Evaluating the Effect of Data Missingness on Fairness of Machine Learning Algorithms

Yuri Bukhradze , Gyuwan Kim, Matthew Yom, Alec Panattoni

March 2023

Abstract

One of the most common ways that biases can appear in machine learning is data missingness, which is the defined as the presence of data that is not stored for the variable of interest. In particular, we will be looking at selection bias. We begin by first establishing our parameters and definitions, defining our definitions of what it means to be fair in the context of machine-learning. Then, we define the different types of missingness, focusing particularly on Missing at Random (MAR), Missing Completely at Random (MCAR), and Not Missing at Random (NMAR). For our project, we will be creating semi-synthetic data to replicate each type of missingness we will be looking at.

1 Introduction

In pursuit of evaluating the ability of machine learning models to accurately produce results without bias, we are set to explore the ability of data missingness – presence of data that is not stored for the variable of interest – to affect the degree at which models are able to adequately respond to biases in the dataset. We define bias as the failure of our dataset to accurately reflect the reported metrics for the entire population; in particular, we are primarily interested in selection bias, which occurs when the data collection process itself introduces inaccurate presentation of population of interest, thus causing our ML model to produce results that are accurate to the data, but are not necessarily representative of the real world. Selection bias, due to the nature of the process, is prone to leaving out relevant data, thus creating missingness. Hoping to determine how such missing data and patterns of such missingness are able to affect the accuracy of model predictions, we set out to study how different types of missingness and the different ways of handling them affect the ability of the model to make both accurate and fair (i.e. representative of the population in question) predictions.

In order to continue with our discussion, it is important to lay out the definitions of concepts that our crucial to our research; in particular, it is necessary to accurately define algorithmic fairness in machine learning. While fairness can be subjective and have a fairly broad definition, in the machine-learning community fairness is typically defined as the process of correcting and rectifying biased algorithms (occuring as a result of misrepresentation of protected attributes including, but not limited to, race, ethnicity, gender, religion, sexual orientation, disability, and class) in existing machine learning models. Missingness in data can stem from a number of reasons, such as participants' data being intentionally omitted, or participants dropping out of studies, or other reasons. In general, participants whose data is missing tend to be people in marginalized communities. Without such bias mitigating processes, missingness in data can skew the data to make it systematically biased against said communities. An example of a fairness-reliant model that would benefit from the resolution of missingness is a dataset that has multiple features: some belong to the set of sensitive attributes, those that we consider to be subject to bias; and others do not belong to that set, but can cause or be caused by the sensitive attributes. In this case, we assume that the data for the sensitive attribute is not missing. However, missing values for other attributes can be correlated with the sensitive attributes; improper handling of this missingness (e.g. dropping of missing tuples) can directly affect the outcome of the model, predicting different results for groups that might be misrepresnted by the partially missing data. Therefore, it is important for us to resolve the connections between attributes and how the missingness plays a role in the distribution of the data in order to build a model that is able to adequately respond to the existing biases in the data. Otherwise, the algorithm that is not taking missingness into account can produce results that are biased towards a certain group.

As mentioned prior, data missingness is another terminology for the common phenomenon of missing data, when data is omitted from the dataset for some variables either in a systematic pattern or with no discernible pattern whatsoever. There are three different types of data missingness:

1. Missing Completely at Random (MCAR) occurs when the data missing are independent from the observed and unobserved variables. In other words, the missing data is completely independent from all other attributes, including its own value. Because the missing data is not associated with the dataset, we can manipulate this data by removing the data or imputing the data. Our syntehtic dataset featuring MCAR data was generated using a Bernoulli variable given probability p (by default set at 0.2), which determines whether a specific data point would be ommitted or not.

2. Missing at Random (MAR) occurs when the probability of the missingness is dependent on some other values in the dataset but not within the values of the column where the data is missing. In this case, it is best practice to probabilistically impute based on the value(s) of the attribute(s) that this column is dependent on. To generate MAR dataset, we used Normal distribution for the generation of probability, passing it further through a Sigmoid function in order to bring it within the range of valid probability. For categorical variables, each value was first randomly associated with

a specific probability using normal distribution; for numerical variables, the points were normalized to Z-value. These techniques were applied to the *dependent* column which determines the probability for the *affected* column (where the data is actually being made missing).

