

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Christian Fricke

Missing Fairness:

The Discriminatory Effect of Missing Values in Datasets on Fairness in Machine Learning

Master's Thesis
Espoo, October 5, 2020

Supervisor: Professor Aristides Gionis
Advisor: Assistant Professor Indrė Žliobaitė

Author:	Christian Fricke	
Title:	Missing Fairness: The Discriminatory Effect of Missing Values in Datasets on Fairness in Machine Learning	
Date:	October 5, 2020	Pages: 66
Major:	Machine Learning, Data Science and Artificial Intelligence	Code: SCI3044
Supervisor:	Professor Aristides Gionis	
Advisor:	Assistant Professor Indrė Žliobaitė	
<p>As we enter a new decade, more and more governance in our society is assisted by autonomous decision-making systems, enabled by artificial intelligence and machine learning. Recently, an increasing amount of academic and general-audience publications have made aware of negative side effects accompanying such systems under the umbrella term of algorithmic fairness. While most of the articles focus on a small number of well-studied cases, to the best of our knowledge, none have dealt with large real-world datasets one might use to train models on in an industrial setting.</p> <p>Datasets are collections of observations recorded by humans, including many different forms of biases. Many proposed solutions to combat the structural discrimination focus on the detection and mitigation of unfairness in datasets and machine learning models. The readily available implementations and services adhere to the common practice of complete-case analysis by filtering samples containing missing values. This often leads to ignoring large portions of recorded data, further increasing subgroup imbalances and biases.</p> <p>In this thesis, we analyze a sparse real-world dataset and the effect of missing values on the predictive power and measurable discrimination of models trained upon it. We start with a brief review of the current literature on the topic of algorithmic fairness, that is, causes of unfairness in form of various biases, as well as the most current fairness definitions and measures. For our dataset, we acquired self-reported law school admissions data based on a popular internet platform in the USA. We explore patterns of missingness in the data and ways of imputing values based on established methods prior to training and tuning our models. Finally, we evaluate the performance of the models with respect to well-established fairness measures and detect a significant decrease of discriminatory biases for the subset of data with missing values.</p>		
Keywords:	Fairness, Missing Values, Data Imputation, Algorithmic Bias	
Language:	English	

Acknowledgments

I wish to thank my supervisor Aristides Gionis for giving me the freedom to explore the research space of fairness in machine learning, and my advisor Indrė Žliobaitė for patiently guiding me throughout this endeavor with invaluable insights, motivation, and faith in my abilities.

Most of all, I wish to thank my family for their encouragement and my partner Helėn for her unwavering love and continuous support.

Espoo, October 5, 2020

Christian Fricke

Contents

1	Introduction	5
1.1	Background and Motivation	5
1.2	Research Question and Scope	6
1.3	Structure of the Thesis	6
2	Literature Review	8
2.1	Algorithmic Fairness	8
2.2	Missingness and Fairness	14
2.3	Other Research	16
3	Data and Missingness	18
3.1	Origin and Markup	18
3.2	Sensitive Attributes	22
3.3	Missingness Patterns	23
4	Methodology	28
4.1	Fairness Measures	29
4.2	Encoding Strategies	33
4.3	Imputation	34
4.4	Classification	37
5	Experimentation	45
5.1	Dataset Fairness	45
5.2	Model Fairness	47
6	Conclusion	54
	Bibliography	56
A	Appendix	63
A.1	Complete Raw Dataset Attribute List	63
A.2	Protected Group Acceptance Statistics	65
A.3	Optimized Hyper-Parameters	66

1 Introduction

1.1 Background and Motivation

Algorithmic fairness has become an inescapable topic not only in the research community but also in the mainstream news media. Popular topics range from seemingly trivial gender-biased translations [1] to existential issues such as discriminatory recidivism prediction in the US court system [2]. The field of fairness, accountability, and transparency in machine learning has spawned dedicated scientific conferences (FAT/ML¹ and ACM FAccT²). Several general-audience books, published by prominent practitioners and researchers [3–5], allude to the unseen threats of rampantly deployed assisted and automated decision-making systems across societies. These systems determine the outcome of loan applications, instruct preventive policing, inform parole and bail decisions, screen job applications, and govern many other aspects of our lives. Enabled by the rise of digital data collection, the extensive use of data-driven decision-making has been shown to discriminate against groups of people [6, 7]. The negative act of discrimination is to treat people differently based on their membership to some group rather than their individual merit [8]. Even an innocuous and objective algorithm design can yield discriminating predictions based on the already biased historical datasets.

Model fairness thus depends on the quality of the data and how they are processed. Missing values are a natural part of almost every dataset due to systematic errors, intentional omission, removal, or other interventions. The most common treatment of missingness in datasets is the elimination of partial observations or entire (incomplete) attributes. Martínez-Plumed et al. [9] present the first comprehensive analysis of the relationship between missing data on fairness. Their study of three scientific datasets indicates that the subsets containing missing data are fairer when compared to their complete counterparts. An examination of common practical programming

¹ <https://www.fatml.org>

² <https://facctconference.org/organization.html>

implementations of mitigation and fairness evaluation methods reveals a strong preference for eliminating partial observations. This practice worsens the quality of statistical analyses performed on the data, extending beyond just the predictive capability of the models as all parts of the modern data pipeline are affected.

1.2 Research Question and Scope

The aim of this thesis is to locate and analyze a novel dataset relating to the current fairness discussion, and to verify the main findings of Martínez-Plumed et al. [9] — records containing missing values are usually fairer than the rest.

This study contains a brief review of algorithmic fairness, including fairness definitions and measures, and its relationship to missingness in datasets. We pay close attention to missingness patterns and how missing values are treated using various methods, such as imputation. While the main objective is not necessarily to provide the most accurate classifier possible, we deem it imperative to investigate optimal model performance on differently treated subsets of data. Furthermore, we separately analyze the unfairness contribution of each chosen predictor along with its information gain. Our main research question relates to the effect of missing values on a particular fairness measure:

How does the application of imputation methods on a sparse dataset effect the discriminatory bias of machine learning models?

1.3 Structure of the Thesis

Following this introduction, we review the current academic literature on algorithmic fairness in Chapter 2. We list an overview of the most common statistical and individual fairness definitions and briefly discuss the proven impossibility of satisfying both simultaneously. Chapter 3 is dedicated to the analysis of a large and sparse dataset obtained from the MyLSN website [10]. We carefully assess the schema, determine sensitive attributes and potential features as predictors, as well as inspect missing values and their correlations with the protected classes and binary target class. In Chapter 4 we define the theoretical foundations of our experiments and detail the methodological choices as they pertain to the selection of fairness measures, data encoding

strategies, imputation methods, and classification models. Empirical results are presented in Chapter 5, which focuses on the evaluation of numerically measured discrimination for different subsets of data. Finally, we conclude our work in Chapter 6 and propose ideas for further research taking advantage of the provided dataset.

The source code for this thesis alongside the raw and pre-processed datasets are made publically available on GitLab³.

³ <https://gitlab.com/cfricke/missing-fairness>

2 Literature Review

In this chapter, we highlight some of the most recent research on fairness in the scientific machine learning literature. This overview attempts to cover the origins of the fairness debate and the causes of unfairness, as well as quantitative and qualitative definitions and measures of fairness. We present a brief overview of the most common measures, and further explore the influence of missingness in datasets and the current efforts to understand its effects on fairness in machine learning.

2.1 Algorithmic Fairness

The field of algorithmic fairness may be relatively young, but the question of what is quantitatively fair and unfair has been posed several decades ago. Hutchinson and Mitchell [11] trace back the origins of the current fairness definitions to the testing communities of education and hiring in the 60s and 70s. They point out similarities of that period to the research today, such as formal notions of fairness based on subpopulations [12], the incompatibility of certain fairness criteria [13], and the limitations of quantitative definitions [6, 12, 14]. Test items are compared to features and item responses are equated with their values. Framed as a racial issue, standardized tests were blamed to maintain and justify the denial of economic opportunities to African Americans and other minorities. What followed these accusations was the search for a legal characterization of unfairness, or rather the unfair discrimination in university admittance procedures and employment practices.

2.1.1 Causes of Unfairness

Chouldechova and Roth [15] identify common causes for seemingly unfair treatment exhibited by machine learning models. These range from human biases encoded in the data we train them with to the need to explore, that

is, to perform non-optimal actions in order to fill gaps in the knowledge that data provides. Another example of unfairness relates to the equal treatment of unequal subpopulations. When a majority and a minority population show different distributions in the data, training a group-blind classifier to minimize the overall error, but it is not able to fit both optimally, it will fit the majority one. Consequently, the minority population will suffer a larger amount of errors. This is perhaps best captured by a quote commonly attributed to Aristotle and his notion of “treating like cases alike” [16]:

“There is nothing so unequal as the equal treatment of unequals.”

Sensitive attributes in a dataset identify designations for the majority and minority populations. Even when experts recognize the need to carefully select predictors and avoid the inclusion of such attributes, other features may encode the sensitive categorization and act as proxies [6]. Then there are of course biases caused by missingness in datasets [9], skewing the representation of subgroups even further. Worse yet, there may be various reasons for missing values to occur beyond the general availability of information, as we explore below. Our study shows that it is often difficult to detect and reason about predictors that contain bias and contribute to negative discrimination.

2.1.2 Definitions of Fairness

In their work, Friedler et al. [17] define the solution to a fairness problem as the mapping between construct and decision space. Construct, observed and decision space are metric spaces for which there exists a distance to be measured between individuals, a process to translate indirectly observable features to concrete ones, and a task to find an optimal outcome, respectively. An example as it pertains to this thesis is the expected performance in universities. The success in secondary and post-secondary education may serve as the construct space, while we quantitatively observe GPA and LSAT scores. The admission decision can take the form of a threshold boundary based on these performance indicators.

Friedler et al. continue to define the fairness of a specific task to prescribe desirable outcomes: a mapping from construct to decision space is fair when individuals with close proximity in one are also close in the other. This is similar to the Lipschitz mapping defined by Dwork et al. [12], which inspired the former. They continue with the description of an axiomatic worldview that essentially equates the construct and observed spaces, and as a consequence creates

structural bias due to the noisy non-uniform transformation. Structural bias is the result of unequal treatment of groups — the categorization of individuals based on certain characteristics, such as race, gender, age, or religion. More specifically, we observe structural bias when the mapping between construct space and observed space creates a significant distortion (group skew) between the inner-group differences from one space to the other. In the context of admission decisions, Santelices and Wilson [18] show that SAT verbal questions are not a valid measure of ability for the African American subgroup. Thus, as a feature in the observed space, these scores may be considered a direct result of structural bias. Another worldview suggests an equal distribution of features in the construct space for all groups, where the structural bias present in the observed space is not necessarily related to the differences in the construct space. The choice of features and axioms informs the specific fairness goals set by an authority. An admission office might judge an applicant based on their level of achievement and intelligence purely by means of commonly agreed-upon scores, and accepts systematic group differences in the observed space as explainable inaccuracies. [17]

Informed by these relationships, they differentiate between direct discrimination and non-discrimination, where there exists a significant group skew in the mapping from observed to decision space, and an insignificant group skew from construct to decision space, respectively. The former is sometimes referred to as “fairness”. Friedler et al. describe a set of mechanisms providing fairness guarantees equivalent to the two prevalent families of definitions in the literature. These are as follows:

Group Fairness. Often called statistical definitions of fairness, these operate on a fixed set of protected demographic groups partitioned by attributes such as race, gender, age, religion, and more. A decision mechanism is then to guarantee approximate parity of a specific statistical measure across all groups. These measures are commonly defined in terms of ratios of positive and negative classification rates, as we showcase in Section 2.1.3 below.

Individual Fairness. Here, instead of averaging over groups, constraints bind on pairs of individuals. While individuals with a small distance according to some measure are treated similarly, those that are far apart should be treated differently. The definition of a similarity measure is task-dependent and non-trivial, requiring knowledge of implied relationships between features and labels. [12, 15]

Under the axiomatic worldviews mentioned above, the individual and group fairness mechanisms aim to provide guarantees with respect to fairness and non-discrimination. However, as statistical definitions of fairness merely average members of protected groups, differences in structured subgroups and between individuals fail to be recognized. Unfairness may also go unnoticed for those at the intersection of multiple groups, or groups that have yet to be defined or recognized by the legislative as requiring protection. Dwork et al. [12] and Kearns et al. [19] further show weaknesses of these notions of fairness, while Kleinberg et al. [13] and Friedler et al. [17] prove a fundamental impossibility concerning both, individual and group fairness. According to the latter studies, “it is impossible to simultaneously equalize false positive rates, false negative rates, and positive predictive value across protected groups” [15]. For example, for the protected group race, if more Caucasians are admitted than African Americans given the same proportions of applicants, the system might appear biased towards African Americans, thus violating group fairness. Contrarily, if a male and a female applicant with similar scores and attributes otherwise receive different admission results, the individual fairness is compromised.

Binns [20] argues that while individual and group fairness are not simultaneously achievable mechanisms, they are simply different views sharing the same moral and political concerns. Group fairness as defined in the algorithmic fairness literature is represented by the philosophical corollaries of egalitarianism and anti-discrimination, and individual fairness by the corollary of consistency. They introduce a third principle based on another one of Aristotle’s maxims, coined by Schauer [16] as “individualized” or “particularized” justice. It differs from individual fairness in the sense that the latter still treats individuals on the basis of belonging to some group. This group is bound by the task-relevant metric space of the model and the similarity metric used to calculate the distances between each individual. The outcome is a grouping of clusters, which is considered unfair according to individual justice, because every case must be assessed on its own, irrespective of prior knowledge.

The authors [17, 20] question the legitimacy with which decision-makers make consequential decisions based on algorithms. They challenge their ability to correctly make societal and normative assumptions about discrimination exhibited in a specific context and successfully account for them in a machine learning model. Binns even goes as far as to suggest that some assumptions cannot be reflected in data, which may invalidate the use of an automated decision-making system entirely. However, in cases where the application of machine learning models is feasible, both authors claim that it is important to focus on the

difference between worldviews instead of structuring the assumptions about them in terms of individual or group fairness.

