific to a particular missing value mechanism. Figure 1 shows the flowchart for our fixing algorithms when both selection bias and missing values are present in the data set. First, we need to determine the missing value mechanism. If the missing values are under MAR, including MCAR, we use stratified resampling then reweighting. If the missing values are under MNAR, we use uniform resampling then reweighting. If the missing value mechanism is unknown, we treat it as MNAR and use uniform resampling then reweighting.

Algorithm 2 shows the high level fixing algorithm. The input into the algorithm is the set of observations $\mathbf{R}$, the sensitive attribute S, and the output label Y. The output of the algorithm is $\mathbf{R}'$, our transformed data set after resampling and W, the weights after reweighting.

Stratified resampling and reweight Kamiran and Calders suggest preferential resampling, which focuses on observations that are more influential to decision making such as data points on the support vectors using SVM [32]. When missing values are under MAR, we want to fix selection bias using stratified resampling. In groups that are undersampled, unlike uniform resampling from the previous section, we sample with replacement from observations without missing values, i.e., we only choose observations from $\mathbf{R}_{obs}(s, y) = \{\mathbf{r}_i | u_i = 0, s_i = s, y_i = y\}$ for $i \in \{1, 2, \ldots, n\}, s \in \{0, 1\}, y \in \{+, -\}$ and add them to the original training data. In groups that are oversampled, we randomly drop some observations to reach the optimal number of observations. With stratified resampling, we can reduce the proportion of missing values and then apply our reweighting algorithm based on the number of missing values in each group in the training data after resampling.

Uniform resampling and reweight When missing values are under MNAR, the missing value distribution is biased. If we use stratified resampling, we are adding more bias into the training data. Therefore, with MNAR, in groups that are undersampled, we sample with replacement from all observations, i.e., uniform resampling. With uniform resampling, the proportion of missing values in the resampled data would remain constant, but uniform resampling can mitigate the negative effects caused by selection bias. Then, to mitigate missing value issues, we calculate weights as we show in Sect. 4.1 (Algorithm 2: lines 15–20).

Table 3 Accuracy, F1 score and fairness measures with different types of missing values on synthetic data

	No missing	MCAR	MAR	MNAR
Accuracy	0.758 (0.012)	0.755 (0.013)	0.756 (0.014)	0.759 (0.018)
F1 score	0.738 (0.016)	0.734 (0.021)	0.736 (0.017)	0.739 (0.017)
p%-rule	0.903 (0.09)	0.894 (0.011)	0.749 (0.019)	0.658 (0.02)
FNR for female	0.308 (0.01)	0.302 (0.009)	0.325 (0.012)	0.353 (0.019)
FPR for female	0.375 (0.012)	0.378 (0.011)	0.326 (0.01)	0.312 (0.017)
FNR for male	0.264 (0.008)	0.273 (0.009)	0.246 (0.007)	0.221 (0.007)
FPR for male	0.457 (0.015)	0.448 (0.013)	0.465 (0.016)	0.506 (0.019)
Number of missing values	0	1295	1302	1304
Proportion of missing values	0	0.21	0.212	0.212

Table 4 Distribution of training sets for selection bias

Sensitive value	Outcome	Balanced set	Bias level 1	Bias level 2
Female	Positive	1524	1300	1070
Female	Negative	1475	1700	1920
Male	Positive	1018	1170	1320
Male	Negative	983	830	690
Total		5000	5000	5000
Bias β		0	0.151	0.299

Algorithm 2: Fixing selection bias and missing values

Input: $(\mathbf{R}, S, Y)$
Output: $\mathbf{R}'$ and W
initialize $\mathbf{R}' = \emptyset$ A = missing value mechanism **for** $s \in S$ **do**
  for $y \in Y$ **do**
    if $| E(\boldsymbol{R}(s, y)) | < | \boldsymbol{R}(s, y) |$ **then**
      $\mathbf{R}$ = random_drop($\mathbf{R}(s, y)$) append $\mathbf{R}$ to $\mathbf{R}$'
    end
    if $| E(\boldsymbol{R}(s, y)) | > | \boldsymbol{R}(s, y) |$ **then**
      if $A = MNAR$ **then**
        $\mathbf{R}$ = bootstrap($\mathbf{R}(s, y)$) append $\mathbf{R}$ to $\mathbf{R}$'
      else
        $\mathbf{R}$ = bootstrap($\mathbf{R}(s, y, U = 0)$) append $\mathbf{R}$ to $\mathbf{R}$'
      end
    end
  end
end
for $i \in \{1, 2, \ldots, n\}$ **do**
  if u_i=*0* **then**
    $w_i = \frac{|\mathbf{R}(S=s_i, Y=y_i)|}{|\mathbf{R}(S=s_i, Y=y_i, U=0)|}$
  else
    $w_i = \frac{|\mathbf{R}(S=s_i, Y=y_i, U=0)|}{|\mathbf{R}(S=s_i, Y=y_i)|}$
  end
end
return $\mathbf{R}'$, W

5 Building intuition using synthetic data examples

In order to build some intuition about the impact of missing values and selection bias on fairness, we setup some simple experiments on synthetic data[5] that vary the distribution and the proportion of missing values, as well as the amount of selection bias. We use logistical regression to predict labels using all the non-sensitive features, and we use the p%-rule and error rate balance to measure fairness of the classifier.

All the data sets for this simple experiment contain four categorical non-sensitive features, one outcome, and one sensitive group. The outcome variable has two values, − and +. As an example, a positive outcome may indicate getting approved for a loan. A sensitive attribute may be gender with two values, `male` and `female`, and `male` may be considered the privileged group, i.e., men are more likely to a get positive outcome than women. In this experiment, we do not want to discriminate based on gender.

Missing values To illustrate the effect of missing values, we choose a non-sensitive feature and remove some of the existing feature values, making them missing values. We build four data sets: (1) one without missing values (this serves as our best case in terms of classification accuracy and fairness), (2) one with missing values under MCAR, (3) one with

[5] We will present how we create synthetic data in Sect. 6.1.

missing values under MAR, and (4) one with missing values under MNAR. We build our machine learning models using 5000 observations and evaluate the models using five fold cross-validation. All four data sets with missing data have a similar number of missing values, approximately 20% of total number of observations.[6]

Table 3 shows the average and standard error (in parentheses) for accuracy, F1 score, and the fairness measures (rows) for all four data sets (columns). From the results, we see that different mechanisms for generating missing values have different impacts on the F1 score and the fairness measures. In terms of prediction accuracy/F1 score, they are similar to the best case, i.e., no missing values. All three mechanisms have less than a 1% change in accuracy and F1 score. In contrast, the missing values, especially MAR and MNAR have a larger impact on fairness. Fairness using the p%-rule decreases by less than 1% for MCAR, approximately 15% for MAR and almost 25% for MNAR. The false negative rate (FNR) increases and false positive rate (FPR) decreases for the minority class (females), particularly for missing values under MNAR. On the other hand, FPR increases and FNR decreases for the majority class (males). The differences in the error rate between the majority class and the minority class are larger for missing values under MAR and MNAR, especially for MNAR. This example highlights the importance of understanding the generative process of the missing values when trying to understand the impact of missing values on the fairness of the machine learning classifiers.