3. Not Missing at Random (NMAR) indicates missingness that is dependent on the value of the variable itself. Following the previous example, NMAR missingness would occur if the data about salary for people with lower pay was missing – i.e, the low value of the pay itself causes the data to be missing (for example, due to unwillingness to report low pay). NMAR type of missingness cannot be accurately evaluated statistically and instead requires subject domain knowledge to hypothesize possible causes of missingness – however, it is possible for us to generate an NMAR dataset by creating a probabilistic threshold for the values that will be made missing. The technique for simulation was similar to that described in MAR, but without the differntiation of dependent and affected columns: the probability is calculated on the affected column itself, and this determines the missingness of the value itself, given the definition of NMAR.

When we choose a dataset to use for this project, we can never accurately identify the type of missingness of each attribute containing missingness. We can only speculate as to what type of missingness is present within these attributes. This brings on the question: how can we study the effects of handling missingness for each different type if we can't know for sure which type of missingness an attribute is in the first place? The solution to this issue comes with creating our own semi-synthetic data. In other words, we can produce each type of missingness for an attribute and guarantee that the attribute is that type of missingness based on how its values are generated. In order to create MCAR missing data, we can randomly drop values within the attribute if a generated probability is less than a chosen threshold. For NMAR data, we must do something similar, in that for each distinct value in the missingness attribute, we must choose a different threshold to designate whether a value in the attribute will be let missing. This way, missingness depends on the value of the attribute itself (and possibly other attribute(s), if the values of others are included in the probabilistic designation). In order to generate a probability based on this value, we must use some kind of function. If the value is categorical, we can use a One-Hot Encoder, and generate a separate probability for each OHE column. If it is a numerical value, we can simply use a sigmoidal function to produce probabilities once our threshold is set. Lastly, we must be able to create MAR data. This is done using the same strategy that was used for creating NMAR data, except we are only using the values of the other attribute(s) that the missingness of the attribute is dependent on. In the same sense, we have different values with different thresholds within the other attribute(s), which will determine whether each value of the missingness attribute is indeed missing.

For reasons of simplicity, the missingness attribute will remain the same for each type of missingness. In other words, the same column will be used for creating each type of missingness. In order to test how handling of different types of missingness affects fairness notions, we'll have to use each individually. These include dropping all missing values in the attribute, imputing based on the mean of the attribute, and probabilistically imputing based on a single other attribute. Each technique will be used for each type of missingness independent from one another to observe how our chosen fairness notions change.

2 Methods

The methodology for finding fairness notions in this model is implemented using a combination of one-hot encoding (OHE), logistic regression, and fairness metrics. The dataset has been divided into training data and test data. The training data is used to fit the logistic regression model, and the test data is used to evaluate the model's fairness. One-hot encoding is applied to the categorical variables in the training data. This is done using the OneHotEncoder from the sklearn library, which transforms categorical variables into numerical variables by creating a new binary column for each unique category. The OHE is then fit to the training data, and the resulting transformed data is concatenated with the non-categorical features. Next, a logistic regression model is fit to the training data. The logistic regression model is used to predict the target variable, "substantiated", based on the OHE-encoded categorical variables and the non-categorical variables. The model is fit using the "balanced" class weight, which helps balancing the weight of each class in the target variable. Once the model is fit, the OHE-encoding is applied to the test data, and the non-categorical variables are concatenated with the OHE-encoded categorical variables. The logistic regression model is then used to make predictions on the test data. Finally, three fairness metrics are calculated to evaluate the fairness of the model. These metrics are the statistical parity difference, the average odds difference, and the equal opportunity difference. Statistical parity difference measures the difference in mean prediction outcomes between two groups. Average odds difference measures the difference in odds of positive prediction outcomes between two groups. Equal opportunity difference measures the difference in true positive rates between two groups.