2.1.3 Brief Overview of Fairness Measures

When it comes to specific measures, there exists a mapping from the empirical assumptions described above to individual and group fairness metrics that are ultimately compatible. There is a plethora of articles providing complete overviews of these fairness measures from a variety of different viewpoints [8, 21, 22]. Based on the definitions of discrimination in the legal domains, Barocas and Selbst [6] distinguish between:

Disparate Treatment. Formal (direct) discrimination, where an individual’s membership in a protected group is explicitly used as a predictor in the model, and consequently affects the treatment of the group in an unfavorable manner.

Disparate Impact. Similarly, individuals are treated negatively based on their membership in a protected group, but rather indirectly by a seemingly neutral policy.

In the context of algorithmic fairness and the above legal definitions, models trained on data with sensitive attributes removed from the set of predictors may not lead to direct discrimination, but can still yield unjustified adverse effects on members of protected classes [13]. Informed by these notions, Pessach and Shmueli present an up-to-date synopsis of the common metrics in the literature pertaining to classification tasks. Below we briefly describe the most prominent of these measures.

Disparate impact is an attempt to mathematically represent its legal namesake. It is formally defined as the ratio between positive prediction rates for two sub-populations, privileged and unprivileged:

$$\frac{P^+(D_{S_j \neq a})}{P^+(D_{S_j = a})} \geq 1 - \varepsilon,$$

where P^+ is the probability of positive outcomes, D is a data matrix of observations, S_j represents the sensitive attribute of a protected class, and a is the value for the privileged group; for a detailed taxonomy refer to Section 4.1. A high value implies similar rates and fair treatment across the groups. The

ε determines the legally required proportion of positive predictions, e.g., a large value is less restrictive and allows for more unequal treatment of the groups. [23]

Demographic parity, also referred to as *statistical parity*, utilizes the difference of the positive prediction rates for both groups, instead of their ratio:

$$|P^+(D_{S_j=a}) - P^+(D_{S_j \neq a})| \leq \varepsilon$$

Here, a low value implies similar rates and fair treatment across the groups. Analogous to *disparate impact*, this measure's fairness guarantee depends on the chosen value for ε by decision-makers and the legislative. When we remove ε from the equation, we derive the raw measure called *statistical parity difference*. This metric is useful when comparing various models with each other. [12]

As already pointed out by Dwork et al. and proven by others [13, 17], these two measures do not ensure fairness on their own. For instance, an individual in one group may be treated differently than a similar individual in the other group. To require *disparate impact* or *demographic parity* further limits the utility of a learned classification model. When the outcome is correlated with the sensitive attribute, or the predictors act as proxies, an ideal classifier ($\hat{Y} = Y$) would be considered discriminatory. However, contrary to other group fairness measures, *disparate impact* and *demographic parity* can be equally applied to standalone datasets and the predictions of a model.

Equalized Odds is meant to rectify the shortcomings of the previous two measures. A classifier satisfies *equalized odds* if the predictions and the sensitive attribute are independent conditional on the true outcomes. For binary outcomes, predictions and protected groups, it is formally defined as:

$$|P(\hat{y} = c^+ | D_{S_j=a}, y = c) - P(\hat{y} = c^+ | D_{S_j \neq a}, y = c)| \leq \varepsilon, \quad c \in \{c^+, c^-\},$$

where $\hat{y} = c^+$ represents a positive prediction, and $y \in \{c^+, c^-\}$ are true observations of positive and negative outcomes, respectively. This measure computes the absolute difference in the false positive ($y = c^-$) and true positive ($y = c^+$) rates of the privileged and unprivileged groups, bounded by some threshold ε . The smaller ε , the fairer is the learned classification model. Under the two constraints put forward by this definition an ideal classifier will not necessarily be considered discriminatory, as long as the true observations are unbiased. [24]

Equal opportunity, similar to *equalized odds*, measures the difference in the true positive rates of the two groups, but forgoes the false positive rates altogether. As a relaxation of the former constraint, the focus of non-discrimination on advantageous outcomes is formally described by the equation

$$|P(\hat{y} = c^+ | D_{S_j=a}, y = c^+) - P(\hat{y} = c^+ | D_{S_j \neq a}, y = c^+)| \leq \varepsilon.$$

That is, a binary classifier satisfies *equal opportunity* with respect to the sensitive attribute and true outcomes if the difference in the true positive rates of the privileged and unprivileged groups are not larger than the threshold ε . [24]

According to Corbett-Davies and Goel [7], this measure fails when the base rates significantly differ between groups. All of the above measures are based on the parity of statistical measures across the protected class to satisfy the notion of group fairness. For the final measure in their overview, Pessach and Shmueli present the most common individual fairness metric by the same name.

Individual fairness, as noted in the previous section, requires individuals with similar characteristics to be treated similarly. It is formally defined by

$$|P(\hat{y}_a = c | x_a, s_a) - P(\hat{y}_b = c | x_b, s_b)| \leq \varepsilon, \quad d(a, b) \approx 0,$$

where a and b are two individuals, c represents a specific outcome, and $d(a, b)$ is some task-dependent distance metric. Individuals with a small distance are treated similarly, while those that are further apart, are treated differently. The potential inclusion of individual attributes other than the sensitive variable is an improvement over group fairness definitions. However, as previously mentioned, the need to define a similarity measure is non-trivial, and additionally requires further assumptions about the relationship between predictors and label [15].

2.2 Missingness and Fairness

Martínez-Plumed et al. [9] address the presence of missing data as one of the major issues in research and practice. Almost every second dataset of the UCI machine learning repositories [25], a popular collection of datasets for data science research, contains missing values. The authors ask whether or not missing data and fairness are related. The accidental absence of information is just as much a concern as the intentional removal of data due to various

reasons, such as privacy and interventions. The different causes of missingness guide the numerous practices of dealing with missing values in datasets. These causes generally follow common patterns [26], such as:

Partial completion (attrition). Also referred to as breakoff, a partial completion occurs in records for sequentially reported attributes or features. This is a phenomenon that is typically associated with longitudinal studies. During the process of collecting data over time, user fatigue, or object-related problems may fail to provide useful data after a certain point in the study. The resulting ordinal dependency of attributes can be used as a basis for treating the missing values.

Missing by design. In this scenario the missingness mechanism is known and intentional, thus statistically treatable in the analysis. We differentiate between *contingency attributes*, when not all characteristics are equally applicable to all objects, and *attribute sampling*, where attributes are recorded randomly for each observation due to efficiency or other reasons.

Item non-response. Some variables are missing for some observations. In user-related studies, we further consider these categories of missingness: *not provided*, intentionally left out for various reasons; *useless*, provided data yields no information due to suitability, legibility or otherwise; *lost*, caused by errors during data processing due to equipment failure, data corruption, or otherwise.

As pointed by Martínez-Plumed et al., these categories originate from the field of survey research focusing on questionnaires and interviews involving human subjects. They are, however, equally applicable to more general data collection practices. We discuss the various missingness patterns in datasets as generated by devices such as *item non-response* in Section 3.3. Most theoretical models and practical machine learning libraries and applications require that missing values are addressed as part of the pre-processing stage. The most prevalent methods that deal with missing values all instruct their removal from the dataset in some fashion. For example, the practice of discarding all observations tainted by missingness is referred to as complete-case analysis. In Section 4.3, we enumerate several techniques to address missing values in incomplete datasets.

In their work, the authors attempt a mapping between the causes of missingness and unfairness. For each of the above patterns, they match a corresponding group of biases as found in the current literature. These range from selection,

measurement, and self-reporting bias to algorithm bias. They make an important realization upon examining the possible combinations of factors responsible for missing values as they occur in the data pipeline: missing data is unlikely to be distributed evenly between different subpopulations. Subject to the treatment of missingness in the dataset, this might lead to unwanted effects on fairness in the resulting model. There is evidence to suggest that “people might intentionally omit information as a natural coping mechanism when there is a belief that a truthful and complete answer might lead to a discriminatory and unfair decision.” [9] Overall, their assessment of the relationship between causes of missingness and fairness is inconclusive. We can loosely connect causes with one another, but ultimately fail to reason about an actor’s true intentions. Under observation, someone might purposefully alter their behavior, or choose to supply only some information to be favorably classified by either human or system.

Irrespective of the mapping, they discover that most applications and libraries that are categorically concerned with analyzing and mitigating unfairness in machine learning models simply remove the rows or columns containing missing values. A fact that is particularly surprising after the inclusion of (even naively) imputed data shows an overwhelming reduction of the discrimination exhibited by datasets and models. For their analysis, Martínez-Plumed et al. choose various well-studied data collections with relatively little missingness. In this thesis, we expand upon their work by incorporating a large and sparse repository of user data one might utilize in an industrial setting, where the majority of the modeling pipeline consists of pre-processing inaccurate, incomplete, or inconsistent data from various sources.

2.3 Other Research

There are many other studies related to the field of fairness in the machine learning literature. We do not address the topic of fair representations [27, 28], a theoretical approach, whereupon transforming a biased dataset containing features that correlate with protected groups a new dataset is produced, in which the protected attributes are statistically independent of other features. Closely related to fair representations is the research concerned with fair adversarial learning [29, 30], with a focus on achieving group fairness by equalizing type I/II errors (see Section 5.2 for definition). There is evidence that limits the fairness potential of representation learning based on the efficacy

of the transformation, that is, how well transformed features can be separated from protected characteristics [15].

Another important aspect we do not detail further pertains to specific dynamics of fairness [31]. While the implicit goal of algorithmic fairness is to detect and correct discrimination in automated decision systems and assisted decision-making, there have not been many studies on the long-term effect on the target environment. Mouzannar et al. explore basic equality constrained policies, and under which conditions unconstrained ones can reach equality on their own. More importantly, they study the change of designated features in the construct space over time when applying these policies.

Kusner et al. [32] present in their work a novel framework to model fairness utilizing causal inference. With *counterfactual fairness* measure they deem a decision fair if an outcome is consistent for an individual belonging to a specific protected group and their membership altered to a different demographic. The change of the sensitive attribute and affected variables represent a counterfactual worldview. A label independent of descendants of the sensitive attribute in a causal graph satisfies this model’s fairness constraint. This measure is limited by the additional structural assumptions required to build the graph, the analysis of which may actually support fairness enhancing policies and deepen the understanding of the underlying discriminatory biases. Loftus et al. [33] provide a detailed analysis of recent approaches to causality-based fairness. Their mathematical formalization of fairness within causal frameworks “provides tools for making the assumptions that underlie intuitive implicit notions of fairness explicit.”

3 Data and Missingness

We begin with a schematic description of the data and identify key attributes before carefully selecting features to build our classification models with. Further, we analyze patterns of missingness in the data and how the missing values in the predictors correlate with the sensitive attributes and our class label.

3.1 Origin and Markup

In the literature of algorithmic fairness exists a small collection of well-studied datasets, which are used to assess and measure the bias and discrimination of machine learning models. Pessach and Shmueli [21] review the most commonly used ones in a recent study, many of which contain only one or two sensitive attributes and range from a few hundred up to hundreds of thousands of records. The popular census datasets rely on sampling frames to count the population. This practice leads to smoothed values for better extrapolation, to ensure fair inclusion and representation, and mitigate non-response. Such properties make the data less suitable for the purpose of investigating bias and discrimination in a real-world setting, where data may not necessarily represent all groups and fair sampling is rarely ever achieved [12, 34–38]. Another frequently appropriated genre for studying fairness is university admissions. However, out of privacy concerns and general sensitivity, these datasets are often not publicly available.

In this thesis, we want to focus on large and sparse data, which may be utilized in a modern data science pipelines to train binary classifiers on. The MyLSN dataset, provided by Rankin [10], aims to help the legal community, including law school applicants, students and lawyers, better understand the law school admissions process. It is extracted from an online platform [39] founded in 2003 as a free, publicly accessible database of user-supplied law school applicant information with the intent of helping other applicants judge their chances in

Table 3.1: Pre-processed MyLSN dataset schema description for application- and user-based attributes. The type (1) of data is either categorical (c) or numerical (n). Mandatory (2) and sensitive (3) fields are indicated separately, as well as whether or not an attribute is selected as a predictor (4).

Attribute	1	2	3	4	Description
<i>applications</i>					
app_id	c	✓			unique application identifier
app_cycle_id	c	✓			unique user cycle identifier
school	c	✓		✓	school identifier
status	c	✓			current status of application
dt_complete	c				date of completion
dt_sent	c				date of submission
dt_decision	c				date of decision received
scholarship	n				monetary amount of scholarship
early_decision	c			✓	applied for early decision
fee_waiver	c			✓	received fee waiver
accepted	c				applicant is accepted (label)
attending	c				received fee waiver
<i>users</i>					
user_id	c	✓			unique user identifier
cycle	c	✓		✓	admissions cycle
african_american	c		✓		inferred African American
lgbt	c		✓		inferred LGBT
minority	c		✓		underrepresented minority
sex	c		✓		self-identified sex
gpa	n			✓	LSAC grade point average
lsat	n			✓	law school admission test score
state	c			✓	place of residence
years_out	c			✓	years since undergrad. degree
non_traditional	c			✓	self-identified as non-traditional
international	c			✓	foreign citizenship
teach_for_america	c			✓	inferred TFA status
veteran	c			✓	previously served in the military

the upcoming law school admissions cycle. Mr. Rankin has provided us with additional user data, which is made available upon request.

On the LSN [39] website, a new user creates an account with a minimal amount of mandatory fields, including a username as part of the `user_id` attribute in the dataset. They have the option to enrich their profile with additional information, such as the college name or type of their undergraduate education, the major, GPA and LSAT scores, and class rank. Demographic information includes city and state of residence, a free-form field for racial identification, and binary choices for gender and minority membership. Other free-form fields are available for extracurricular activity and additional information and updates. These fields contain arbitrary input and are used to infer attribute values in the MyLSN dataset.

The user then proceeds with opening one or several applications per school for the current admissions cycle, where only the school and the current status are non-optional fields. While “Accepted”, “Rejected”, “Waitlisted”, “Waitlisted, Accepted”, “Waitlisted, Rejected” and “Pending” are considered valid applications in the dataset, the initial status is set to “Intend to Apply” as a default. There are separate date options available for the completion, submission, reception, and decision of an application. Separately, the user might indicate the amount of scholarship received, whether they applied for an early decision, or they received a fee waiver for their application. The latter requires a student to commit to attending the school if they get accepted at an earlier deadline, which is often paired with scholarship benefits.