Selection bias To illustrate the effect of selection bias, we construct three synthetic data sets: 1) one without selection bias (Balanced), 2) one with relatively little selection bias (Bias Level 1), and 3) one with more bias (Bias Level 2).[7] All three data sets contain 5000 observations. We use each of them to train a logistic regression model. We then evaluate each model on three test sets containing 1000 observations that have the same three bias levels as the training sets.

Table 4 shows the distribution of the sensitive attribute for the three data sets. In the balanced data set, the number of positive and negative cases for each class value is similar. No subgroup is undersampled or oversampled. In the Bias Level 1 data set, all the groups are slightly oversampled or undersampled, between 3% and 5%. In the Bias Level 2 data set, some groups such as females with a negative outcome are oversampled at a high rate, approximately 15% higher, and other groups such as females with positive outcome are undersampled at a rate that is approximately 15% lower. We can see that this leads to a bias of 0.151 for the Level 1 data set and 0.299 for the Level 2 data set.

Table 5 shows the average value and standard error of classification accuracy, F1 score, and fairness measures for the three data sets. The accuracy and F1 scores across the three data sets are about the same, within a percent of each other. In the balanced data set, the p%-rule is highest and the error rates between male and female classes are the most balanced. In the Bias Level 1 data set (the less biased data set), the fairness measures are a little worse than the balanced set. For example, the p%-rule is 15% lower. In the Bias Level 2 data set (the more biased data set), the fairness measures are substantially worse. Here, the p%-rule is close to 40x lower. All three data sets contain the same number of observations. The only difference is the distribution of the observations across different groups. We see that the different amounts of selection bias lead to similar classification accuracy. However, the impact on fairness differs substantially.

In general, these simple experiments illustrate that selection bias and missing values should not be ignored when analyzing fairness.

6 Experiment setup

In this section, we describe our experimental setup. We begin by presenting the details of our data sets, both real world and synthetic. We then present our evaluation measures for both learning and fairness.

6.1 Data set

COMPAS data set The COMPAS recidivism risk data set [31] contains 7214 observations with 14 continuous and categorical features. For our experiments, the goal is to predict whether the individual recidivated. If an individual is not recidivated, we label that as a positive outcome. If an individual is recidivated, we label that as a negative outcome. This data set contains 3963 observations with the positive outcome and 3251 observations with the negative label. Race is the sensitive attribute, with black as the unprivileged group and non-black as privileged group. Non-sensitive categorical features include age (binned into three age groups, less than 25, 25–45, greater than 45), the number of prior crimes, the COMPAS score, and the severity of the charge.

Adult data set The UCI Adult data set [20] contains 48,842 observations with 15 features. For our experiments, the goal is to predict whether an individual has an income greater than $50,000 per year. This data set is more imbalanced than the COMPAS data since there are 11,687 observations with a positive outcome and 37,155 with a negative outcome. Gender is the sensitive attribute with male as the privileged group and female as the unprivileged group. Non-sensitive categor-

[6] We use 20% here to create an example showing how missing values can affect fairness. This example is to build intuition. Our full empirical analysis is presented in Sect. 7.

[7] We chose these three bias levels as an example to show how selection can affect fairness. A complete empirical analysis is presented in Sect. 7.2.

Table 5 Accuracy, F1 score, and fairness measures using training sets with different level of selection bias (β)

	Balanced set	Bias level 1	Bias level 2
Accuracy	0.768 (0.015)	0.77 (0.015)	0.771 (0.016)
F1 score	0.764 (0.015)	0.762 (0.014)	0.769 (0.018)
p%-rule	0.914 (0.018)	0.778 (0.021)	0.541 (0.029)
FNR for female	0.367 (0.02)	0.423 (0.025)	0.466 (0.028)
FPR for female	0.298 (0.017)	0.269 (0.024)	0.209 (0.016)
FNR for male	0.259 (0.019)	0.239 (0.021)	0.199 (0.023)
FPR for male	0.437 (0.025)	0.468 (0.024)	0.504 (0.029)

ical features include age (binned into decades) and education level. In the adult data set, about 6% of the observations contain missing values across three features. In our experiment, none of our sensitive and non-sensitive categorical features contains missing values.

Synthetic data set Our synthetic data generator has a number of parameters we can adjust including: (1) the number of non-sensitive features, (2) the number of observations (3) the correlation between non-sensitive features and the outcome variable, the sensitive attribute, and the other non-sensitive features, (4) the binning strategy (equal frequency, equal width, etc.), (5) the number of bins, (6) the bias level of the selection bias, (7) the proportion of missing values, and (8) the missing value mechanism (MAR, MNAR, or MCAR). As an example, Table 6 shows different synthetic data sets in which the number of bins and the proportion of missing values are varied, and the other parameters are kept constant. Details about parameters of the synthetic data are provided in Appendix A. Table 6 shows the average p%-rule scores on the synthetic data sets using five fold cross-validation. We only report the average p%-rule score from the five fold CV because the standard error is very small (within 0.02) across all the different settings. We can see that when the number of bins is four, the data fit the best, i.e., they are the least sensitive. The synthetic data we create here contain 20,000 observations with 10,000 positive outcome labels and 10,000 negative outcome labels. (Details about the distribution are provided in Appendix A.) This data set allows us to control the properties of the data more than we can with the real-world data sets.

6.2 Evaluation method

Missing values We use five fold cross-validation to evaluate the performance of our classifiers trained on data sets with missing values. We first generate missing values with different missing value mechanisms and proportions of missing values on the original data sets. We then randomly shuffle the missing values and perform cross-validation. In the first fold, we take the first 20% of the shuffled data set as test set and the rest as training set. We then run the fix algorithm and train the classifiers, again, using five fold cross-validation. We validate the classifiers on the test set to obtain the accuracy, F1-score, and fairness measures.

Selection bias We use a slightly different approach for cross-validation for these experiments. First, we randomly shuffle the original data set. Then, in the first fold, we take the first 20% of data as the test set and the rest (80%) as the training set. In this task, because we want to measure how different bias levels can affect fairness, we need to manipulate the training data to simulate selection bias. To do so, for each bias level, we use the training set as a seed to generate multiple training data sets. With the seed training data, we randomly drop some observations in each group until the desired bias level is reached. We use the same procedure to generate the test data sets of different bias levels. Then, using each training set, we train a classifier and we validate the classifier on all the test sets with varying levels of bias. In the second fold, we take the next 20% as the test set seed and the rest as the training set seed. We then use the seeds to generate training and test data sets and we repeat this for each fold.

7 Experiment

In this section, we begin by investigating the impact of missing values and selection bias on both accuracy and fairness as the levels of missing values and selection bias change. We then apply our fairness fixing algorithm to assess the improvement in terms of fairness and the impact on accuracy and F1-score. We conduct our empirical evaluation using categorical data from the COMPAS recidivism risk data set, the UCI Adult data set, and a synthetic data set as described in Sect. 6.1. We add an additional feature to Calmon et al. that captures the missing values [14]. We then implement our approach for calculating the weighting and sampling schemes, using this adjusted data to train the classifier. We make all of our code available for further research in this area. For all of the experiments, we use the fixed data to train a logistic regression classifier and compute the accuracy, F1-score, and the fairness measures disparate impact ("p%-

Table 6 p%-rule score on synthetic data sets with a varying number of bins and a varying proportion of missing values

Number of bins	$\lambda = 0$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.15$	$\lambda = 0.2$	$\lambda = 0.25$
2	0.496	0.488	0.473	0.487	0.473	0.468
3	0.846	0.838	0.778	0.818	0.804	0.788
4	0.904	0.859	0.814	0.764	0.748	0.717
5	0.842	0.809	0.68	0.599	0.567	0.592
6	0.782	0.698	0.655	0.677	0.703	0.679

rule") and error rate balance using five fold cross-validation (see Sect. 6).