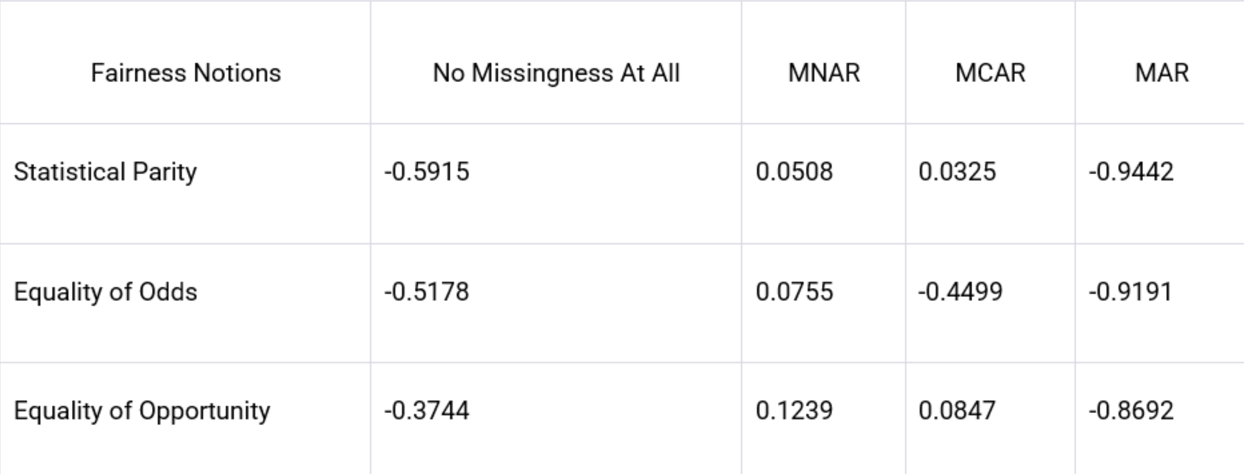
NMAR

the probability distribution for NMAR can be written as: $P(R|Y_{obs}, Y_{mis}, q)$ This says that the probability of whether a position in R is 0 or 1 depends on both Y obs and Y_{mis}, and this relationship is governed by q.

MAR

The probability distribution for MAR can be written as: $P(R|Y_{obs}, q)$. This means that missingness depends only on Yobs, and this relationship is governed by q.

3 Results



Fairness Notions	No Missingness At All	MNAR	MCAR	MAR
Statistical Parity	-0.5915	0.0508	0.0325	-0.9442
Equality of Odds	-0.5178	0.0755	-0.4499	-0.9191
Equality of Opportunity	-0.3744	0.1239	0.0847	-0.8692

The findings presented in the table allow for a comparison of three distinct fairness conceptions for various kinds of missing data. statistical parity, equality of odds, and equality of opportunity are the concepts of fairness, and they are all frequently used measurements of algorithmic fairness in the fields of machine learning and artificial intelligence.

Statistical parity measures the difference between the probability of one group being preferenced over the other; the closer the value is to 0, the better is the degree of fairness in the algorithm. Without any missingness introduced to our dataset, SP is measured at -0.5915, indicating a mid-high degree of bias. With the missingness introduced NMAR, statistical parity drops close to zero; similar is for missingness MCAR, indicating that these types of missingness distribution correlate to higher degree of fairness in the algorithm. MAR, which relies on interdependence between variables, leads to high score of -0.9442, indicating a high degree of bias in terms of choosing based on a sensitive attribute.