Table 3.1 contains the schema of the final pre-processed dataset used in the empirical study. A complete list of all original fields provided in the dataset can be found in Appendix A.1. Most of the fields are self-explanatory, however, `african_american`, `lgbt` and `teach_for_america` are inferred based on special markers in the user profile as mentioned above, e.g., “african”, “aa”, or “black”, “gay”, “bisexual”, or “lgbt”, and “teach for america”, or “tfa”, respectively. Since not all of the websites’ fields are made available in the dataset, we use the inferred fields as is. Consequently, we discard the free-form `race`, `schooltype` and `major` fields, as the analysis and classification of contextual markers is beyond this study’s objectives. In their current form, each holds several thousand of varying values, providing little to no information gain for the classification task.

Originally part of the application dataset, we identified 4 cases where the same user had both positive and negative values inferred for the African American

attribute. The designation was consistent for a given `cycle` and likely caused by mismatching profile information between applications.

Furthermore, the user can freely choose whichever value for the four date attributes, past, present, or future. We discard the `decision_ts` timestamp as it lacks the actual time information in its current form and thus is equivalent to the decision date. In the same instance, we reduce the `complete_ts` to its date value, disregarding inconsistencies.

Using the application status, we create our target variable `accepted`, where all occurrences of “Ac” and “AcWa” are positive, and “Re” and “ReWa” are negative. The user dataset contains partially redundant information and requires further data hygiene concerning the non-available indicators for several attributes, e.g., 0, “nan” and “none” for `gpa`. We dismiss the user table LSAT attributes, as the first is equal to the reference in the applications data and `lsat2/lsat3` are only present in less than 2% of the applications for which users retook the exam.

Finally, we merge the applications and user records, ignoring users without any applications, and applications without users, as well as applications that have neither been accepted nor rejected according to our target label. Records are further filtered for applications from users who applied to the same school multiple times over multiple cycles. We choose to keep only the most recent application from each user per school. Note that repeated applications tend to have a higher acceptance rate.

As shown in Table 3.1, most of the data are categorical, with only `school` and `status` being mandatory — excluding the unique identifiers for users and applications. We identify 4 sensitive, protected attributes, namely `african_american`, `lgbt`, `minority` and `sex`. These are automatically disqualified as potential modeling attributes and are utilized for the discrimination analysis below. Additionally, we dismiss `scholarship` as a predictor due to its strongly correlated with the positive class of our target label, detailed in the correlation plot below. It is important to recognize the origin and markup of the available binary attributes, such as `minority`, `international`, etc. Inference depends on the occurrence of certain markers in the source data. Other binary attributes are voluntary fields. A positive flag indicates a strong signal from the user, whereas a negative flag is a weak signal and either indicates a true negative response or merely a lack of response. This further complicates the descriptive statistics of various populations based on these attributes when compared to other categorical data, e.g., `sex` and `years_out`.

Table 3.2: Summary statistics of the MyLSN dataset

	accepted	rejected	total
no. of applicants	37,919	24,834	73,200
no. of applications	187,867	83,000	444,882
avg. no. of applications per applicant	~ 5	~ 3	~ 10

Over its lifetime, the database has accrued almost half a million records. Table 3.2 shows an overview of the summary statistics of the data. The total number of applicants is further differentiated based on the outcome of the application and may include users with both, accepted and rejected applications. The combination of all accepted and rejected applications make up the entire dataset utilized in this study. We observe an average of 9.731 applications per user, which is close to the median of 9. Note that for each label the number of applicants compared to the number of applications does not follow the same ratio due to the varying number of accepted vs. rejected applications per user.

3.2 Sensitive Attributes

As mentioned above, we identify various protected attributes. Table 3.3 contains an overview of the distribution of sensitive groups with respect to applicants and applications. We label each subgroup with the lower number of applicants as the unprivileged minority, that is, positive (1) for African American, LGBT and minority, and female (f) for sex, respectively. Among the groups themselves, there exists a large imbalance when compared to their counterparts. Underrepresented minorities may also belong to other groups, such as LGBT, or African American, without having been correctly inferred in the original data. This mixture of protected characteristics makes it difficult to compare any statistics or fairness outcomes for the four groups separately. Therefore, we may only compare each minority group to its majority counterpart.

Table 3.3: Summary of protected attributes. \odot indicates missingness, i.e., the absence of a value. The groups are African American (aa), LGBT, sex, and minority (mnr). Note that the majority of attributes do not provide a missingness indicator to discern between a negative or non-response.

		applications	applicants	acceptance rate
aa	0	259,487	37,731	0.739 ± 0.278
aa	1	11,380	1,506	0.706 ± 0.320
lgbt	0	268,661	38,985	0.738 ± 0.280
lgbt	1	2,206	253	0.734 ± 0.268
sex	m	134,344	18,485	0.720 ± 0.277
sex	f	91,716	13,456	0.744 ± 0.283
sex	\odot	44,807	7,303	0.773 ± 0.277
mnr	0	236,759	34,639	0.744 ± 0.275
mnr	1	34,108	4,600	0.692 ± 0.310

3.3 Missingness Patterns

As discussed in Chapter 2, fairness in machine learning depends on the quality of the data, and the quality of processing of said data. Missing values in data are a considerable problem not just in research but also in the industry, to which the solution is most commonly the deletion of the samples or entire attributes they occur in. Lavrakas [26] differentiate between three main patterns and their causes, namely *partial completion*, *missing by design*, and *item non-response* (see Section 2.2 for details). Van Buuren [40, p. 34] makes a broad distinction between two types of missing data, namely *intentional* and *unintentional* missing data. He further classifies their differences with the subcategories of unit and item non-response.

Most of the datasets used in the scientific fairness literature contain only small amounts of missing values and have been thoroughly analyzed. As outlined by Martínez-Plumed et al. [9], various patterns may apply depending on the data and the collection method used. In comparison, the MyLSN dataset is extremely sparse for fields that support true missing values and contains a large number of weak signals, where we cannot ascertain a user’s true intentions. The latter poses a problem specifically with respect to class imbalance. Some of the recorded data we discriminate as belonging to a negative class may actually belong to a positive one. This is an example of item non-response, where the answer is not known, or the question either ignored or overlooked. Inference

for African American and LGBT attributes fall into the same category. The user may have failed to indicate their affiliation with one of these groups in their profile data, leading to gaps in the inference.

According to Little and Rubin [41], missingness mechanisms, while not always known to the analyst, are crucial because the properties of missing data methods depend very strongly on the nature of the dependencies in these mechanisms. They differentiate between three general types of missingness mechanisms.

MCAR (missing completely at random). Missing values are independent of known (observed) and unknown (unobserved) parameters and occur at random. Individual rows are assumed to be independent and identically distributed. This implies that the causes of missingness are unrelated to the data.

MAR (missing at random). Missing values depend on observed data only, and not on unobserved components. MAR is a less restrictive and more realistic assumption than MCAR and allows for pure likelihood and Bayesian inferences without modeling the missingness mechanism.

MNAR (missing not at random). Contrary to MAR, some missing values depend on unobserved data, requiring explicit models for a complete analysis. This may be the case if neither MCAR nor MAR holds.

This distinction is important when evaluating the efficacy of missing data methods. As is often pointed out [40–42], Rubin’s theory lays down the conditions under which missing data methods can provide valid statistical inference. While some simple methods provide easy solutions under the MCAR assumption, they may yield biased estimates when the data are MAR or MNAR.

For our dataset, we choose a popular R implementation¹ of Little’s MCAR global test [43] that simultaneously evaluates mean differences on every variable in the dataset. Enders [42, pp. 17–21] notes, testing whether all variables are consistent with MCAR is not entirely useful, because some missingness is likely to be systematic. He highlights several issues with Little’s multivariate extension of the t -test approach: (1) it does not identify the specific variables that violate MCAR, (2) the test assumes that the missing data patterns share a common covariance matrix, and (3) suffers from low power; when the number of variables that violate MCAR is small, the relationship between data and missingness is weak, or the data are MNAR. This can lead to a false sense of

¹ <https://cran.r-project.org/web/packages/BaylorEdPsych/>

3 Data and Missingness

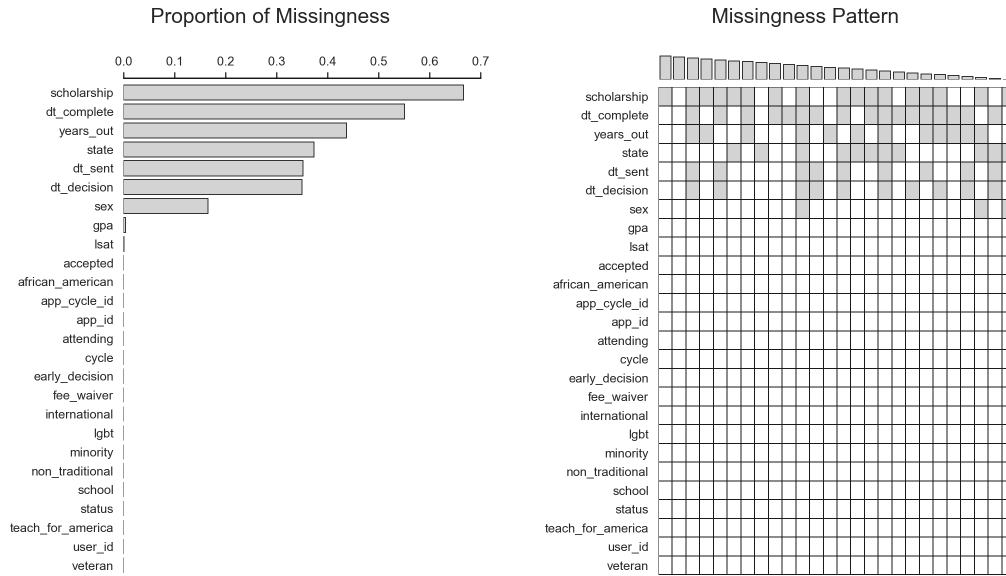


Figure 3.1: Histogram (%) of missing values per attribute (left) and an aggregation plot for the most frequent missingness pattern (right). Of the 252 different patterns in total, we visualize only the cases representing more than 1% of records within the entire dataset.

security about the missing data mechanism, he continues. Despite these issues, and given the missingness pattern exhibited by the MyLSN data (illustrated below), we employ the test and reject the null hypothesis of MCAR with p -values < 0.001 . Consequently, removing or altering the records with missing values may negatively bias the resulting sample.

Figure 3.1 depicts the sparsity of each attribute alongside an aggregation plot of the existing combinations of missing and non-missing values for which the total number of records is greater than 1%. The observed pattern falls into the category of general missingness, more specifically item non-response, which is typically handled by imputation methods [41, p. 10]. We postulate a connection between missing values and the protected attributes, whereas the introduced bias from listwise or column deletion may have a significant effect on fairness. Given the above pattern and confirmed assumptions for collections like the LSN dataset, imputing missing values with simple and iterative imputation methods should generally decrease potentially present negative bias.

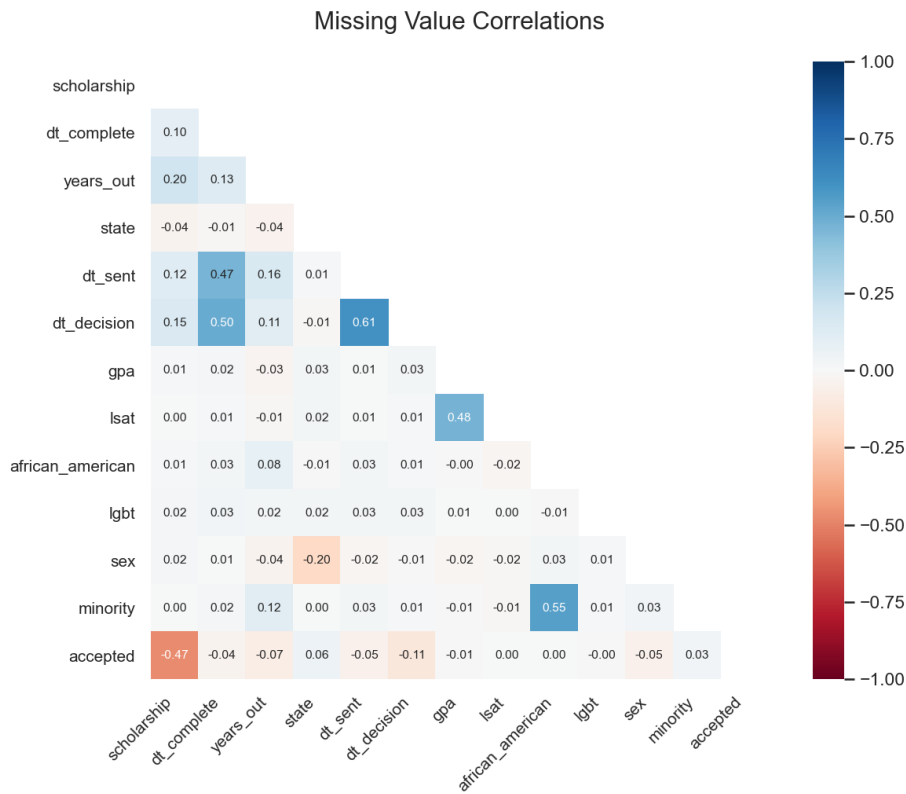


Figure 3.2: Pearson’s correlations of the MyLSN dataset. We include discretized attributes that exhibit missingness (1 if missing, 0 otherwise), the protected attributes (1 if privileged, 0 otherwise), and the binary target class (1 if accepted, 0 if rejected).

3.3.1 Correlations

Finally, we conclude the data analysis by visualizing the Pearson’s correlations between missingness, the privileged protected groups, and the positive label, as depicted in Figure 3.2. Two attributes with missing values that completely overlap have a correlation value of 1, whereas -1 indicates a complete absence of any overlap. In this visualization, we choose to hide the diagonal for clarity, as an attribute’s indicator fully correlates with itself. The matrix confirms a general lack of significant correlations between individual attributes, except for a few cases, such as dates. This is partly due to the heavy class imbalance for the majority of the protected variables.

The three timestamps seem to follow a natural order, whereas a missing date of application completion (`dt_complete`) is followed by the absence of its transmission (`dt_sent`) and lastly the date indicating the reception of a decision (`dt_decision`) from the school. A lack of providing any of the above dates also indicates further absent values for provided scholarships and years since graduation. The latter includes a value for “in undergrad”, making the field relevant for every applicant and not just the non-traditional ones. The failure to indicate a potential scholarship confirms another visible relationship in the matrix, namely the negative correlations between a successful (accepted) application and the dates. An applicant is more likely to provide complete information if an application is accepted.