This section is organized as follows. We begin by presenting our empirical evaluation for data containing varying levels of missing values, then data containing different amounts of selection bias, and lastly, data containing both missing values and selection bias.

7.1 Missing values and fairness

For this set of experiments, we analyze the impact of different proportions and mechanisms of missing values in each of our data sets. We consider all three missing value mechanisms: MAR, MCAR, and MNAR. Figure 2 shows accuracy, F1 score, and fairness measures for each data set, where each row shows results for a different data set. Going from left to right, the subfigures show accuracy and F1 score, p%-rule, false negative rate (FNR), and false positive rate (FPR). Unlike Sect. 5 where we set proportion of missing values to 20% ($\lambda = 0.2$) to build the intuition, in this section, we try multiple values for λ. For all subfigures, the x-axis shows the proportion of missing values and the y-axis shows the score being measured. Each line in the figure shows a different missing value mechanism (blue = MAR, red = MCAR, green = MNAR). As we mentioned in Sect. 6, we use five fold CV for evaluation and we obtain one set of measurements for each test set. In this figure, because the standard error is very small across all folds (within 0.01–0.02 for accuracy and F1 score, within 0.02 for p%-rule and error rates), each line represents average value across all measurements in all five folds.

In each data set, the F1 score is lower than the accuracy because each data set is imbalanced. In the COMPAS and synthetic data sets, the number of observations with a positive outcome is higher than the number of observations with a negative outcome (about 10% more). In the adult data set, the distribution is much more imbalanced with the number of observations having a positive outcome occurring three times more often than the number of observations with a negative outcome. Across all three missing value mechanisms, accuracy scores and F1 scores decrease as proportion of missing values increases. In general, the accuracy and F1 score of the classifiers remain fairly consistent across missing value mechanisms when the proportion of missing values is less than 20%. In terms of fairness, for missing values under MCAR the p%-rule and error rates are fairly consistent as the proportion of missing values increases. This is not surprising since with MCAR missing values are uniformly randomly distributed in the data set, and therefore, the missing values do not introduce bias.

With the MNAR missing value mechanism, as the proportion of missing values grows, the i%-rule gets smaller and the error rate differences between the privileged group and unprivileged group get larger. Unlike MCAR, MNAR does introduce bias into the data set. Thus, MNAR has a significantly larger impact on fairness than MCAR, with a drop in p%-rule of 15–32% depending upon the data set. It is also interesting to see the impact of the false negative rate on the underprivileged group—it is higher across all proportions of missing values and the gap widens as the proportion of missing values increases. With the MAR missing value mechanism, the impact on fairness measures is in between that of MCAR and MNAR. In this case, missing values depend on the observed data. Therefore, it is possible that more missing values exist in one sensitive group when compared to the other, meaning that the missing values become proxies for the sensitive attribute. Our results are consistent with previous studies in the statistical community about missing values. For example, Dong et al. suggest that MCAR is less of a threat to statistical inferences than MAR or MNAR, and MNAR is the largest threat [19].

Effectiveness of Reweighting on MAR and MNAR We now evaluate the effectiveness of our fixing algorithms on missing values under MAR and MNAR. Recall, that our fixing algorithm assigns a higher weight to observations without missing values and a lower weight to observations with missing values so that the classifier can learn more information from observations without missing value. Because MCAR has very little impact on fairness, we only evaluate our fixing algorithm on MNAR and MAR missing value mechanisms. Figure 3 shows the accuracy score and fairness measures before and after reweighting on MAR distributed missing values, and Fig. 4 shows the accuracy score and fairness measures before and after reweighting on MNAR distributed

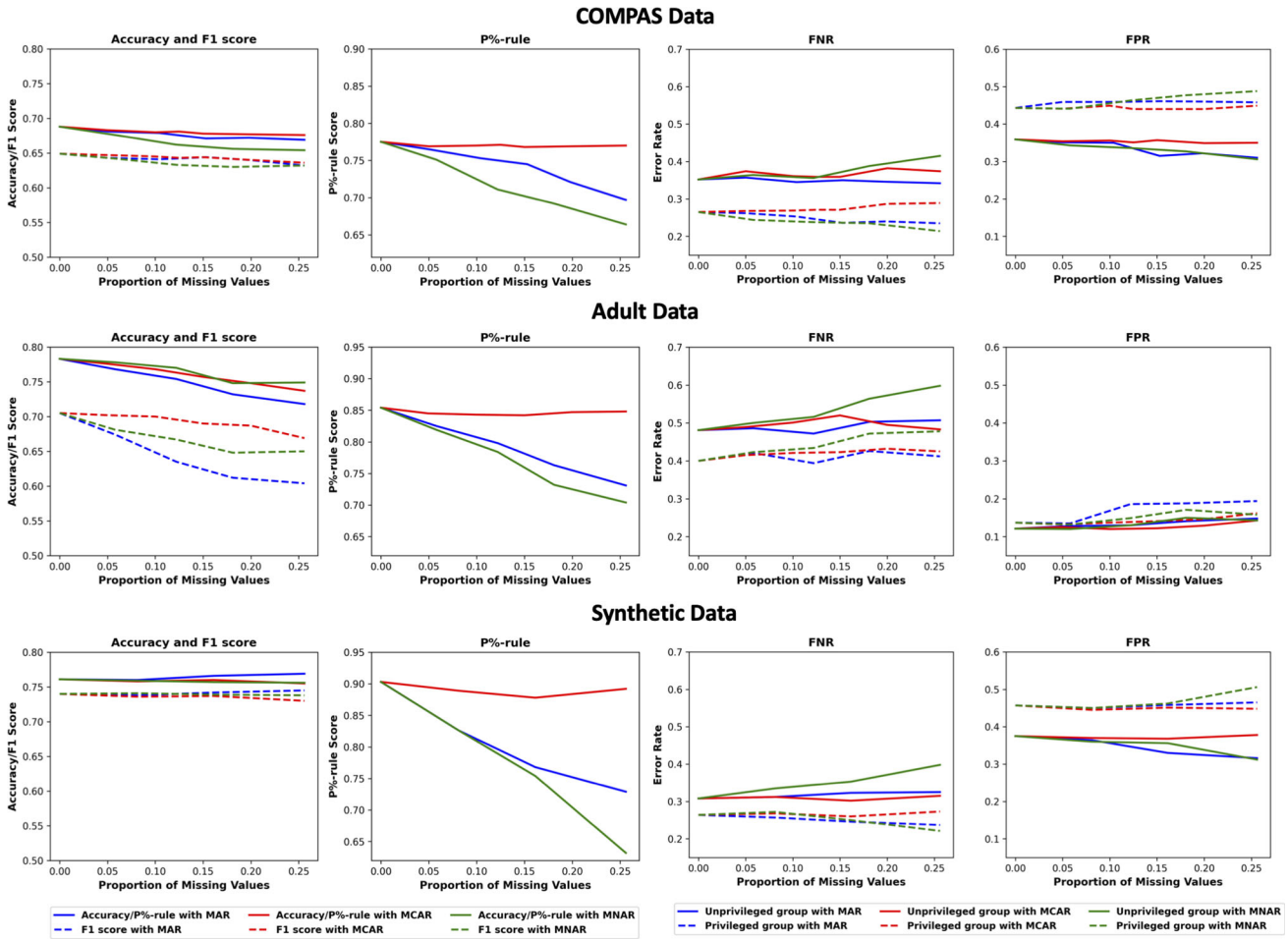

Fig. 2 Accuracy, F1 score, and fairness measures across missing value mechanisms

missing values. In both figures, the subfigures have the same x and y axes as Fig. 2. In both figures, we report the average of measurements from cross-validation because the standard error is very small (within 0.02 for all metrics). The blue line represents the non-fixed results, and the red line represents the fixed results using reweighting.