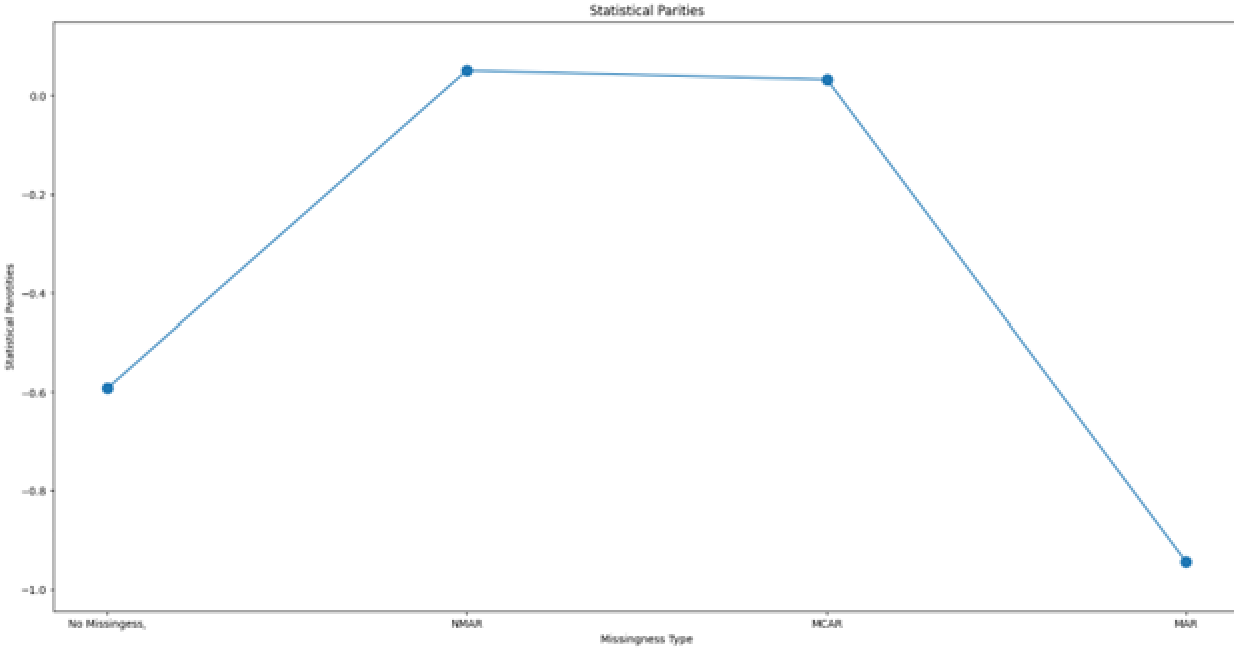
For our outcomes, we want to achieve a The outcomes demonstrate that the statistical parity measure for the first scenario, "fairness assumptions for No Missingness At All," is -0.59, indicating a modest bias in favor of one side. The equality of odds measure is -0.52, which also denotes bias. The model performs better in this area according to the equality of opportunity metric, which has the most favorable results (a value of -0.37).

In contrast to the first scenario, the second scenario, "fairness ideas for NMAR," exhibits better fairness outcomes. Positive measurements of statistical parity and equality of odds show that neither group is the target of the model's bias. The equality of opportunity metric, with a value of 0.12, is still comparatively low.

The final scenario, "Fairness Concepts for MCAR," yields a range of outcomes. In contrast to the equality of odds measure, which is negative and indicates bias, the statistical parity measure is positive, showing no bias. The equality of opportunity measure has a 0.08 value, which is low but favorable.

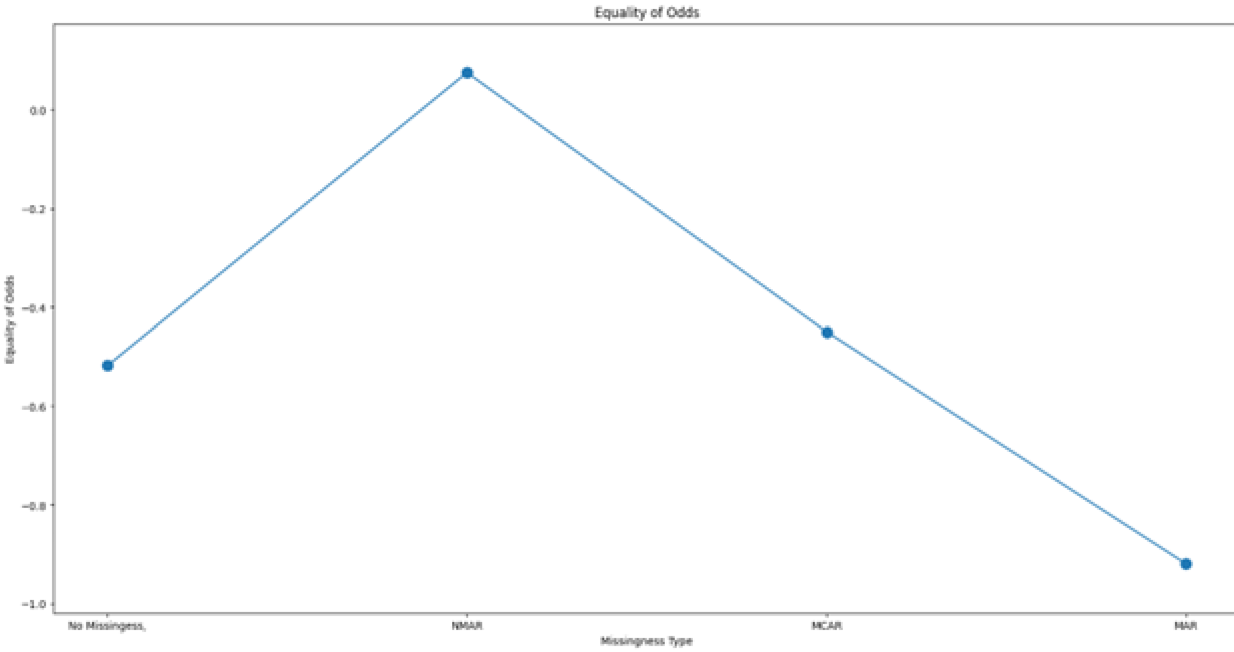
The worst outcomes in terms of fairness can be seen in the final scenario, ”fairness ideals for MAR.” Both the statistical parity and equality of odds measures are far from 0 (-0.9442 and -0.9191 respectively), demonstrating a clear bias in favor of one group. With a value of -0.87, the equality of opportunity measure is likewise unfavorably skewed.