Other observations include the relationship of supplying a numeric value for the LSAT and the GPA, as well as the evident inclusion of an African American affiliation in the minority subgroup.

4 Methodology

Our empirical study is conducted in multiple steps. We start by defining our choice of fairness measure used to evaluate the bias exhibited in the raw input data, as well as the trained models. This allows for a consistent comparison of negative discrimination before and after machine learning methods are applied. In addition to the global fairness indicator, we further investigate potential conditional discrimination in our dataset. This analysis allows us to separate explainable from purely negative discrimination by observing that part of the differences in the probability of acceptance for the various subgroups may be objectively explainable by other attributes in the data [44].

Next, we explore various data encoding methods to ensure an optimal outcome for imputation and model training. These procedures depend heavily on the markup of the variables and are not necessarily compatible with mixed form data, while also posing severe limitations for the choice of our classification model. In Chapter 5 we provide results for the best encoding, only.

We then proceed with the simple and iterative imputation of the pre-processed MyLSN dataset. While one might critique most fairness-oriented machine learning libraries for deleting records with missing values, there are several limitations when it comes to the practical application of imputation methods. Even simple or naive imputation can pose technical obstacles when dealing with certain data types. More involved procedures that impute missing values by means of predictive models can amplify already present negative bias in the data if the variables containing missing data are related to the target class or the protected attributes. For our study, we compare the effect of both categories of algorithms on the measurable performance of the trained classifiers and the negative bias with respect to all subgroups.

Finally, we discuss our choice of classification method and performance measures, before performing several rounds of hyper-parameter optimization and model evaluation. The tuning of our models allows for the discovery of an optimal balance between model accuracy and negative discrimination mitigation. In the previous chapter, we have already highlighted the predictive attributes used in

our models. We utilize as many predictive features as possible since our model choice essentially performs a feature selection within the training step.

As noted in Chapter 1, for reproducibility purposes we make our source code publicly available online on GitLab¹, a web-based DevOps tool that provides free git-repository management, among other offerings. The entirety of our data analysis, classification experiments, and model evaluation is composed using literate programming in org-mode². Our implementation consists of source code written in the Python programming language, utilizing the popular PyData collection of libraries. The most notable of these are scikit-learn, a machine learning project originally developed by Pedregosa et al. [45], and pandas, a fast and efficient data analysis and manipulation tool [46]. A complete list of libraries can be found in the accompanying Nix³ expression, which enables the distribution of highly reproducible environments for Unix systems, such as Linux and macOS. The documentation and implementation of the above open-source software can be consulted for technical details of the procedures and models referred to in this and the following chapter.

4.1 Fairness Measures

As we already discussed the many notions of fairness and its various measures in Chapter 2, we start by precisely describing the definition of fairness and its measure we employ in this work. To make the study more easily comparable, the taxonomy is aligned with the reference article by Martínez-Plumed et al. [9] and other closely related publications.

Let $D = (X, Y)$ be an $n \times p$ data matrix, where X is a set of categorical and numerical attributes, and Y is the classification label or class attribute. The elements of D are denoted by d_{ij} with $i = 1, \dots, n$ samples and $j = 1, \dots, p$ variables in total. For each sensitive, or protected attribute $S_j \in S \subset X$ exists a set of predominantly binary values V_j , e.g., $V_{\text{minority}} \in \{0, 1\}$, where 1 indicates the membership of a minority and 0 indicates the exclusion from it. In most cases, one of these values is considered the *privileged* group, which is commonly used as a reference to measure the extent of bias, or negative discrimination towards the *unprivileged* group, or groups as the remaining designations in V_j . Similarly, we differentiate between the perceived positive (that is desirable, or

¹ <https://gitlab.com/cfricke/missing-fairness>

² <https://orgmode.org/>

³ <https://nixos.org/>

favorable) and negative (undesirable, or unfavorable) outcome for the values $C \in \{c^+, c^-\}$ of the label Y . Note that in multiclass problems c^+ usually marks a single favorable outcome, while c^- includes all others. In the college admission setting $c^+ = 1$ denotes the positive outcome of an accepted application. Let \odot represent potentially missing values in an unlabeled sample x , which itself is a tuple of values in V_j for each attribute X_j . Together with a class value y from C , $\langle x, y \rangle$ forms a labeled instance. We denote with \hat{y} the expected outcome or prediction with respect to some ground truth y in a decision problem, defined as a mapping $x \rightarrow \hat{y}$. Examples x are sampled from a dataset D , using the notation $x \sim D$. We further define $D_{X_j=a}$ as the set of instances x such that $X_j = a$, with $a \in V_j$. Equivalently, the labeled datasets of all positive and negative samples are denoted by $D_{y=c^+}$ and $D_{y=c^-}$, respectively. Let \hat{D} be the dataset with estimated labels. Finally, we define the probability of a positive outcome $P(y = c^+)$ when $x \sim D$ by $P^+(D)$.

4.1.1 Statistical Parity Difference

The majority of fairness metrics are defined in terms of comparing predicted and true labels, as highlighted in Chapter 2. Given the task of analyzing the effect of missing values in datasets on fairness in machine learning, our choice of fairness measures is restricted to those that can be equally applied to datasets and the predictions of a model.

The two most popular fairness metrics that satisfy this limitation are *disparate impact* and *demographic parity*, also known as *statistical parity difference* (SPD) [12, 21]. While the former was created to satisfy the legal notion of disparate impact, both are concerned with the proportion of positive outcomes for the privileged and non-privileged groups. Formally, statistical parity difference is defined for an attribute S_j with the privileged value a as follows:

$$\text{SPD}_j^+(D) = P^+(D_{S_j=a}) - P^+(D_{S_j \neq a})$$

When measured for \hat{D} , it ensures that positive predictions are assigned to the groups at a similar rate. Hence, a lower value indicates fairer treatment. When $\text{SPD} > 0$, the privileged group has an advantage over the unprivileged one, and vice versa for $\text{SPD} < 0$. If the sensitive attribute S_j is binary and the values of C are swapped for the label Y , then the sign of the SPD flips [9].

$$\begin{aligned}\text{SPD}_j^-(D) &= P^-(D_{S_j=a}) - P^-(D_{S_j \neq a}) \\ &= 1 - P^+(D_{S_j=a}) - (1 - P^+(D_{S_j \neq a})) = -\text{SPD}_j^+\end{aligned}$$

The choice of the privileged group does not change the value of the metric, but rather the sign of the SPD indicates which group receives a larger benefit over the other. One disadvantage of this measure is “that a fully accurate classifier may be considered unfair, when the base rates (i.e., the proportion of actual positive outcomes) of the various groups are significantly different” [21, p. 4]. However, for the purpose of this work, our conclusions do not depend on reaching an optimal value for the metric we choose, but rather the observation of a positive or negative difference between the value for the dataset and our models.

4.1.2 Conditional Discrimination

While statistical parity provides us with the means to measure the general group fairness with respect to positive outcomes, we are also interested in understanding which attributes cause the observed overall negative bias and the extent to which the bias is explainable by other attributes. As we show in Chapter 5, while the overall discrimination exhibited by the dataset may appear negligible, individual features disproportionately contribute to the final value. Even weak correlations of features and sensitive attributes may have a large effect on the discrimination encoded in models trained upon the dataset.

Žliobaitė et al. [44] have previously explored this issue of conditional non-discrimination in classifier design. Their work depends on the assumptions that (1) sensitive and explanatory attributes are nominated externally by law or domain experts, (2) the explanatory attribute is not independent from the sensitive attribute and provides objective information about the class label, and (3) the bad discrimination in the data is directly influenced by the sensitive attribute. Within the scope of this thesis, we may observe multiple explanatory attributes simultaneously while analyzing cases of a single binary sensitive attribute independently from others. Therefore, we separately evaluate the conditional discrimination for each of the four protected variables in our dataset.

The measurable discrimination, here denoted by \mathfrak{D} , is defined in terms of the statistical parity difference as above, in Subsection 4.1.1. Further, Žliobaitė,

Kamiran, and Calders describe the difference in the probabilities $P^+(D_{S_j=a})$ and $P^+(D_{S_j \neq a})$ as a sum of the explainable and bad discrimination as follows:

$$\mathfrak{D}_{\text{all}} = \text{SPD} = \mathfrak{D}_{\text{expl}} + \mathfrak{D}_{\text{bad}} \quad \longrightarrow \quad \mathfrak{D}_{\text{bad}} = \mathfrak{D}_{\text{all}} - \mathfrak{D}_{\text{expl}} \quad (4.1)$$

Coincidentally, they derive the definition for the explainable part of the total discrimination using a toy example about university admissions using gender as the sensitive attribute. That is, the explainable discrimination is the difference between acceptance of the privileged and unprivileged group

$$P^*(y = c^+ | E_i) = \frac{P(y = c^+ | E_i, S_j = a) + P(y = c^+ | E_i, S_j \neq a)}{2},$$

if the acceptance rate for all individuals with a fixed value of the explanatory attribute E_i in both groups is equal, independently of the privileged group:

$$\begin{aligned} \mathfrak{D}_{\text{expl}} &= \sum_{i=k}^k P(E_i | S_j = a) P^*(y = c^+ | E_i) - \sum_{i=k}^k P(E_i | S_j \neq a) P^*(y = c^+ | E_i) \\ &= \sum_{i=k}^k (P(E_i | S_j = a) - P(E_i | S_j \neq a)) P^*(y = c^+ | E_i) \end{aligned}$$

where $E \in \{E_1, \dots, E_k\}$ denotes the set of explanatory attributes. Finally, by substituting the definitions in Equation 4.1, the bad discrimination is calculated as follows:

$$\begin{aligned} \mathfrak{D}_{\text{bad}} &= P^+(D_{S_j=a}) - P^+(D_{S_j \neq a}) \\ &\quad - \sum_{i=k}^k (P(E_i | S_j = a) - P(E_i | S_j \neq a)) P^*(y = c^+ | E_i) \end{aligned} \quad (4.2)$$

We verify the correctness of our implementation based on the above description by measuring the conditional discrimination of the adult dataset [25] and directly comparing the results to the one obtained in [44].

4.2 Encoding Strategies

The majority of machine learning algorithms, as implemented in commonly available programming libraries, operate on complete numerical data. The same is true for scikit-learn, which is implemented on top of the NumPy API, a library adding support for large, multi-dimensional arrays and matrices, specifically optimized for integer and precision numbers. However, often features are given as categorical and not continuous values, requiring additional treatment, or encoding. This issue is typically avoided by performing a complete-case analysis, which is the practice of removing samples with missing values, or even entire attributes, as discussed in prior chapters. Encoding is the process of converting data into a format required for information processing needs. Concerning the classifier used in this research, these needs include a numerical representation of categorical data. This process is further complicated when accounting for missingness, and the imputation of missing values during various steps of the machine learning pipeline.

There are multiple transformations to consider, e.g., numerical encoding, one-hot encoding, or learned embedding.

Integer Encoding. Also known as numerical, or ordinal encoding, a categorical feature's unique values are mapped to a sequence of random, or ordered integers. The dimensionality of the data is not affected.

One-Hot Encoding. One-of-K, dummy, or one-hot encoding transforms a new binary feature for every categorical label. This may result in a very large number of additional features, depending on the cardinality of the feature.

Learned Embedding. Originally developed as a language modeling and feature learning technique in natural language processing, embeddings translate large sparse vectors into a lower-dimensional space that preserves semantic relationships. For categorical encoding, this technique utilizes neural networks to learn a distributed representation of the categories.

For the purpose of this study, we focus on integer and one-hot encoding, only. While some machine learning algorithms may theoretically support arbitrary categorical values, our implementation of the model expects the input data to be in purely numerical form. Several of the categorical features in the MyLSN dataset thus require a transformation, specifically the `cycle`, `school`, `state` and `years_out` attributes.

There are several difficulties to note with respect to missingness and imputation. We start by one-hot encoding the feature vectors, for which the individual cardinality ranges from 4 to 217 unique values. With the entire dataset one-hot encoded, we arrive at 316 input vectors, compared to the original 12 predictors detailed in Table 3.1, Chapter 3. While we are able to encode the complete-case subset of our data without any issues, encoding the incomplete data prior to imputation is not very useful. By applying this method, the features to be encoded are removed from the data, that is, all missingness is lost and samples with missing values receive a zero bit for all newly created binary features. Furthermore, special attention is required when splitting the data into training and test sets. The unique set of values for a given attribute may differ, depending on the split strategy used, resulting in differently shaped input data for the respective phases. Therefore, we need to transform the data after imputing the individual attributes, but prior to the training of our models.

All of the experiments are then repeated with numerically encoded data, for which none of the above issues apply. Every label is replaced with a unique integer as it is encountered in the dataset in sequence, that is, a categorical value at index position 0 is designated the integer 1, the next unique value receives the designation 2, and so forth. A general concern here is the order that integers naturally imply, which disqualifies them for a large number of algorithms that expect continuous input. We choose a popular classification algorithm that treats continuous attribute values as discrete partitions, described in detail in Section 4.4.

4.3 Imputation

It is no surprise that programming libraries remove missing values through listwise deletion by default, or require some form of pre-processing in that regard. Depending on the missingness pattern, several techniques exist to handle incomplete datasets. According to [40], examples of common ad-hoc solutions are the following:

Listwise Deletion. All rows, or samples that contain missing data for selected features are removed from the data prior to training and evaluation. When the dataset is large enough and the missing values are MCAR, samples can be removed without negative consequences. If, however, the MCAR assumption does not hold, this technique will produce a biased model. The predictive power of the estimator generally may be less

accurate the more data are discarded, as is often the case for real-life sparse datasets.

Pairwise Deletion. Also known as available-case analysis, in this scenario we calculate the means and (co)variances on all of the observed data. That is, the mean for each variable is based only on its observed values, while correlation and covariance consider all samples where data are present for the variables under consideration. Similar to listwise deletion, this technique will produce biased estimates if the data are not MCAR.

Indicator Method. The missing values for an attribute are discretized, say, zero for a regression problem and a response indicator is added to differentiate observed from unobserved data. As is the problem with such categorical data, the new feature lacks any order or connotative information and cannot be utilized by most machine learning methods. In this case, the original attribute is commonly removed after its missingness is captured by a binary indicator.