For MAR, we see that our reweighting algorithm is effective when the proportion of missing values is greater than 5–10%. In those cases, we can see a 5% to 20% improvement in fairness on the real-world data sets and synthetic data with little decrease in accuracy and F1 score (less than a 4% decrease). The impact on fairness is larger for all the data sets if the missing values are generated from the MNAR mechanism. This is not surprising given that it is correcting a larger bias. In all figures, we can see a trade-off between fairness and performance. For example in Fig. 3, in the subfigures showing accuracy and F1 score, the blue lines are higher than the red lines for a trade off in fairness scores. The trade-off is relatively smaller in the COMPAS and synthetic data than the trade off in the adult data. In the COMPAS data, when the proportion of missing values is between 10 and 25%, the p%-rule scores after reweighting are higher than the p%-rule without missing values (about 3% increase). Such improvement comes with a trade-off in performance that both accuracy and F1 score are lower when the proportion of missing values is between 10 and 25% (about 4% decrease in F1 score from 0.65 to 0.61 and about 3% decrease in accuracy from 0.695 to 0.665).

7.2 Selection bias

In the previous section, we mentioned that selection bias happens when some groups in the sample are oversampled or undersampled, i.e., sensitive attributes and outcomes are not sampled independently. In this subsection, we investigate how selection bias impacts fairness and analyze how well our proposed resampling fixing algorithm mitigates the negative effect of selection bias.

In each data set, we use the five fold cross-validation evaluation method described in Sect. 6. Similar to missing values, we report average value of measurements from the cross validation because the standard error is small for each metric

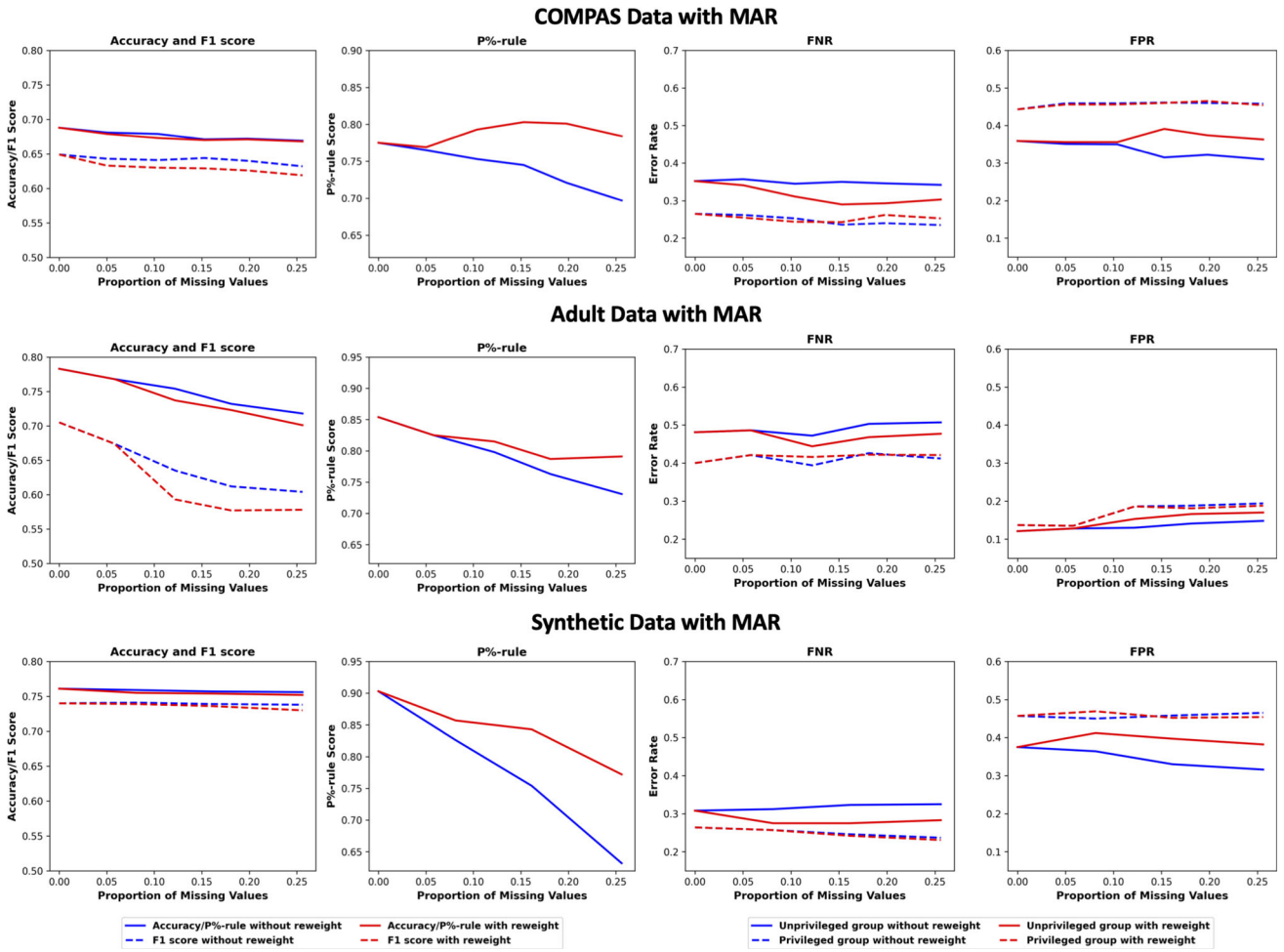

Fig. 3 Accuracy, F1 score, and fairness measures with MAR before and after reweighting

(within 0.02 for accuracy and F1 score, 0.03 for $p\%$-rule and 0.02 for error rates). Figure 5 shows the accuracy, F1 score, and fairness measures for different level of selection bias on the test sets. The x-axis shows the level of selection bias, and the y-axes show the accuracy with F1 score and fairness measures. The blue line represents the results before using our fix algorithm, and the red line represents the results after using our fix algorithm. Across three data sets, the accuracy and F1 scores are fairly consistent with less than a 2% change across different levels of selection bias. On the other hand, the fairness measures, including the $p\%$-rule and the error rate balance, decrease as the amount of selection bias increases. The improvement of our fixing algorithm is over 40% when the selection bias is over 20%. Such improvement can be explained by the balance of the training data. Imbalanced training data can have significant impacts on fairness. For example, Buolamwini and colleagues [11] show that some facial recognition algorithms result in gender and race bias, e.g., darker-skinned females have significantly lower accuracy than others because the data used to train the algorithms contain fewer samples for some demographic groups. Our resampling algorithm can mitigate these types of biases in the training data. This result shows that it is possible to maintain similar F1 scores and accuracy levels as that of the biased data even after the bias has been reduced.