Statistical Parity The equality of results for several groups of people is referred to as statistical parity. In other words, statistical parity states that results of the model should be comparable for various groups of people, such as men and women or individuals of various racial or ethnic backgrounds. This idea of fairness contributes to ensuring that the model does not discriminate against any one group.



Equality of Odds

Equal odds refers to the true positive rate (TPR) and false positive rate (FPR) being the same for various groups of people. In other words, the rate at which members of various groups are rightly or incorrectly classified should be the same. This idea of fairness makes sure that no group is unfairly targeted by the model and that everyone has an equal chance of being accurately categorized.

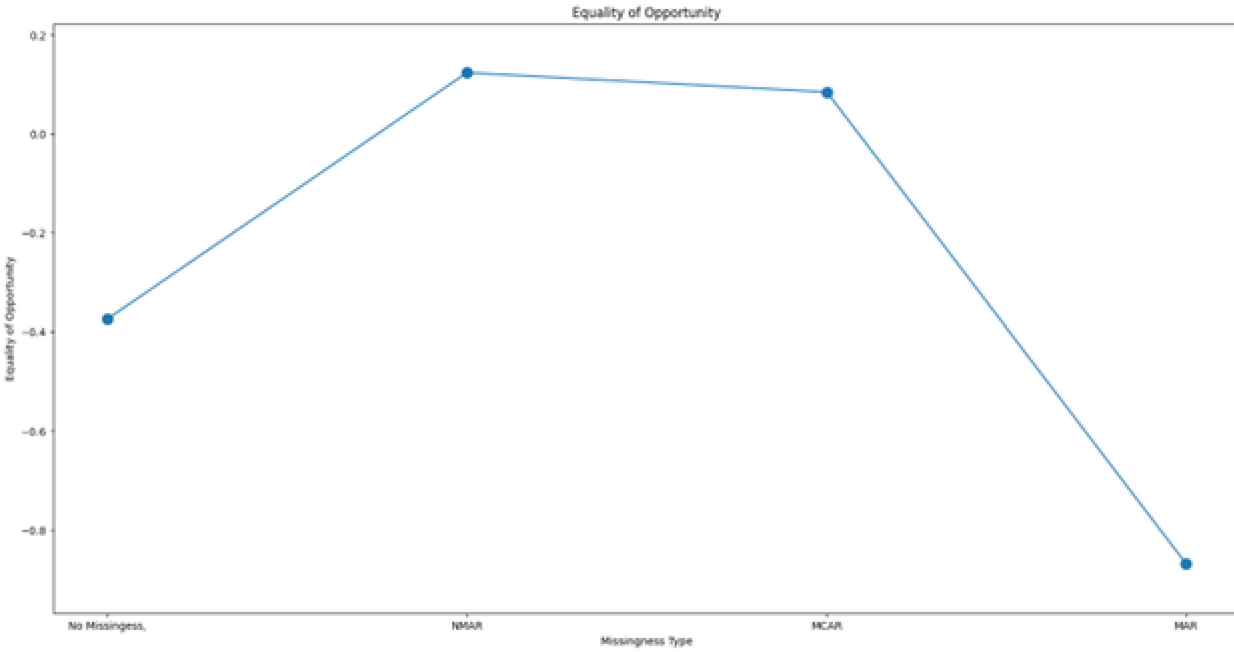


Equality of Opportunity

Equality of Opportunity allows for individuals within the ”advantaged” group to be given an equal chance of correctly classified. In other words, we want the True Positive Rate to be equal for every attribute in the protected class. Equality of opportunity is a type of fairness where each sample, model should be treated similarly, with no bias or preference, except where particular distinctions can be explicitly justified.