Imputation. Contrary to the techniques above, imputation methods try to infer a meaningful value where it is missing by replacing them with statistical point estimates, or more sophisticated means such as predictive models, incorporating other attributes in the data. Specifically, the goal of imputation is to “draw synthetic observations from the posterior distribution of the missing data, given the observed data and given the process that generated the missing data.” [40, p. 39]

We extend our notation for what follows informed by the author. The response indicator R of our data matrix D is defined as an $n \times p$ 0–1 matrix, where r_{ij} denotes their elements similar to d_{ij} above. We collectively define as D_{obs} the observed data, while its complement D_{mis} , contains all missing elements $d_{ij} \in D$ where $r_{ij} = 0$. With D_j we describe the j^{th} attribute, or column. D_{-j} denotes the complement of D_j , that is, all columns in D except D_j . Like Little and Rubin [41], van Buuren continues to distinguish between various types of missingness patterns:

1. A pattern with a single variable D_j with missing data is called *univariate*, whereas *multivariate* missing data contains multiple such variables.
2. A *monotone* pattern contains variables D_j with missing data that can be ordered such that all variables D_k with $k > j$ are also missing. Non-monotone patterns are also called general patterns.

3. In a *connected* pattern any observed data point can be reached from any other observed data point by sequentially moving horizontally or vertically across the $n \times p$ grid.

Our dataset falls into the multivariate, connected category, as can be seen in Figure 3.2, where each attribute with missingness contains some random overlap with one another. If a pattern is not connected, some variables are not able to provide correlation coefficients required for the imputation of multivariate missing data. The author discusses several other issues, such as some predictors D_{-j} containing missing values themselves, circular dependencies of attributes due to correlations, impossible value combinations that might occur, and the problem of perfect predictions for categorical data. Furthermore, there is the concept of ignorability as it pertains to the missing data assumptions (MCAR, MAR, and MNAR) and the construction of imputation methods, which we will not discuss in this study. [40, p. 38,89,111–112]

Like Martínez-Plumed et al. [9], we posit that large improvements for both, performance and bias, can be achieved even using single imputation, such as *unconditional mean imputation*, as long as the attributes are uncorrelated with the class or the protected group. Formally, we estimate missing values d_{ij} with $r_{ij} = 0$ by the mean \bar{d}_j of the recorded values of D_j , if V_j are numerical, and the mode, otherwise. As both authors [40, 41] point out, the result of an unconditional mean imputation underestimates the true variance of an attribute due to imputing missing values at the center of the distribution. As a consequence, while providing a quick solution to problems with little missing data, this technique should generally be avoided. Regardless of the advice, we apply it to merely exemplify potential benefits over common listwise deletion. However, issues for even simple imputation lie in the identification of numerical vs. categorical features after the data has already been encoded. For ease of implementation, we choose the mode for both, numerical and categorical data.

In order to improve on the previous study [9] further, we additionally employ a fully conditional imputation method, namely, *multivariate imputation by chained equations* (MICE) [47]. Algorithm 1 describes the procedure in detail. Given a dataset of observations and missing values we specify a model for each variable j and fill in starting imputations \dot{D}_j^0 by drawing randomly from the observations. For a fixed number of iterations $t = 1, \dots, M$, we sequentially define the currently complete dataset \dot{D}_{-j}^t . Then, we draw a parameter $\dot{\phi}_j^t$ and fill a new instance of imputed values \dot{D}_j^t from the new conditional

Algorithm 1 MICE algorithm for imputation of multivariate missing data according to van Buuren [40]

Input: Dataset $D = (D_{\text{obs}}, D_{\text{mis}})$, estimator P , number of iterations M

Output: Imputed dataset \dot{D}

```

1: for  $j = 1$  to  $p$  do
2:    $P(D_j^{\text{mis}}|D_j^{\text{obs}}, D_{-j}, R) \leftarrow$  specify imputation model for  $D_j$ 
3:    $\dot{D}_j^0 \leftarrow$  fill by random draws from  $D_j^{\text{obs}}$ 
4: end for
5: for  $t = 1$  to  $M$  do
6:   for  $j = 1$  to  $p$  do
7:      $\dot{D}_{-j}^t = (\dot{D}_1^t, \dots, \dot{D}_{j-1}^t, \dot{D}_{j+1}^{t-1}, \dots, \dot{D}_p^{t-1})$ 
8:      $\dot{\phi}_j^t \sim P(\phi_j^t | D_j^{\text{obs}}, \dot{D}_{-j}^t, R)$ 
9:      $\dot{D}_j^t \sim P(D_j^{\text{mis}} | D_j^{\text{obs}}, \dot{D}_{-j}^t, R, \dot{\phi}_j^t)$ 
10:   end for
11: end for
12: return  $\dot{D}$ 

```

density. Scikit-learn’s `IterativeImputer` is inspired by van Buuren’s MICE implementation in R, that by default produces a single imputation instead of multiple. We can perform multiple imputations by sampling from the Gaussian predictive posterior of the fitted estimator (Bayesian ridge regression) for each imputation.

Finally, like the authors above, we caution that imputation may lead to a false sense of data being complete. In non-ignorable cases, an estimator trained on real and imputed data may exhibit substantial biases [41], especially when performing single imputation.

4.4 Classification

As machine learning models are generally incapable of dealing with missingness, they instead rely on a complete-case analysis of the data by removing all incomplete samples or features. Besides, many require continuous data, misinterpreting even numerically encoded categorical data as ordinal. The algorithmic fairness literature has its own set of requirements for the choice of our model, further complicating the matter. A severe limitation comes in the form of model interpretability, which excludes popular kernel methods such as

the support vector machine and neural networks. These are referred to as “black box” models, which, when tuned correctly, perform extraordinarily well, but are difficult (if not impossible) to interpret [48] and tend to be computationally demanding. Instead, we focus on tree and tree ensemble algorithms. This group of classification methods provides various degrees of explainability, a sound theoretical basis, and a proven track record of strong predictive performance.

4.4.1 Decision Trees

A widely used model able to process missing values is the decision tree algorithm, specifically the classification and regression trees (CART), originally developed by Breiman et al. [49]. As the name implies, and contrary to other supervised learning algorithms, decision trees can be applied to both, regression and classification problems. The criteria for splitting nodes and pruning the tree are the only differences when constructing a tree for either purpose. For our use case, we limit the definition to binary classifiers. Additionally, decision trees choose attributes based on their predictability, conveniently performing feature selection for us.

Formally, a tree T_0 is grown by recursively splitting and pruning N instances $\langle x, y \rangle \sim D$ into subtrees $T \subset T_0$ based on a node impurity measure $Q_m(T)$ and some cost complexity criterion $C_\alpha(T)$. The splitting procedure continues until one of several conditions are met, such as the maximum depth of the tree, or each leaf node containing a minimum amount of samples. A terminal node m represents a region R_m with $N_m = \#\{x_i \in R_m\}$ observations. For a label Y with K outcomes, we define the proportion of class k instances in node m as

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k).$$

Observations in m are classified as $k(m) = \operatorname{argmax}_k \hat{p}_{mk}$, the majority class in the node. The various node impurity measures $Q_m(T)$ include misclassification error, Gini index, and cross-entropy. The former typically guides the pruning process, while either of the latter informs the tree growth. Cross-entropy, for example, is defined as follows:

$$Q_m(T) = - \sum_{k \in Y} \hat{p}_{mk} \log \hat{p}_{mk}$$

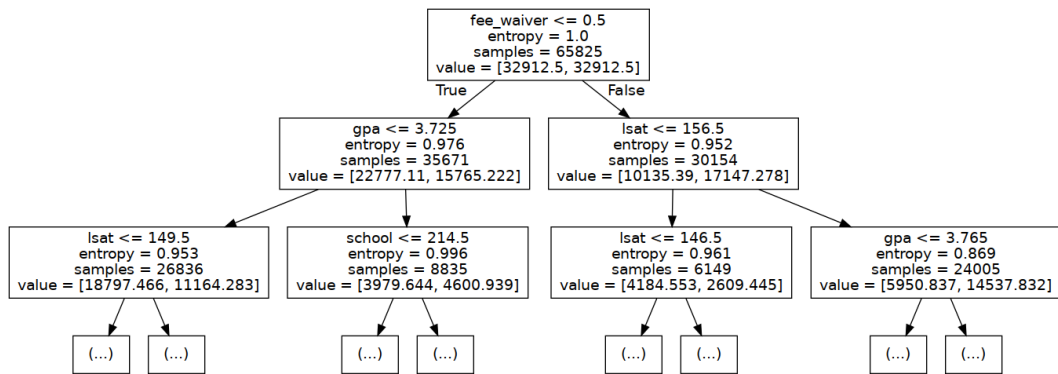


Figure 4.1: First two layers of a sample binary decision tree classifier trained on the integer encoded MyLSN dataset.

For the binary classification problem, where p is the proportion in the second class, the measure becomes $-p \log p - (1 - p) \log(1 - p)$. The tree is pruned by collapsing internal (non-terminal) nodes. With $|T|$ as the total number of terminal nodes in T , the goal is to find a subtree $T_\alpha \subseteq T_0$ that minimizes the cost complexity criterion

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|,$$

for the given tuning parameter $\alpha \geq 0$, which balances the size of the tree and the goodness of fit. When $\alpha = 0$ (the default in scikit-learn) the full tree is chosen. [50, p. 305–310]

Figure 4.1 illustrates the first two layers of an example decision tree trained on our encoded dataset. A node contains the chosen predictor variable and its cutoff, the value for the cross-entropy impurity measure, as well as the number observations and their corresponding split. Each subtree further discriminates the number of accepted ($c^+ = \text{True}$) and rejected ($c^- = \text{False}$) applications based on the predictor providing the largest reduction in impurity. By following the spanning of the if-else rule set, the classification of a new sample becomes a simple task. In a section below we present a metric to measure and visualize the overall importance of each predictor in the tree.

The implementation provided by the scikit-learn library is an optimized version of Breiman’s CART algorithm, but at the time of writing supports neither categorical nor missing values. Unlike Martínez-Plumed et al. [9], who utilize

the R programming language together with the `rpart` package⁴, we are limited to applying our tree-based methods to the subset of complete samples before, and the entire dataset only after, imputing missing values.

Decision trees are sensitive to small changes in the data, tend to easily overfit, and thus exhibit poor out-of-sample performance. While we are not necessarily interested in achieving the maximum prediction accuracy possible, it remains important to demonstrate the effect of mitigating bias and discrimination on modern classification techniques.

4.4.2 Random Forests

Random forests [51] is an ensemble learning method that builds a large collection of decision trees. For classification, the trees form a committee and cast a vote for the predicted class. The algorithm is based on bagging, or bootstrap aggregation, an ensemble meta-algorithm by Breiman for reducing the variance of an estimated prediction function. Its performance is on par with boosting, another ensemble meta-algorithm, but simpler to train and adjust. This makes random forests a popular choice among practitioners. [50]

Algorithm 2 Random forests according to Hastie et al. [50]

Input: Dataset D , number of baggings B , minimum node size n_{\min}

Output: Ensemble of trees $\{T_b\}_1^B$

```

1: for  $b = 1$  to  $B$  do
2:    $\mathbf{Z}^* \leftarrow$  draw bootstrap sample from  $D$ 
3:    $T_b \leftarrow$  initialize random forest tree with  $\mathbf{Z}^*$ 
4:   while  $N_m > n_{\min}$  for any terminal node  $m$  in  $T_b$  do
5:      $\{X_j\}_r \leftarrow$  select a random subset  $r$  from all  $p$  variables
6:      $X_{\text{best}} \leftarrow \operatorname{argmin}_{X_j} Q_m(T_b)$  from all  $\{X_j\}_r$  splits
7:      $(N_{m_L}, N_{m_R}) \leftarrow$  split  $m$  using  $X_{\text{best}}$ 
8:      $T_b \leftarrow T_b \cup (m_L, m_R)$ 
9:   end while
10:   $\{T_b\} \leftarrow \{T_b\} \cup T_b$ 
11: end for
12: return  $\{T_b\}_1^B$ 

```

As detailed in Algorithm 2, the general idea is to grow several decision trees and build an approximately unbiased model with low variance by reducing the

⁴ <https://cran.r-project.org/web/packages/rpart/index.html>

correlation between the estimators. The algorithm starts by randomly drawing a bootstrap sample from the dataset. With this sample, we grow a single tree as described in the previous section. However, instead of the entire features space, we select a random subset of predictors to evaluate the best possible split for each terminal node until the stopping criterion is met. While the documented algorithm performs a majority vote of the classification output from each tree, the scikit-learn implementation averages their probabilistic predictions instead.

For the upcoming chapter it is important to point out the inherent use of out-of-bag (OOB) sampling in the algorithm, which produces an error estimate that “is almost identical to that obtained by N -fold cross-validation.” [50, p. 592]

4.4.3 Cross Validation

Models are typically evaluated using N -fold cross validation (CV), where data is resampled N times into training and test data. While the class attribute in our dataset is not overly imbalanced (69% of accepted applications), the subgroups balance paints a very different picture. The various protected groups are significantly underrepresented, as can be seen in Table 3.3. In combination with the slight class imbalance, the standard 5-fold CV produces an estimator predicting a large number of falsely accepted applications. A better strategy is provided by the *stratified shuffle split*, a variation of the shuffle split, which creates training-test splits by preserving the same class frequency in each fold as exhibited in the complete dataset. It samples randomly over the entire dataset, somewhat alleviating changes in the treatment of subgroup over time.

Furthermore, we diverge from the standard Pareto principle of 80/20 splits and instead make use of an optimal value based on the scaling law for the validation and training set size ratio [52, 53]. The law states that the optimal size of the test set should be inversely proportional to the square root of the number of free adjustable parameters. While it is an artifact of our experiments using one-hot encoded data (with more than 300 additional features), we retain its use as it compares well to the default $1/N$ ratio.