7.3 Selection bias and missing values

In the previous two subsections, we discussed the impact of missing values and selection bias individually. In this subsection, we want to consider both real-world data issues together to understand how fairness is affected when both issues occur simultaneously. We then evaluate the effectiveness of our fixing algorithms for improving fairness in this situation.

For ease of exposition, we focus this evaluation on the COMPAS data set. While we do not present the results from the other data sets, they are comparable to the results we show for the COMPAS data set. We present two different experiments. In the first experiment, the bias level is constant, and we vary the proportion of missing values. In the

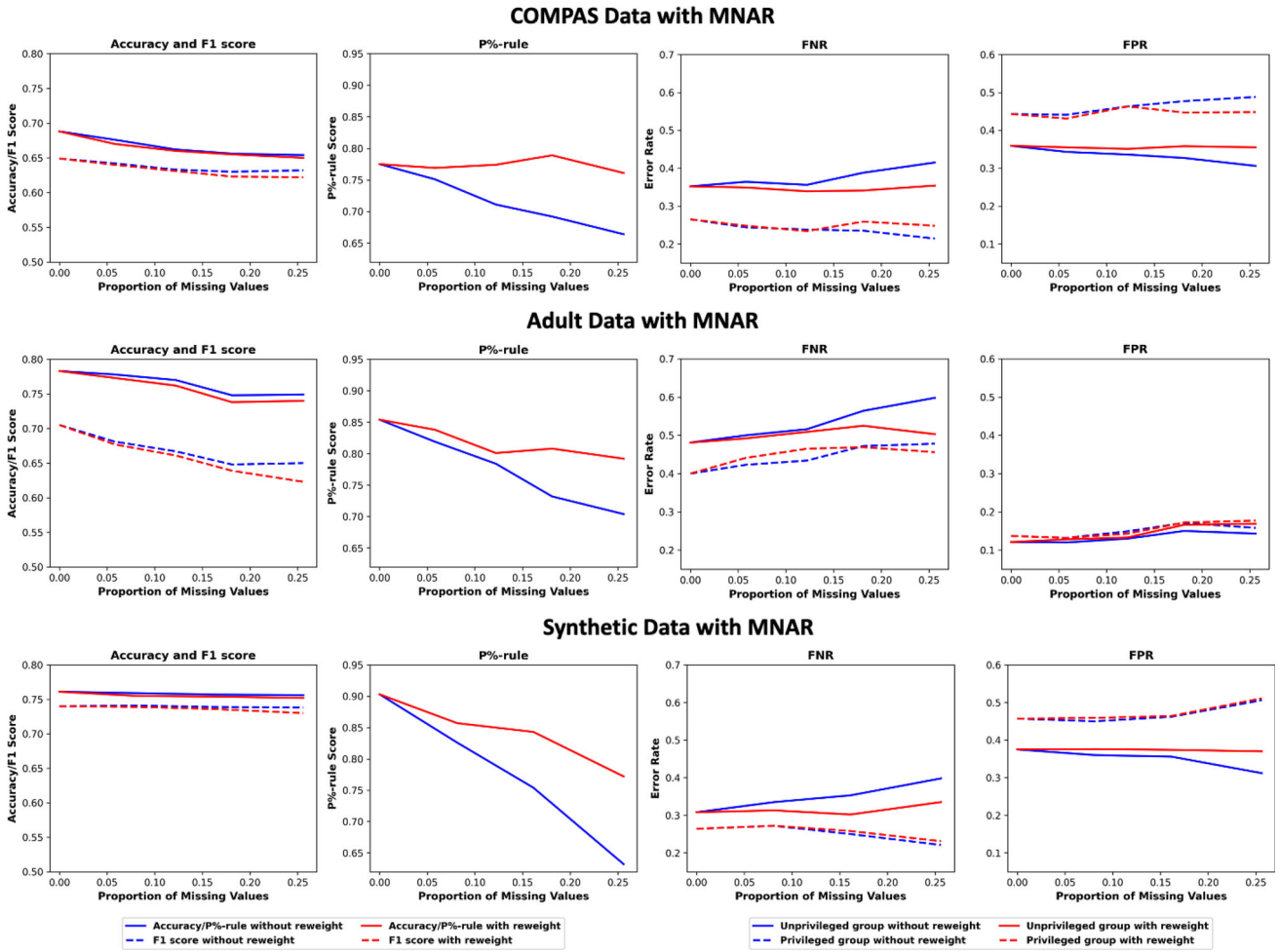

Fig. 4 Accuracy, F1 score, and fairness measures with MNAR before and after reweighting

second experiment, we keep the proportion of missing values constant and vary the level of selection bias. In each experiment, we consider two missing value mechanisms: MAR and MNAR. We choose not to include MCAR because in Fig. 2 we have shown that with MCAR, the fairness measures and performance measures do not change much across various proportion of missing values. Also in theory, missing values under MCAR do not introduce any bias and the only negative effect of MCAR is the reduction in sample size. We expect that with both selection bias and MCAR missing values, the fairness and performance measures will be the same as the case with only selection bias. Our evaluation uses the same cross-validation technique as the selection bias experiments with one difference. At the beginning of the process, before shuffling, we create missing values with different missing value mechanisms and proportions.

Figure 6 shows the accuracy, F1 score, and fairness measures with selection bias and MNAR missing values using COMPAS data. Similar to figures presented in the previous sections, we show the average values. For these experiments, the standard errors are within 0.03 for all metrics. In the top four figures, we keep the selection bias level constant and vary the proportion of missing values. In the bottom four figures, we keep the proportion of missing values constant and vary the level of selection bias. The blue line shows the results when the fixing algorithm is not applied. The red line shows the results when the fixing algorithm is applied. In this figure, we see similar results to those presented when we investigated missing values and selection bias individually in Sects. 7.1 and 7.2 . The figure shows that the proportion of missing values and the selection bias level can both negatively affect fairness. We see that there is an improvement in fairness of over 35% when the fixing algorithm is applied, while the accuracy and F1 score decreases by less than 3%. Once again, we see a large payoff in terms of fairness with very little impact on the performance of the classifier.