Equality of Opportunity

Equality of opportunity refers to a situation where everyone has an equal chance of success, regardless of factors such as race, gender, or background. In this context, the graph is showing how the level of equality of opportunity is affected by the type of missing data present in the scenario being analyzed.



4 Discussion

5 References

6 Contributions

Alec Panattoni: Contributed to writing introduction, structuring repository, docker-related tasks, wrote code for missingness generation, set up targets, wrote test data creation notebook, and wrote data cleaning code

Yuri Bukhradze: Contributed to writing the introduction and the section on the types of missingness. Worked on the code that handled creation of visualizations and also contributed to the early versions of the functions for generating the synthetic dataset. Helped with interpreting the results and worked on writing the conclusion. Added the appendix.

Gyuwan Kim: Contributed to writing the types of missingness and equality of opportunity. Worked on the code to create models and evaluate our fairness metrics on each of the different types of missingness. Also wrote the code to visualize the output of our results.

Matthew Yom: Contributed to writing the Abstract at the top of the page, and writing the sections on the methodology and the results of our data.

7 Appendix

The following is the problem statements of our project proposal from the first quarter for reference on our progress.

7.1 Broad Problem Statement

The goal of this project is to describe how different types of missingness affect fairness in Machine Learning (ML). Missing data is very problematic in the world of machine learning because we cannot create reliable machine learning models when there is missingness in the data. To handle missing data, it is not as simple as dropping all tuples that contain missingness. Instead, we need to first understand why certain data is missing. Furthermore, missing data can cause problems when creating ML models with considerations for fairness: the ability of the model to adequately respond to the biases present in the training dataset.

However, what exactly are fairness-based ML models, and what does it mean to create a model that is fair? While fairness can be subjective and have a fairly broad definition, in the machine-learning community fairness is typically defined as the process of correcting and rectifying biased algorithms (of protected attributes including but not limited to race, ethnicity, gender, religion, sexual orientation, disability, and class) in existing machine learning models. Missingness in data can stem from a number of reasons, such as participants' data being intentionally omitted, or participants dropping out of studies, or other reasons. In general, participants whose data is missing tend to be people in marginalized communities. Without such bias mitigating processes, missingness in data can skew the data to make it systematically biased against said communities. An example of a fairness-reliant model that would benefit from the resolution of missingness is a dataset that has multiple features: some belong to the set of sensitive attributes, those that we consider to be subject to bias; and others do not belong to that set, but can cause or be caused by the sensitive attributes. In this case, we assume that the data for the sensitive attribute is not missing. However, missing values for other attributes can be correlated with the sensitive attributes; improper handling of this missingness (e.g. dropping of missing tuples) can directly affect the outcome of the model, predicting different results for groups that might be misrepresnted by the partially missing data. Therefore, it is important for us to resolve the connections between attributes and how the missingness plays a role in the distribution of the data in order to build a model that is able to adequately respond to the existing biases in the data. Otherwise, the algorithm that is not taking missingness into account can produce results that are biased towards a certain group.

In our consideration of missingness, it is important to note that we can only infer why certain data is missing; there is no approach that would allow us to accurately recreate the missing data. Therefore, contextual and domain knowledge must be applied to make the best guess regarding the source of missingness and the correlation of missing points to other attributes. The way that missingness is handled depends on our assessment of the ”type” of missingess for each attribute, which is a classification of the relationship between attributes that is causing the missingness among the data. In this paper, we aim to observe how different types of missingness compare in terms of fairness notions; more specifically, we are interested in exploring how differing types of missingness affect fairness notions, metrics that exist to evaluate to what extent the model is able to approach biases in the data.

7.2 Specific Problem Statement

Over the course of this capstone class, we have learned several ways to ensure fairness in our data through different techniques, whether they be pre-processing, in-processing or post processing techniques. In the paper ”The Importance of Modeling Data Missingness in Algorithmic Fairness: A Causal Perspective”[**islam**], The authors discuss each type of technique, and their benefits and drawbacks. In addition, they apply each methods in the paper as a demonstration of the benefits of such algorithms and what they can do. As a result, we decided to try and apply it ourselves and tested them out on different datasets for our own research. After our initial research, we learned the impact of missing data and how such missing information lead to biases about a target population. With this, creating models and estimating populations based on missingness leads to the question of whether or not the models we create are a true representation of the population we would like to study. While this paper did provide useful information about how missingness can impact fairness in different ways, it was not clearly exemplified through real-world examples. Moreover, it was not shown through example the effect that different types of missingness have on overall fairness notions.