Lastly, there seem to be different strategies for classification under imputation. Our reference [9] imputes the entire dataset prior to training their models⁵. Kuhn and Johnson [54], however, describe imputation in this setting as a

⁵ https://github.com/nandomp/missingFairness/blob/191ee5b6146a5b643cd5fd070f77cf387cf3ea31/MissFairnes_main.R

predictive model within a predictive model. Contrary to the statistical literature, we are not concerned with generating correct hypothesis testing procedures to make valid inferences, but rather with improving the accuracy of our predictions. In the context of classification, we should be estimating missing values of predictors by using the information of other predictors before training the model to avoid data leakage. The additional predictive imputation models add uncertainty, which depends on the various training and test splits one might use to evaluate the parameter configuration of an estimator. We therefore separately impute the missing values as a transformation step prior to training the model on the resampled data and average the results. [54, p. 42]

4.4.4 Hyper-parameter Optimization

For the initial tests, we rely on the parameter documentation of the implementation at hand. Scikit-learn’s selection of sensible defaults tends to follow the guidelines set by the original algorithm authors and current research methodologies. To locate the best value combinations possible for our data and models, we apply a common strategy for hyper-parameter optimization, randomized search [55]. The method is chosen over grid search because it finds models that are as good or better over the same domain by using significantly fewer computational resources. This, in turn, enables the coverage of a much larger configuration space. For most datasets, only a small number of hyper-parameters matter, which also tends to differ from dataset to dataset.

A recent study by Mantovani et al. [56] explores important hyper-parameters for tuning decision trees. For CART-based implementations, the minimum size of terminal nodes (`min_samples_leaf`) and the number of observations required for splitting (`min_sample_split`) stand out above other stopping criteria such as maximum tree depth (`max_depth`). The latter should ideally not be restricted, due to the poor performance of shallow trees when compared to their fully grown counterparts. We also optimize the impurity measure (`criterion`) to determine the best split quality. Additionally, if the dataset is imbalanced for some classes, adjusting weights (`class_weight`) inversely proportional to class frequencies for all observations may yield better results.

The optimization of random forests, similar to decision trees, focuses on largely the same set of hyper-parameters, with the addition of the number of base learners (`n_estimators`) and the number of randomly selected features for evaluating the best split (`max_features`). As Breiman [51] notes, with its natural resilience to overfitting, the former can potentially be very large,

Table 4.1: Confusion matrix for university admissions

actual class	predicted class	
	$\hat{Y} = 1$ (accepted)	$\hat{Y} = 0$ (rejected)
$Y = 1$ (accepted)	true positive (TP)	false negative (FN)
$Y = 0$ (rejected)	false positive (FP)	true negative (TN)

while tuning the latter has a much greater effect on the performance of the model [50, 57]. We perform the optimization with all of the above parameters to avoid potentially weak models due to default values. As for the optimization target, instead of sticking to the most widely used metric in the fairness literature, as detailed in Section 4.4.5, we utilize a metric that more evenly measures the quality of binary classifications. For reproducibility purposes, we point out the constant random seed of 0, which is passed to every algorithm and procedure containing any form of randomness. Given the same data, even the hyper-parameter optimization with random search should yield the same results for every repetition. A complete overview of all optimized hyper-parameters after 100 of 16,200 possible permutations can be found in Appendix A.3.

4.4.5 Model Performance Metrics

In a classification setting the model evaluation measures are commonly based on the confusion matrix described in Table 4.1. The fundamental ones are *recall*, the true positive rate (TPR), *specificity*, the true negative rate (TNR), and *precision*, the positive predictive value (PPV). While the above provides a basic overview of binary classification performance, the F_β score, and Matthews correlation coefficient (MCC) serve as more informative metrics in the literature. The F_β score is defined in terms of *precision* and *recall* as their weighted harmonic mean, where a β parameter smaller or larger than 1 emphasizes either the former or latter, respectively. Even though we are able to fine-tune the scorer, its value can be misleading for imbalanced labels. The MCC, on the other hand, takes into account all positive and negative outcomes and is generally regarded as a balanced measure irrespective of minor or major class imbalances. Its values range from -1 to 1, where the former indicates a complete failure to predict accurately and the latter indicates a perfect classifier. Another well-known metric is *Accuracy*, which works well enough for datasets with balanced classes. For problems with a large class imbalance, it can provide

a distorted insight in favor of the majority class, just like the F_β score. Despite its shortcomings, this measure is the most popular metric in fairness related studies and is thus included in ours as well. Below we list the exact definitions of the metrics used to evaluate the models in this work.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$
$$\textit{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Additionally, to gain a better understanding of key contributors in our decision trees and random forests, we analyze the variable importance plots. These are the split-criterion improvement values attributing to the splitting variable X_{best} , accumulated over all trees separately. The randomized split-variable selection increases the chance that a variable gets included in the random forest. Scikit-learn’s implementation computes the importances as the normalized total reduction of cross-entropy brought by a feature.

Furthermore, we calculate the permutation importances by measuring the decrease in a baseline score for randomly shuffling a single feature. Random forests exhibit low variance due to the bootstrapping procedure during training. As impurity-based variable importances are computed on the statistics derived from the training data, an estimator able to overfit using a certain variable may yield a high importance value, even if the variable is not able to accurately predict the target. Analyzing the permutation importances mitigates some of the limitations of the variable importances and provide us with a better picture of what variable actually holds predictive value.

5 Experimentation

We present the results of the experiments conducted using the methodology laid out in the previous chapter. First, we analyze the inherent bias in the dataset and the progression of discrimination in the trained estimator on imputed data. The predictive model is trained and evaluated on three different subsets of the data to illustrate the difference in fairness with respect to the performance measures. We complete our study with an inspection of the conditional discrimination exhibited in our dataset and how individual biased features contribute to the estimators' performance.

5.1 Dataset Fairness

The results of the initial fairness analysis of the dataset are shown in Table 5.1. We calculate the SPD for different subsets of the data regarding the four protected groups. These include all rows of the imputed dataset, including samples with originally missing values ($X \cup \odot$), rows that only contain imputed values ($X \cap \odot$), and lastly the complete-case subset of the dataset barring any missingness ($X \setminus \odot$).

Missingness, while affecting almost the entire dataset in 9 variables, covers 66% of the samples for 4 of our chosen predictors. Each minority class besides $S_{\text{sex}} = \text{female}$ is extremely underrepresented, as evidenced by majority and minority missing columns in the table. Note that the protected groups do not exclude each other; membership in one does not prevent an individual from belonging to another. For instance, there are 6 users with a total of 48 applications that are part of all four minority classes. 37 (73%) of these samples contain missing values. Interestingly, the entire dataset itself contains only marginal discrimination towards both minority (unprivileged) and majority (privileged) classes, shown as positive and negative numbers, respectively. Here, imputation plays no significant role, yet, as it only pertains to chosen features that contain missing values, none of which are sensitive attributes,

Table 5.1: Statistical parity difference (SPD) for different subsets of data with 66% (178,328 of 270,867) rows containing missing values (\odot) affecting 9 columns, 4 of which are used as features. The subsets include: all rows ($X \cup \odot$), rows containing missing values ($X \cap \odot$) and rows without any missing values ($X \setminus \odot$). The minority missing column shows the number of samples with missingness and their relative percentage for the minority (unprivileged) class. Negative values indicate a discriminatory bias towards the majority (privileged) class. Numbers in bold represent the fairest subset.

protected attribute	majority	minority missing	SPD		
			$X \cup \odot$	$X \cap \odot$	$X \setminus \odot$
african_american	0 (96%)	6,061 (53%)	0.0039	-0.0576	0.0807
lgbt	0 (99%)	982 (45%)	-0.0023	0.0659	-0.0508
minority	0 (87%)	18,831 (55%)	0.0469	0.0181	0.0906
sex	male (59%)	56,172 (61%)	-0.0401	-0.0493	-0.0263

nor the target label. Shifting the focus to the samples containing missing values, removing the complete cases, the SPD values increase and are reversed for the privileged and unprivileged groups of African Americans and LGBT individuals. For the `minority` protected group the bias decreases, whereas males are slightly more discriminated against. In the complete-case subset, the direction of the discrimination is the same as for the entire dataset, but the values are significantly larger, with `sex` being the exception. The latter group exhibits strictly reversed discrimination for all subsets, that is, a bias against the privileged majority group of males. All groups follow the same pattern, where the SPD value for all rows lies in between the two exclusive subsets, averaging it roughly according to their proportions. The fairest value, the SPD closest to zero, differs depending on the attribute, but generally falls in a spread of -6 to 9%. The largest difference is 14% for African Americans. This outcome differs from the findings by Martínez-Plumed et al. [9], where the fairest subset is the one containing only the missing samples for all three datasets. While the missingness in their data follows a similar pattern, the percentage of samples with missing values is significantly lower and showed slightly larger correlations. Additionally, the causes of missingness may also differ, specifically for sensitive attributes. As mentioned in Chapter 3, we do not have complete knowledge about group membership; we can only assume that a positive flag implies a deliberate choice, or in the case of `african_american` and `lgbt` correct inference. The ambiguity and uncertainty present in the MyLSN data mirror real-world datasets used by data scientists in the industry.

5.2 Model Fairness

Table 5.2: Statistical parity difference (SPD), Matthew correlation coefficients (MCC) and *Accuracy* (ACC) averaged for 10 stratified shuffle splits using optimized random forests as the predictive model. Missing values are separately imputed for each training set prior to fitting the model via imputation methods (IM), *unconditional mean imputation* (mean) and *multivariate imputation by chained equations* (mice). All categorical variables are numerically encoded. We include the fairness metric values for the dataset (data) as reference. The notation follows Table 5.1.

IM	SPD			MCC			ACC		
	$X \cup \odot$	$X \cap \odot$	$X \setminus \odot$	$X \cup \odot$	$X \cap \odot$	$X \setminus \odot$	$X \cup \odot$	$X \cap \odot$	$X \setminus \odot$
african_american									
data	0.0039	-0.0576	0.0807						
mean	0.2090	0.1457	0.2473	0.7043	0.6920	0.6653	0.8718	0.8645	0.8597
mice	0.2074	0.1418	0.2473	0.6999	0.6829	0.6653	0.8707	0.8621	0.8597
lgbt									
data	-0.0023	0.0659	-0.0508						
mean	-0.0166	0.0439	-0.0771	0.7043	0.6920	0.6653	0.8718	0.8645	0.8597
mice	-0.0169	0.0477	-0.0771	0.6999	0.6829	0.6653	0.8707	0.8621	0.8597
minority									
data	0.0469	0.0181	0.0906						
mean	0.1728	0.1522	0.1865	0.7043	0.6920	0.6653	0.8718	0.8645	0.8597
mice	0.1709	0.1451	0.1865	0.6999	0.6829	0.6653	0.8707	0.8621	0.8597
sex									
data	-0.0401	-0.0493	-0.0263						
mean	-0.0166	-0.0268	-0.0082	0.7043	0.6920	0.6653	0.8718	0.8645	0.8597
mice	-0.0163	-0.0248	-0.0082	0.6999	0.6829	0.6653	0.8707	0.8621	0.8597

After training the random forests with the parameters set according to the hyper-parameter optimization (Appendix A.3), we evaluate the performance and fairness metrics on the holdout sets of each stratified shuffle split and record the averaged results for all subsets of the data. We separate the training and test sets irrespective of the existence of missing values. Therefore, all splits maintain roughly the same proportion of samples with and without missing values as the whole dataset. Missing values are imputed for each training set in a pre-process transformation step, as noted in Section 4.3. The SPDs, Matthew correlation coefficients and *Accuracy* are captured in Table 5.2 for all sensitive attributes and imputation methods separately. Since none of the protected

variables are part of the set of chosen predictors, the performance measures yield the same results for all of the 4 groups. They differ merely with respect to the imputation method applied to the data. We present the results for the numerically encoded dataset only, as the sequence of transformations is generally more sound when compared to our binary encoding strategy. Random forests benefit from variables with high cardinality when compared to the many binary counterparts, as we highlight further below.

First, we observe a substantial increase in discrimination for the sensitive attributes `african_american` and `minority`. While the strong correlation between these two variables explains the pattern, the extent of the amplification is not as easily explained. As we show in Section 5.2.1, depending on the protected group, there exists a different interplay of predictors with various degrees of negative discrimination. Besides, those with high predictability may also contribute the most harmful bias, as well as introduce a significant amount of missingness. This complex cycle of influence should make the requirement for such an in-depth analysis obvious.

Regarding our choice of fairness measure, we consider the smallest absolute value the fairest. Discrimination goes both ways, as fairness is a zero-sum game—one group’s advantage is another group’s disadvantage. The ideal condition is the wholly equal treatment of all groups, majority and minority alike. The nature of the MyLSN dataset, given the ambiguity of voluntary and partially inferred fields, complicates the determination of a common pattern of discrimination among the various minorities. While the model with the lowest SPD for the aforementioned correlated groups is trained on the subset of samples with missing values, the opposite is true for the other two. In the case of `lgbt`, the negative discrimination in the same subset of data may be explained by the low number of samples available. For the `sex`, the value and missingness distribution are approximately equal between the various subsets. Here we observe a consistently lower SPD when compared to the one measured in the dataset. In contrast to all other groups, the model does not learn to negatively discriminate further but rather equalizes the treatment of majority and minority. This may be partially due to the more balanced subgroups and the already favored minority.

The model performance, on the other hand, depends directly on the number of samples available. Under both imputation methods, the MCC and *Accuracy* are the largest for the dataset including all rows, and smallest for the complete-case subset. The unconditional mean imputation provides consistent inference based on the distribution of each incomplete variable, whereas multivariate

Table 5.3: Normalized confusion matrix for random forests evaluated on mean imputed data ($X \cup \odot$) according to Table 5.2. We include the number of accepted and rejected applications per split for the actual class to visualize the label imbalance. For target class ratios of accepted and rejected applications refer to Table 3.2.

actual class	predicted class	
	$\hat{Y} = 1$	$\hat{Y} = 0$
$Y = 1$ (54,233)	0.894 ($\sim 48,484$)	0.178 ($\sim 9,654$)
$Y = 0$ (23,960)	0.106 ($\sim 2,540$)	0.822 ($\sim 19,695$)

imputation by chained equations produces a separate model for each one. The trees grown in the random forest exhibit even less variance in the former setting and predict slightly better with a lower false positive rate. Consequently, this treatment exacerbates the learned discrimination in the model for those groups with sufficiently many samples containing missing values. As the performance increases, the fairness value worsens.