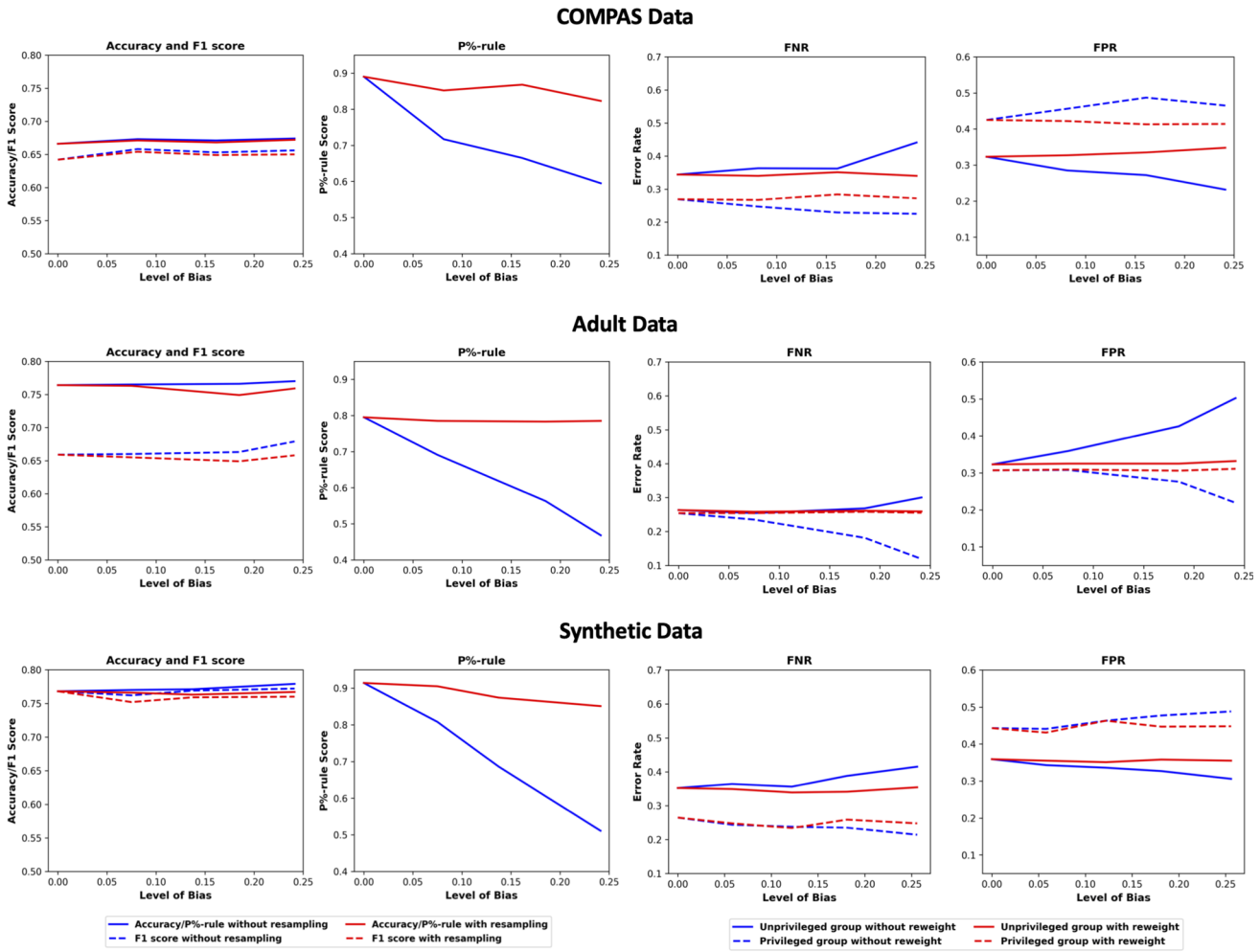

Fig. 5 Accuracy, F1 score, and fairness measures with selection bias before and after resampling

7.4 Effectiveness of uniform resampling and stratified resampling

In Sect 4.3, we compare the different resampling strategies—uniform resampling and stratified resampling when there are missing values. If the missing values are under MNAR, we choose to use uniform resampling to avoid adding more bias from the biased missing values. If the missing values are under MAR, we choose to use stratified resampling where we only resample from observations without missing values to reduce the proportion of missing values.

As we discuss in Sect. 3.3, there is no existing method to determine whether missing values are under MAR or MNAR. If the distribution of missing values is unknown, the best strategy is to treat the mechanism as MNAR and use uniform resampling. In Fig. 6, the red line shows the accuracy and fairness measures with MNAR missing values after applying uniform resampling and reweighting. The accuracy scores decrease a little as the bias increases, but the fairness measures improve significantly, over 35%. In Fig. 7, the red line shows the accuracy and fairness measures when the missing values are distributed as MAR after applying uniform resampling and reweighting. In this case, we do not know the distribution of missing values. The green line shows the accuracy and fairness measures after using stratified resampling. In the figure, the accuracy scores are similar with and without the fixing, but the fairness measures are higher for the green line (stratified resampling) than the red line (uniform resampling) and the blue line (no fixing). If we know the distribution of missing values is MAR, we can use a stratified resampling method to get the best performance from the fixing algorithm. If we do not know the distribution and use the uniform resampling method, the result is not optimal, but it is still much better than the fairness measures without applying any fixing algorithms.

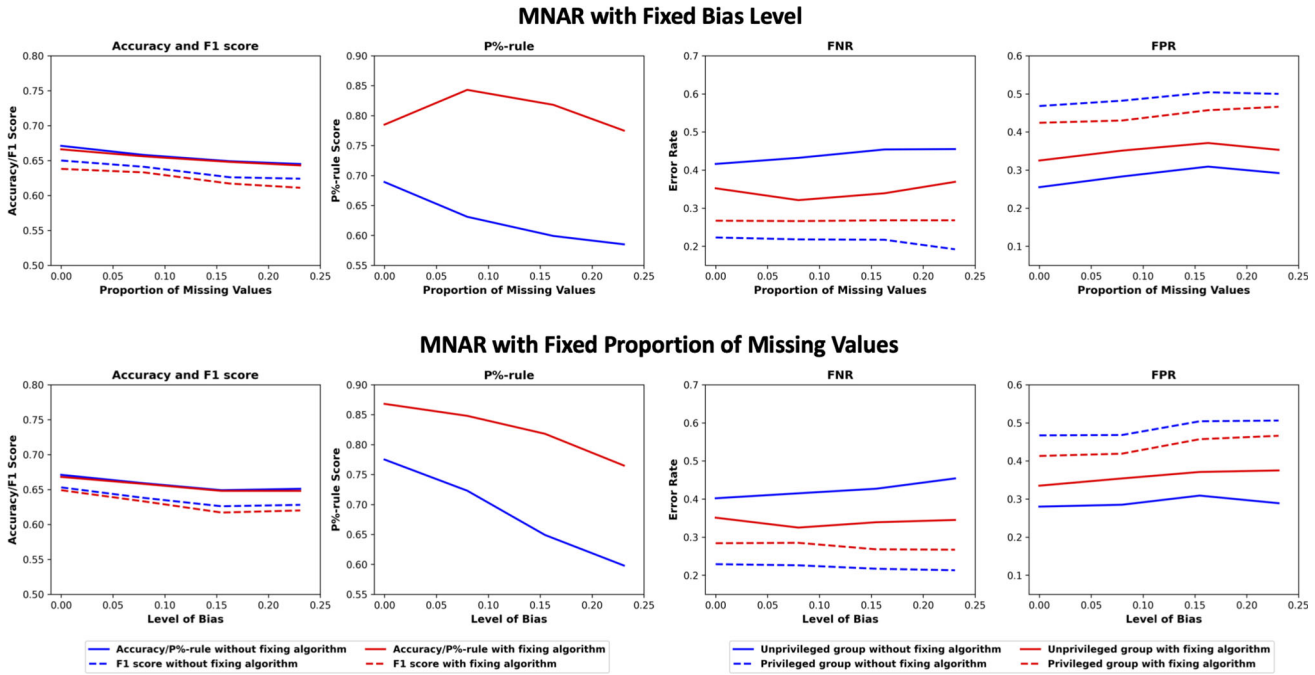

Fig. 6 Accuracy, F1 score, and fairness measures with selection bias and MNAR before and after applying fixing algorithm using COMPAS data