Previous courses had taught us the different types of missingness; Missing at Random, Missing Completely at Random and Missing Not at Random. Missing at Random is the type of missingness in which missing values in a column have a clear relationship with values in another column. In order to determine if an attribute has this type of missingness, statistical tests like permutation testing are often used to determine if the distributions of outcomes are not due to chance. An example of this could be the ”Name” column being associated with low test scores in a self-reported test scores dataset. The name of students with low test scores is missing because they likely don’t want others to know that they got that low test score. Missing Not at Random is the type of missingness in which missingness is dependent on the attribute in question itself. With the example of the recidivism classifiers, people might leave their ethnicity missing because they are afraid their ethnicity will be used against them. Therefore, they left the value missing because of the value itself. Missing Completely at Random is the type of missingness in which their is no real identifiable reason for why the data is missing. In other words, it is simply a random sample, and we cannot expect all data to be present in any random sample.

Now that the types have been explained, it can be explained how each type is generally handled. For MNAR (or NMAR) data, because the missing data depends on the column itself, we could impute by modeling the data of the missing column itself. An example of this would be probabilistically imputing based on that column's values and their probabilities. For MCAR datas, we can completely drop the values, or estimate missing values using aggregatations of the attribute like the mean, since missingness is essentially due to random sampling. Lastly for MAR, because missingness depends on values from other columns, we can estimate the value by, again, probabilistically imputing, but this time with subsets differing according to the values of the column this one is dependent on.

Before we do any of this, we need to produce data that will reliably reflect each type. In order to do so, we will be semi-synthetically creating our data. There are many different libraries provided which provide us with ways to create the missing data. We could also hand-craft the data ourselves, but it more likely that we will use libraries to do so. We have not yet discussed with our mentor which libraries will be used to do this, but this will certainly be addressed in week 1 of quarter 2. It is important to mention that when creating the data, we will not leave missingness within our sensitive attributes (only non-sensitive attributes). The reason for this is that there are issues of fairness within sensitive attributes themselves that could confound our results. In order to compare how each type of missingness performed we will need each of the three datasets to be very similar. This is because there needs to be a general baseline for fairness notions so that one dataset is not already much more fair than the others before anything is done to the data. Therefore, the same attributes and distributions that values are drawn from will be used for each. The only significant difference between the three will be the type of missingness in each. Once we have created the data, we will handle each form of missingness as previously mentioned.

Once missingness has been handled accordingly for each type, we will be producing predictions with a single model to test how certain fairness notions change between them. In terms of fairness notions, we plan to use statistical parity, equality of odds, and equality of opportunity. Statistical parity, which occurs when the following equation is satisfied:

$$P(R = +|A = a) = P(R = +|A = b) \forall a, b \in A$$

where A is the set of sensitive characteristics and R is the predictive outcome. In a perfectly fair algorithm, the two sides of the equation are equal; therefore, the closest we get the difference between $P(R = +|A = a)$ and $P(R = +|A = b)$ to 0, the better our intervention technique is at improving fairness. Equality of Odds is a much stricter version of equality of opportunity. Instead of using True Positive Rate to be equal in every attributes, it rather takes in the lowest False Negative Rate and uses them as the guide to make all the rate to be equal in all of the attributes. Equality of Opportunity allows for individuals within the "advantaged" group to be given an equal chance of correctly classified. In other words, we want the True Positive Rate to be equal for every attribute in the protected class. Equality of opportunity is a type of fairness where each sample, model should be treated similarly, with no bias or preference, except where particular distinctions can be explicitly justified.

7.3 Project Output

To explain our potential solutions to addressing the specific types of data missingness, we will be writing a paper on how each type of missingness affects fairness in Machine Learning. After predictions are produced for each type of missingness through a single model, (same model used for each) we will compare and assess our chosen fairness notions for each. Each mentioned before, these include statistical parity, equality of odds, and equality of opportunity. Once we have our results, we will compare each missingness type and reason through what is observed. We will also produce plots and visualizations to accompany our results and conclusions.