With Table 5.3 we include a sample confusion matrix to illustrate the type I (false positive) and type II (false negative) error rates. We achieve these values by optimizing our model parameters using Matthews correlation coefficient instead of *Accuracy*. In our optimization tests, the latter scoring function yields significantly more true positives, while the proportionate increase in false negatives has only a minor influence on the overall outcome.

Finally, as we record the worst performance for the models trained on the subset of data with missingness discarded ($X \setminus \odot$), the discrimination is highest for 3 out of the 4 protected groups. These are coincidentally the ones with the least representation in the dataset. The predictions improve for the complement subset of the data ($X \cap \odot$), providing the fairest outcome for the 2 groups with more recorded cases. We observe the mean fairness value out of the two settings above reflected in the results for all rows ($X \cup \odot$) except for the outlier, *lgbt*. Due to the distribution of the data between the subsets, the optimal outcome depends on the target of the model optimization: if the objective is to yield the fairest classifier, we may need to preferentially sample more ambiguous data and rely on the inferences of a second predictive model.

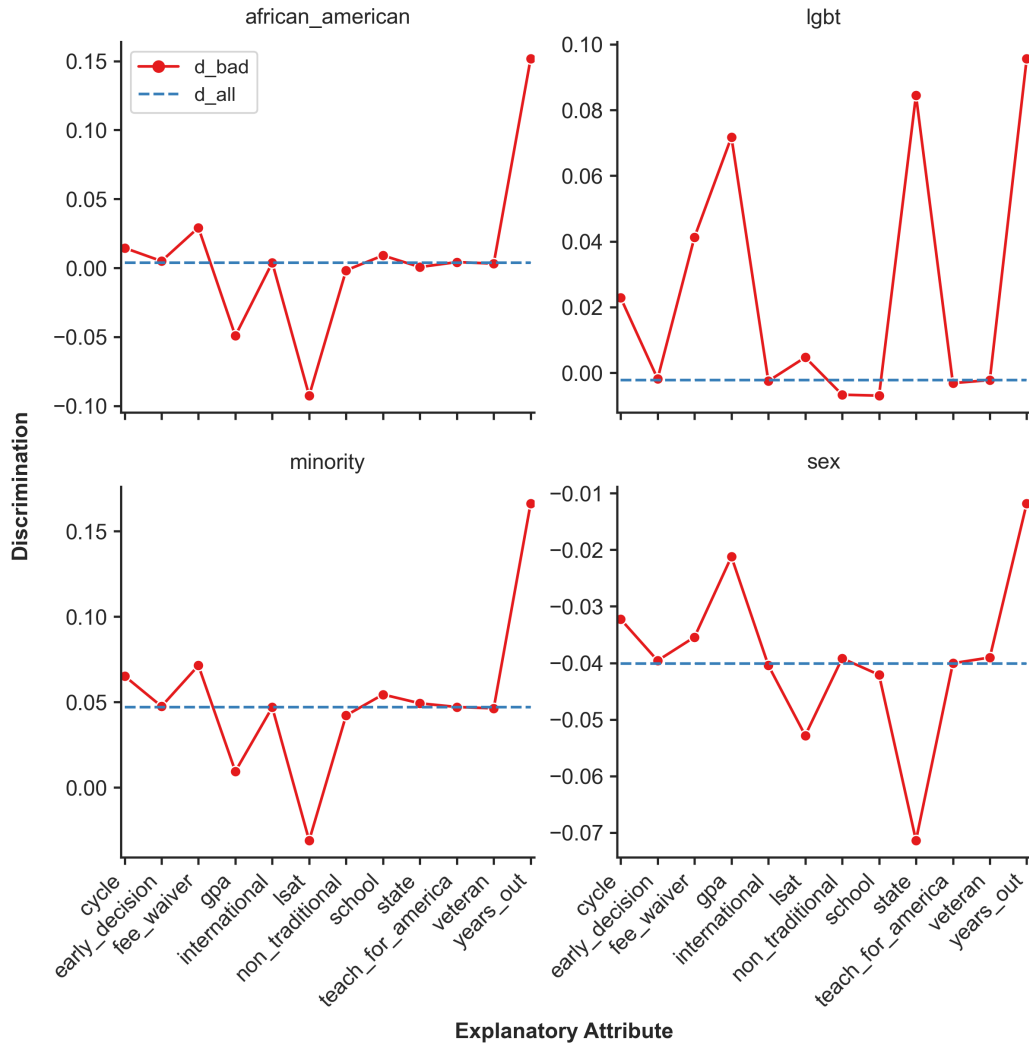


Figure 5.1: Conditional discrimination in the entire dataset for each binary sensitive attribute, with \mathcal{D}_{all} and \mathcal{D}_{bad} denoted by d_{all} and d_{bad} , respectively. Refer to Section 4.1.2 for more details.

5.2.1 Predictor Fairness

Based on the results in Tables 5.1 and 5.2, it may be beneficial to inspect individual predictors for their discriminatory contribution and predictive importance. We start by visualizing the conditional discrimination in the dataset for each binary sensitive attribute in Figure 5.1.

The baseline discrimination ($\mathfrak{D}_{\text{all}}$) is equivalent to the statistical parity difference (SPD). A large difference between $\mathfrak{D}_{\text{all}}$ and the bad discrimination, $\mathfrak{D}_{\text{bad}}$, implies that a group-based difference in admittance depends on the value of the explanatory attribute. The majority of binary predictors contribute little to no bad discrimination themselves—only `fee_waiver` stands out. The opposite is true for the numerical and categorical features with high cardinality. `years_out`, along with `state`, `gpa` and `lsat` explain most of the bad discrimination. These features discriminate the privileged and unprivileged groups in opposing directions, likely nullifying their effect for the entire dataset. It may indicate why the learned models exhibit such an extreme degree of bias, given that these predictors provide the most information about the class label.

The two sensitive attributes `lgbt` and `sex`, for which the designated privileged groups are discriminated against, stand out. From the perspective of demographics, minorities may be more likely to live in, or apply from certain states. `sex` and `state` are also the variables with the largest negative missingness correlation, while the latter’s missing values also correlate the most with being accepted at a university. There seems to be a strong preference for privileged groups depending on the number of gap years. A member of an unprivileged minority is less likely to be accepted based on the values of the `years_out` variable, when compared to the privileged counterpart. These values are missing the most for African Americans and other minorities. In contrast, these groups are favorably treated based on their GPAs and LSAT scores.

Overall, every protected group exhibits a slightly different pattern of conditional discrimination. An exact analysis is hindered by the exclusion of samples with missing values in the calculation of the explanatory discrimination value. A large difference is not necessarily representative for the entire dataset if the variable is missing for the majority of samples. The seemingly arbitrary distribution of $\mathfrak{D}_{\text{bad}}$ values for the LGBT minority with less than 1% of records in the dataset illustrates this phenomenon. Additional tests with these predictors removed show no significant effect on mitigating the bias of the learned estimator, serving as further evidence.

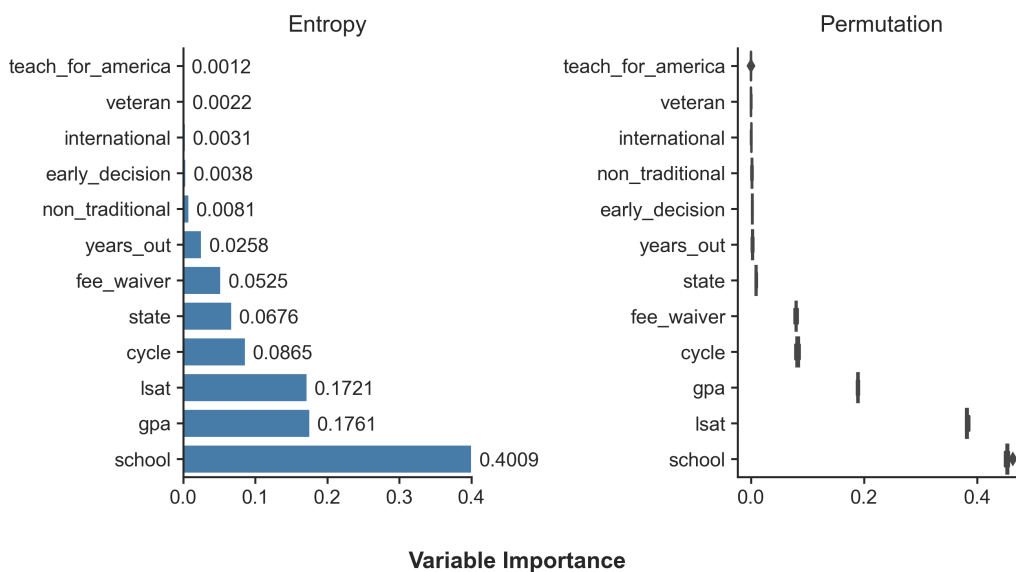


Figure 5.2: Variable importance plots for a random forest grown on a stratified sample of the MyLSN dataset. The mean decrease in impurity (left) is based on the cross-entropy splitting measure, while the permutation importance (right) measures the variable importance with respect to its predictive contribution.

5.2.2 Variable Importances

We assess the importances of variables using the random forest with optimized parameters. It is grown on a single stratified shuffle split of the entire mean imputed dataset ($X \cup \odot$). The permutation importances are computed on the holdout set of the split by comparing the Matthews correlation coefficients. Note that impurity-based importances are biased towards high cardinality features when compared to binary ones.

As shown in Figure 5.2, the best predictors are either numerical or categorical in nature. All binary features, besides `fee_waiver`, contribute comparatively very little to the averaged decrease in cross-entropy across the decision trees in the forest. We observe that the features with a high concentration of missingness such as `years_out` and `state` rank relatively low, as well. Especially in the case of permutation importance, these features seem to add almost nothing to the predictability of the estimator. According to these charts, the choice of `school` has the greatest impact on the admission decision, which might indicate either a strong preference of certain students or merely be a side effect of students with homogeneous value distributions applying to specific schools.

The order of the variables changes from one plot to the next. When measuring the permutation importances, `lsat` is almost equivalent to `school`, the most valuable predictor, while `years_out` and `state` have barely any influence on the coefficients. Finally, the permutation plot reveals a very low variance of the feature contributions between all models. This is likely due to the small number of meaningful variables and the fact that all decision trees within a random forest are similarly constructed using the same set of predictors.

In the previous section, we mention additional experiments, in which we leave out predictors with high values of $\mathfrak{D}_{\text{bad}}$. Continuing this path, we exclude `state` and `years_out`, attributes with a low predictive value and a large percentage of missing values. Training a classifier on the mean imputed dataset ($X \cup \odot$) yields differences in discrimination between -0.0008 and 0.0098. The performance of the model, on the other hand, improves from an MCC of 0.7043 to 0.7075.

6 Conclusion

The purpose of this work was to explore the relationship of missing values and algorithmic fairness, and how different treatments of missingness in a novel, sparse, real-world dataset effect the measured fairness of the original data and the machine learning models built upon it. We accomplished these goals by examining the current fairness literature and expanding on one of the most recent studies on the same subject [9].

In the literature review, we analyzed the definitions of fairness and how they inform decision-making processes supported by algorithms. Many renowned voices in this research area stipulate for a transition away from metric-guided solutions towards assisted decision systems informed by axiomatic worldviews. Our overview of widely used fairness measures focused mainly on statistical metrics as a basis for the comparison of discrimination exhibited in datasets and the predictions of fully trained models. A further investigation into the causes of missingness did not yield any insight concerning effects on fairness but led to the realization that an unequal distribution of missing values would likely influence data and model bias.

Our contributions are twofold. First, we carefully processed a previously unknown collection of user-contributed law school admission data, which we are making available as part of this thesis. We identified four sensitive attributes with 3 underrepresented minority subpopulations that provide the foundation for our fairness research. In an initial statistical analysis, various missingness patterns emerged, spanning multiple attributes we later utilized as predictors in our models. Then, in our experiments, we compared the discrimination exhibited in the three subsets of the data based on the presence and absence of missing values. The group fairness measure, *statistical parity difference*, varies drastically for three out of the four protected classes. The similarity of African Americans and other minorities can be explained by the partial set intersection of both attributes. More surprisingly, once the models had been trained on the subsets of data without missing values and with imputed values, the predictions exposed a drastic increase in discrimination of the latter two

groups. In particular, the models trained on the subset containing only samples with originally missing values appear to somewhat negate the discrimination of the complete-case model. Once the entire dataset is used to grow a random forest classifier, the fairness metric seems to have equalized the two other two subsets' extremes. On the other hand, the model accuracy unsurprisingly improved with each increase in observations.

Furthermore, we explored two different treatments of missing values, *unconditional mean imputation* and *multivariate imputation by chained equations*, and, like Martínez-Plumed et al. [9], were able to affirm our original working hypothesis: the application of naive imputation on a sparse dataset decreases the discriminatory bias of machine learning models by including valuable observations, which would otherwise be discarded in a complete-case analysis. Additionally, with an iterative imputation approach, the discrimination improved even further, while the prediction accuracy of the model worsened. This seemingly obvious trade-off between model fit and group discrimination should not be easily dismissed, as neither model parameters change, nor are any fairness thresholds adjusted. By adding a second predictive model in the form of missing value inference, we introduce a new variable to consider in the design of algorithmic decision systems [15, 20].

Future studies utilizing the dataset we provide along this work could expand on our initial statistical analysis by further exploring the relationship between the predictors and the conditional discrimination. For example, this can be done by performing a hierarchical clustering analysis, sub-sampling records with missing values, or selecting features with high importance and minimal bad discrimination, \mathcal{D}_{bad} . Regarding the treatment of missing values, one might explore other imputation methods or predictive models to infer data more accurately. However, instead of merely shifting the focus to yet another layer of indirection, a structural analysis through the lens of counterfactual fairness may prove more beneficial.

In summary, imputation itself can help to supplement information gain, but more importantly, increases the sample size by incorporating otherwise abandoned observations. The inclusion of partial records reduces negative discrimination towards subpopulations. A major concern of data science practitioners is the negative effect of fairness enhancing measures on the predictive capability of regression or classification algorithms. We showed that even a naive treatment of missing values can significantly improve a classifier's prediction accuracy.