8 Conclusion

This work studies how real-world data issues like missing values and selection bias can negatively affect fairness of categorical sensitive attributes. We begin by expanding the framing of fairness to consider three missing value mechanisms, MAR, MNAR, and MCAR. We then propose fixing algorithms to mitigate the negative effects resulting from missing values and selection bias. We propose a reweighting method for missing values and a resampling method for selection bias. Using two real-world data sets, the COMPAS data and Adult data, as well as a synthetic data set,

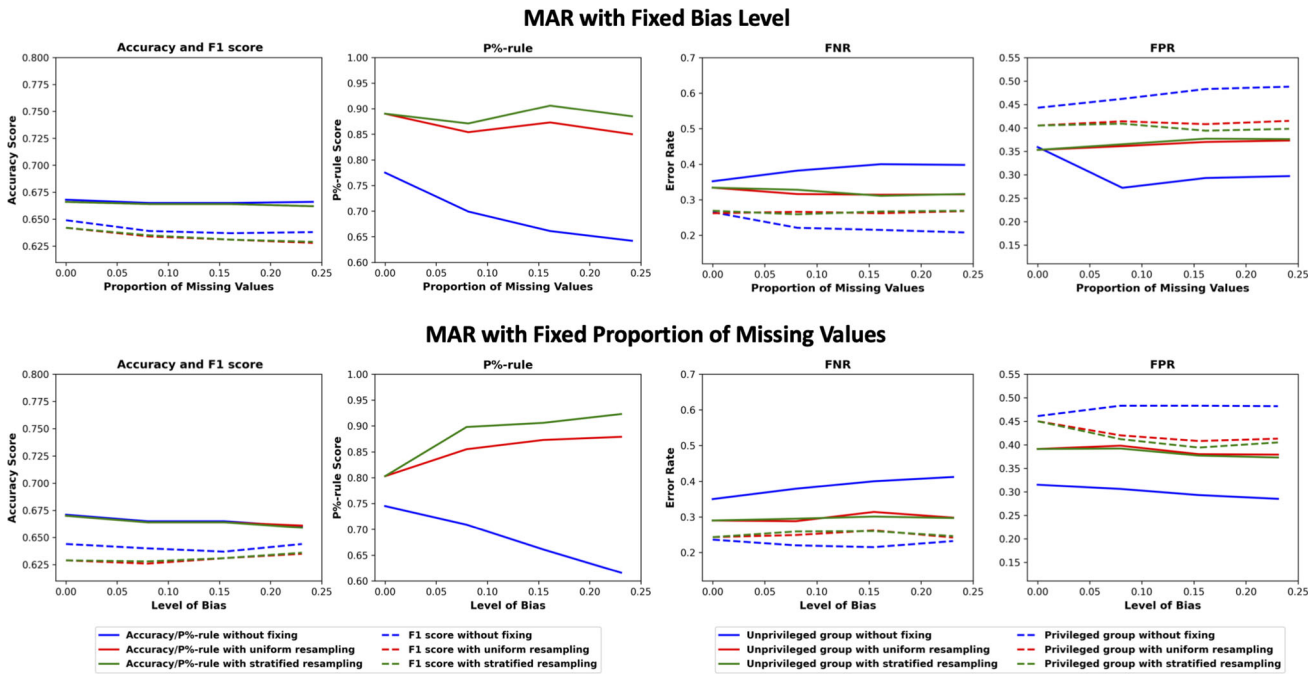

Fig. 7 Accuracy, F1 score, and fairness measures with selection bias and MAR before and after applying fixing algorithm using COMPAS data

we evaluate the impact of missing values and selection bias on accuracy and fairness. For all of our experiments, we vary the level of bias and the missing value mechanism and measure fairness using $p\%$-rule and error rate balance. We find that among all types of missing values, MNAR has biggest impact on fairness and MCAR has the least impact. In other words, all missing value generation mechanisms are not equal with regard to fairness. When considering different levels of selection bias, not surprisingly, our results show that higher levels of bias lead to a larger negative impact on fairness. We evaluate our pre-training fixes for missing values using reweighting, selection bias using resampling and both missing values and selection bias using a combined fixing method that incorporates both reweighting and resampling. We show that our fixing methods are able to significantly mitigate the negative effects on fairness of missing values and selection bias with a small impact on accuracy and F1 score. To support further research in this area, both our code and the synthetic data set have been made available.

While this is an important first step, there are several avenues for future work. First, there are many other real-world data issues other than missing values and selection bias. For example, real-world data can be noisy. Because noise can have various distributions, future work should investigate the impact of these different types of noise on fairness. Second, our fixing algorithms for missing values and selection bias take place during preprocessing. Future work could investigate ways to adjust the learning process to account for these biases. Finally, in additional to continuous and categorical data, there are many other data types including text and image data. It would be meaningful to study different real-world data issues on other commonly used types of data to understand how they impact fairness.

Acknowledgements This research is supported in part by National Science Foundation awards #1934925 and #1934494, the National Collaborative on Gun Violence Research (NCGVR), and the Massive Data Institute (MDI) at Georgetown University. We thank our funders for supporting this research. We would also like to thank Agoritsa Polyzou, Fritz postdoctoral fellow and the rest of the Georgetown Datalab team for their insight.

Funding NSF #1934925, NSF #1934494, National Collaborative on Gun Violence Research.

Declarations

Conflict of interest/Competing interests On behalf of all authors, the corresponding author states that there is no conflict of interest.

Availability of data and material The data and the code are available at https://github.com/GU-DataLab/fairness-and-missing-values

Code availability https://github.com/GU-DataLab/fairness-and-missing-values

A Appendix: Synthetic data distribution

The synthetic data set we used in our empirical evaluation has the following specifications: three non-sensitive variables, 20,000 observations, a constant correlation, equal frequency binning, and a constant bias level. The missing value mechanism is MAR. Figure 8 shows the Pearson correlation score between the outcome, Y, the sensitive attribute, *gender* and the three non-sensitive features. Table 7 shows the distribution of the synthetic data and the bias level β is 0.22.

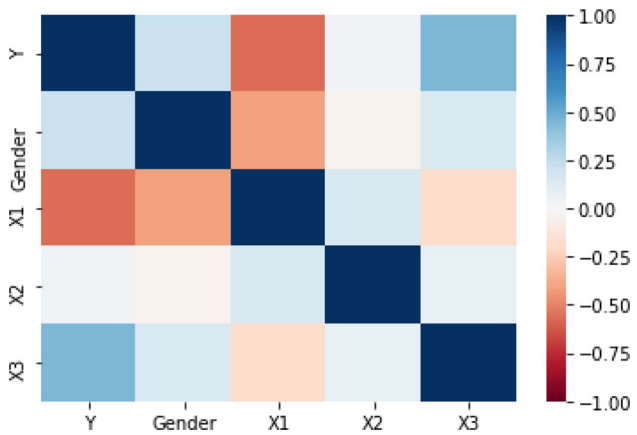

Fig. 8 Pearson correlation score between outcome, sensitive attribute and non-sensitive features

Table 7 Distribution of the synthetic data

Sensitive value	Outcome	Number of observations
Female	Positive	4485
Female	Negative	6684
Male	Positive	5515
Male	Negative	3316
Total		20000
Bias β		0.22

References

1. Acock, A.C.: Working with missing values. J. Marriage Fam. **67**(4), 1012–1028 (2005)
2. Allison, P.D.: Missing Data, vol. 136. Sage Publications (2001)
3. American Bar Association: Disparate-impact claims under the adea (2019). https://www.americanbar.org/groups/gpsolo/publications/gp_solo/2011/september/disparate_impact_claims_adea/
4. Baeza-Yates, R.: Bias on the web. Commun. ACM **61**(6), 54–61 (2018)
5. Bennett, D.A.: How can i deal with missing data in my study? Aust. N. Z. J. Public Health **25**(5), 464–469 (2001)
6. Biddle, D.: Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing. Gower (2005)
7. Binns, R.: Fairness in machine learning: Lessons from political philosophy. In: Conference on Fairness, Accountability and Transparency, New York, NY, USA. pp. 149–159 (2018)
8. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Proceedings of International Conference on Neural Information Processing Systems. pp. 4356–4364. NIPS'16, Curran Associates Inc., USA (2016)
9. Bronner, L.:Why statistics don't capture the full extent of the systemic bias in policing (2020). FiveThirtyEight, June 25, 2020. https://fivethirtyeight.com/features/whystatistics-dontcapture-the-full-extent-of-the-systemic-bias-in-policing
10. Brown,T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dharwal, P., Neelakantan, A.,Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G.,Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C.,Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C.,McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are fewshotlearners (2020). arXiv:2005.14165
11. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR, New York, NY, USA (23–24 Feb 2018)
12. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: IEEE International Conference on Data Mining Workshops, Miami, FL, USA. pp. 13–18 (2009)
13. Calders, T., Verwer, S.: Three naive bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. **21**, 277–292 (2010)
14. Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: Advances in Neural Information Processing Systems 30, pp. 3992–4001. Curran Associates, Inc. (2017)
15. Carpenter, J., Goldstein, H.: Multiple imputation in mlwin. Multilevel Modell. Newslet. **16**(2), 9–18 (2004)
16. Center for Democracy & Technology: Ai & machine learning. https://cdt.org/area-of-focus/privacy-data/ai-machine-learning
17. Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data **5**(2), 153–163 (2017)
18. Chouldechova, A., Roth, A.: The frontiers of fairness in machine learning (2018). arXiv:1810.08810
19. Dong, Y., Peng, J.: Principled missing data methods for researchers. SpringerPlus **2**, 222 (2013)
20. Dua, D., Graff, C.: UCI machine learning repository (2017). http://archive.ics.uci.edu/ml
21. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of Innovations in Theoretical Computer Science Conference, pp. 214–226. Association for Computing Machinery, New York, NY, USA (2012)
22. Farhangfar, A., Kurgan, L., Dy, J.: Impact of imputation of missing values on classification error for discrete data. Pattern Recognit. **41**(12), 3692–3705 (2008)
23. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, pp. 259–268 (2015)
24. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst. **14**(3), 330–347 (1996)
25. Goel, N., Yaghini, M., Faltings, B.: Non-discriminatory machine learning through convex fairness criteria. In: Conference on Artificial Intelligence, New Orleans, LA, USA (2018)
26. Graham, J.W.: Missing data analysis: making it work in the real world. Annu. Rev. Psychol. **60**, 549–576 (2009)
27. Green, B., Chen, Y.: Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In: Conference on Fairness, Accountability, and Transparency. Atlanta, GA, USA (2019)
28. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Proceedings of International Conference on Neural Information Processing Systems, Red Hook, NY, USA, pp. 3323–3331 (2016)
29. Horton, N.J., Kleinman, K.P.: Much ado about nothing. Am. Stat. **61**(1), 79–90 (2007)
30. Jakobsen, J.C., Gluud, C., Wetterslev, J., Winkel, P.: When and how should multiple imputation be used for handling missing data in randomised clinical trials-a practical guide with flowcharts. BMC Med. Res. Methodol. **17**(1), 1–10 (2017)
31. Jeff, L., Surya Mattu, L.K., Angwin, J.: How we analyzed the compas recidivism algorithm (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivismalgorithm
32. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. **33**(1), 1–33 (2012)
33. Kamiran, F., Calders, T., Pechenizkiy, M.: Discrimination aware decision tree learning. In: Proceedings of IEEE International Conference on Data Mining, Washington, DC, USA, pp. 869–874 (2010)
34. Kamishima, T., Akaho, S., Sakuma, J.: Fairness-aware learning through regularization approach. In: IEEE International Conference on Data Mining Workshops, Vancouver, BC, Canada, pp. 643–650 (2011)
35. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Machine Learning and Knowledge Discovery in Databases, Berlin, Heidelberg, pp. 35–50 (2012)
36. Kang, H.: The prevention and handling of the missing data. Korean J. Anesthesiol. **64**(5), 402–406 (2013)
37. Kehl, D.L., Kessler, S.A.: Algorithms in the criminal justice system: Assessing the Use of Risk Assessments in Sentencing. Berkman Klein Center for Internet & Society (2017)
38. King, G., Honaker, J., Joseph, A., Scheve, K.: Analyzing incomplete political science data: an alternative algorithm for multiple imputation. Am. Polit. Sci. Rev. **95**, 49–69 (2001)
39. Kuhn, M., Johnson, K.: Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press, Chapman & Hall/CRC Data Science Series (2019)
40. Little, R., Rubin, D.: Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics, Wiley (2002)
41. Little, R.J.A., Schenker, N.: Missing Data, pp. 39–75. Springer, Boston (1995)
42. Martınez-Plumed, F., Ferri, C., Nieves, D., Hern'andez-Orallo, J.: Fairness and missing values (2019). arXiv:1905.12728

43. Olteanu, A., Castillo, C., Diaz, F., Kıcıman, E.: Social data: biases, methodological pitfalls, and ethical boundaries. Front. Big Data **2**, 13 (2019)
44. Pickles, A.: Missing data, problems and solutions. In: Kempf-Leonard, K. (ed.) Encyclopedia of Social Measurement, pp. 689–694. Elsevier, New York (2005)
45. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. In: Proceedings of International Conference on Neural Information Processing Systems Long Beach, CA, USA, pp. 5684–5693 (2017)
46. Rubin, D.: Multiple Imputation for Nonresponse in Surveys. Wiley, Wiley Classics Library (2004)
47. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
48. Schafer, J.L., Graham, J.W.: Missing data: our view of the state of the art. Psychol. Methods **7**(2), 147–77 (2002)
49. The White House: Artificial intelligence for the american people. https://trumpwhitehouse.archives.gov/ai
50. Van Buuren, S.: Flexible Imputation of Missing Data. CRC Press (2018)
51. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness, pp. 1–7. FairWare '18, Association for Computing Machinery, New York, NY, USA (2018)
52. Wang, T., Wang, D.: Why amazon's ratings might mislead you: the story of herding effects. Big Data **2**, 196–204 (2014)
53. Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness constraints: mechanisms for fair classification. In: Proceedings of International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA. pp. 962–970 (2017)
54. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. Proceedings of International Conference on Machine Learning, Atlanta, GA, USA **28**, 325–333 (2013)