Bibliography

- [1] Schiebinger, L. and Klinge, I. and Sánchez de Madariaga, I. and Paik, H. Y. and Schraudner, M. and Stefanick, M., “Gendered Innovations in Science, Health & Medicine, Engineering and Environment.” 2011–2018. [Online]. Available: <https://genderedinnovations.stanford.edu/case-studies/nlp.html>
- [2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks.” 2016. [Online]. Available: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [3] F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press, 2015.
- [4] C. O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016.
- [5] M. Kearns and A. Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.
- [6] S. Barocas and A. D. Selbst, “Big data’s disparate impact,” *California Law Review*, vol. 104, p. 671, 2016. [Online]. Available: <https://doi.org/10.15779/Z38BG31>
- [7] S. Corbett-Davies and S. Goel, “The measure and mismeasure of fairness: A critical review of fair machine learning,” *CoRR*, vol. abs/1808.00023, 2018. [Online]. Available: <http://arxiv.org/abs/1808.00023>
- [8] I. Žliobaitė, “Measuring discrimination in algorithmic decision making,” *Data Min. Knowl. Discov.*, vol. 31, no. 4, pp. 1060–1089, 2017. [Online]. Available: <https://doi.org/10.1007/s10618-017-0506-1>
- [9] F. Martínez-Plumed, C. Ferri, D. Nieves, and J. Hernández-Orallo, “Fairness and missing values,” *CoRR*, vol. abs/1905.12728, 2019. [Online]. Available: <http://arxiv.org/abs/1905.12728>

Bibliography

- [10] R. Rankin, “Mylsn.info,” 2020, [accessed 2020-01-07]. [Online]. Available: <https://mysn.info/>
- [11] B. Hutchinson and M. Mitchell, “50 years of test (un)fairness: Lessons for machine learning,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. ACM, 2019, pp. 49–58. [Online]. Available: <https://doi.org/10.1145/3287560.3287600>
- [12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, “Fairness through awareness,” in *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, S. Goldwasser, Ed. ACM, 2012, pp. 214–226. [Online]. Available: <https://doi.org/10.1145/2090236.2090255>
- [13] J. M. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” *CoRR*, vol. abs/1609.05807, 2016. [Online]. Available: <http://arxiv.org/abs/1609.05807>
- [14] S. Mitchell, E. Potash, S. Barocas, A. D’Amour, and K. Lum, “Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions,” 2018.
- [15] A. Chouldechova and A. Roth, “The frontiers of fairness in machine learning,” *CoRR*, vol. abs/1810.08810, 2018. [Online]. Available: <http://arxiv.org/abs/1810.08810>
- [16] F. Schauer, “On treating unlike cases alike,” *Constitutional Commentary*, 2018. [Online]. Available: <https://ssrn.com/abstract=3183939>
- [17] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, “On the (im)possibility of fairness,” *CoRR*, vol. abs/1609.07236, 2016. [Online]. Available: <http://arxiv.org/abs/1609.07236>
- [18] M. V. Santelices and M. Wilson, “Unfair treatment? the case of freedle, the sat, and the standardization approach to differential item functioning,” *Harvard Educational Review*, vol. 80, no. 1, pp. 106–134, 2010. [Online]. Available: <https://doi.org/10.17763/haer.80.1.j94675w001329270>
- [19] M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu, “Preventing fairness gerrymandering: Auditing and learning for subgroup fairness,” in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. JMLR Workshop and Conference Proceedings, J. G. Dy and A. Krause,

Bibliography

- Eds., vol. 80. JMLR.org, 2018, pp. 2569–2577. [Online]. Available: <http://proceedings.mlr.press/v80/kearns18a.html>
- [20] R. Binns, “On the apparent conflict between individual and group fairness,” in *FAT* ’20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, M. Hildebrandt, C. Castillo, E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna, Eds. ACM, 2020, pp. 514–524. [Online]. Available: <https://doi.org/10.1145/3351095.3372864>
- [21] D. Pessach and E. Shmueli, “Algorithmic fairness,” *CoRR*, vol. abs/2001.09784, 2020. [Online]. Available: <https://arxiv.org/abs/2001.09784>
- [22] S. Verma and J. Rubin, “Fairness definitions explained,” in *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, Y. Brun, B. Johnson, and A. Meliou, Eds. ACM, 2018, pp. 1–7. [Online]. Available: <https://doi.org/10.1145/3194770.3194776>
- [23] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and removing disparate impact,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu, and G. Williams, Eds. ACM, 2015, pp. 259–268. [Online]. Available: <https://doi.org/10.1145/2783258.2783311>
- [24] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 3315–3323. [Online]. Available: <http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning>
- [25] Dheeru Dua and Casey Graff, “UCI Machine Learning Repository,” 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [26] P. Lavrakas, *Encyclopedia of Survey Research Methods*. SAGE Publications, 2008.
- [27] A. Bower, L. Niss, Y. Sun, and A. Vargo, “Debiasing representations by removing unwanted variation due to protected attributes,” *CoRR*,

- vol. abs/1807.00461, 2018. [Online]. Available: <http://arxiv.org/abs/1807.00461>
- [28] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, ser. JMLR Workshop and Conference Proceedings, vol. 28. JMLR.org, 2013, pp. 325–333. [Online]. Available: <http://proceedings.mlr.press/v28/zemel13.html>
- [29] D. Madras, E. Creager, T. Pitassi, and R. S. Zemel, “Learning adversarially fair and transferable representations,” in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 3381–3390. [Online]. Available: <http://proceedings.mlr.press/v80/madras18a.html>
- [30] B. H. Zhang, B. Lemoine, and M. Mitchell, “Mitigating unwanted biases with adversarial learning,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, J. Furman, G. E. Marchant, H. Price, and F. Rossi, Eds. ACM, 2018, pp. 335–340. [Online]. Available: <https://doi.org/10.1145/3278721.3278779>
- [31] H. Mouzannar, M. I. Ohannessian, and N. Srebro, “From fair decision making to social equality,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. ACM, 2019, pp. 359–368. [Online]. Available: <https://doi.org/10.1145/3287560.3287599>
- [32] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4066–4076. [Online]. Available: <http://papers.nips.cc/paper/6995-counterfactual-fairness>
- [33] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva, “Causal reasoning for algorithmic fairness,” *CoRR*, vol. abs/1805.05859, 2018. [Online]. Available: <http://arxiv.org/abs/1805.05859>

Bibliography

- [34] J. M. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein, “Discrimination in the age of algorithms,” *CoRR*, vol. abs/1902.03731, 2019. [Online]. Available: <http://arxiv.org/abs/1902.03731>
- [35] A. Chouldechova and M. G’Sell, “Fairer and more accurate, but for whom?” *CoRR*, vol. abs/1707.00046, 2017. [Online]. Available: <http://arxiv.org/abs/1707.00046>
- [36] H. Suresh and J. V. Gutttag, “A framework for understanding unintended consequences of machine learning,” *CoRR*, vol. abs/1901.10002, 2019. [Online]. Available: <http://arxiv.org/abs/1901.10002>
- [37] D. Pedreschi, S. Ruggieri, and F. Turini, “Discrimination-aware data mining,” in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, Y. Li, B. Liu, and S. Sarawagi, Eds. ACM, 2008, pp. 560–568. [Online]. Available: <http://doi.acm.org/10.1145/1401890.1401959>
- [38] N. Kallus and A. Zhou, “Residual unfairness in fair machine learning from prejudiced data,” in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. JMLR Workshop and Conference Proceedings, J. G. Dy and A. Krause, Eds., vol. 80. JMLR.org, 2018, pp. 2444–2453. [Online]. Available: <http://proceedings.mlr.press/v80/kallus18a.html>
- [39] Meyer Media LLC, “LawSchoolNumbers,” 2020, [accessed 2020-02-01]. [Online]. Available: <https://lawschoolnumbers.com/>
- [40] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed., ser. Chapman & Hall/CRC Interdisciplinary Statistics. Chapman & Hall/CRC Press, 2018.
- [41] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed., ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2019.
- [42] C. Enders, *Applied Missing Data Analysis*, ser. Methodology in the social sciences. Guilford Publications, 2010. [Online]. Available: <https://books.google.fi/books?id=MN8ruJd2tvgC>
- [43] R. J. A. Little, “A test of missing completely at random for multivariate data with missing values,” *Journal of the American Statistical*

- Association*, vol. 83, no. 404, pp. 1198–1202, 1988. [Online]. Available: <http://www.jstor.org/stable/2290157>
- [44] I. Žliobaitė, F. Kamiran, and T. Calders, “Handling conditional discrimination,” in *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, D. J. Cook, J. Pei, W. Wang, O. R. Zaïane, and X. Wu, Eds. IEEE Computer Society, 2011, pp. 992–1001. [Online]. Available: <https://doi.org/10.1109/ICDM.2011.72>
- [45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [46] Wes McKinney, “Data Structures for Statistical Computing in Python,” in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.
- [47] S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in r,” *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011. [Online]. Available: <https://www.jstatsoft.org/v45/i03/>
- [48] C. Rudin, “Please stop explaining black box models for high stakes decisions,” *CoRR*, vol. abs/1811.10154, 2018. [Online]. Available: <http://arxiv.org/abs/1811.10154>
- [49] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth, 1984.
- [50] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*, ser. Springer Series in Statistics. Springer, 2009. [Online]. Available: <https://doi.org/10.1007/978-0-387-84858-7>
- [51] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [52] M. J. Kearns, “A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split,” in *Advances in Neural Information Processing Systems*

- 8, *NIPS, Denver, CO, USA, November 27-30, 1995*, D. S. Touretzky, M. Mozer, and M. E. Hasselmo, Eds. MIT Press, 1995, pp. 183–189. [Online]. Available: <https://doi.org/10.1162/neco.1997.9.5.1143>
- [53] I. Guyon, “A scaling law for the validation-set training-set size ratio,” in *AT&T Bell Laboratories*, vol. 1, no. 11, 1997.
- [54] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2013. [Online]. Available: <https://doi.org/10.1007/978-1-4614-6849-3>
- [55] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2188395>
- [56] R. G. Mantovani, T. Horváth, R. Cerri, S. B. Junior, J. Vanschoren, and A. C. P. de Leon Ferreira de Carvalho, “An empirical study on hyperparameter tuning of decision trees,” *CoRR*, vol. abs/1812.02207, 2018. [Online]. Available: <http://arxiv.org/abs/1812.02207>
- [57] J. N. van Rijn and F. Hutter, “Hyperparameter importance across datasets,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Y. Guo and F. Farooq, Eds. ACM, 2018, pp. 2367–2376. [Online]. Available: <https://doi.org/10.1145/3219819.3220058>
- [58] J. G. Dy and A. Krause, Eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. JMLR Workshop and Conference Proceedings, vol. 80. JMLR.org, 2018. [Online]. Available: <http://proceedings.mlr.press/v80/>

A Appendix

A.1 Complete Raw Dataset Attribute List

Table A.1: Dataset schema description for application- and user-based attributes as contained in the raw dataset files provided by Rankin [10].

Attribute	Description
<i>user-apps.csv</i>	
application_id	unique application identifier
school	school identifier
status	current status ¹ of application
sentmonth	month of submission
decisionmonth	month of decision received
app_id	identifier for users' applications (per cycle)
cycle	academic year of application, e.g., 0304
sent	date of submission
decision	date of decision received
url	LSN [39] profile URL
money	scholarship received
lsat	Law School Admission Test score
gpa	undergraduate LSAC grade point average
urm	self-identified underrepresented minority
ed	applied for early decision
feewaiver	received fee waiver
complete_ts	timestamp of application completion
decision_ts	timestamp of decision received
aa	self-identified as African American (inferred)
attending	accepted offer of attendance

continued on next page

¹ The status can be either “Ac” (Accepted), “AcWa” (wait-listed, then accepted), “AcRe” (wait-listed, then rejected), “Re” (rejected), “Wa” (wait-listed).

Table A.1: (continued)

Attribute	Description
nontrad	self-identified as non-traditional ²
yearsout	self-identified years since undergraduate degree
<i>user-pages.csv</i>	
app_id	identifier for users' applications (per cycle)
cycle	academic year of application, e.g., 0304
url	LSN [39] profile URL
state	place of residence
race	self-identified race
sex	self-identified sex
yearsout	self-identified years since undergraduate degree
lsat	Law School Admission Test score
gpa	undergraduate LSAC grade point average
schooltype	self-identified type of undergraduate institution
major	self-identified undergraduate major
urm	self-identified underrepresented minority
nontrad	self-identified as non-traditional ²
international	self-identified as non-USA citizen
lgbt	self-identified as LGBT (inferred)
tfa	participated in Teach For America (inferred)
military	self-identified as military veteran
lsat1	first LSAT score (same as <code>lsat</code> above)
lsat2	second LSAT score (if exam retaken)
lsat3	third LSAT score (if exam retaken)

² A non-traditional student is generally someone who took time off school between their undergraduate and graduate education.

A.2 Protected Group Acceptance Statistics

Table A.2: Descriptive statistics of protected group acceptance ratios per user. The groups are African American (aa), LGBT, sex, and minority (mnr).

		count	mean	std	min	Q_1	Q_2	Q_3	max
aa	0	37,731	0.739	0.278	0	0.556	0.800	1	1
aa	1	1,506	0.706	0.320	0	0.500	0.800	1	1
lgbt	0	38,985	0.738	0.280	0	0.556	0.800	1	1
lgbt	1	253	0.734	0.268	0	0.583	0.778	1	1
sex	f	13,456	0.744	0.283	0	0.571	0.833	1	1
sex	m	18,485	0.720	0.277	0	0.500	0.778	1	1
sex	⊙	7,303	0.773	0.277	0	0.600	0.875	1	1
mnr	0	34,639	0.744	0.275	0	0.571	0.800	1	1
mnr	1	4,600	0.692	0.310	0	0.500	0.750	1	1

A.3 Optimized Hyper-Parameters

Table A.3: Results for the hyper-parameter optimization of decision trees and random forests on the complete-case MyLSN dataset after 1000 iterations.

Model	Parameter	Range	Optimum
decision trees	min_simple_split	1–10	8
	min_simple_leaf	2–10	7
	max_depth	10, 25, 50, 100, ∞	25
	criterion	entropy, gini	entropy
	class_weight	1, balanced	1
random forests	min_simple_split	1–10	10
	min_simple_leaf	2–10	2
	max_depth	10, 25, 50, 100, ∞	50
	criterion	entropy, gini	entropy
	class_weight	1, balanced	balanced
	n_estimators	10, 50, 100	50
	max_features	$\log_2 p$, \sqrt{p} , $p/2$	$p/2$