ESADE
Business School
Ramon Llull University

To obtain the degree awarded by ESADE in

Master in Business Analytics

Master Thesis 2019-2020

Detecting Unconscious Bias and Discrimination in Algorithms

Student:

Aleksandra Śledziewska

Director or Personal Faculty Advisor:

Marc Torrens Arnal

Associate Professor, Department of Operations,

Innovation and Data Sciences at ESADE

Barcelona, October 2nd, 2020

Abstract

This thesis is intended to introduce and better understand the concepts of bias and fairness of algorithms, and to review open-source tools available online to facilitate bias detection and encourage companies and individuals to audit their algorithms for the presence of discrimination against some protected groups.

The methodology includes the literature review to search for different definitions of fairness and to understand the importance and prevalence of discrimination in algorithms. Secondly, an exploratory analysis with visualizations is conducted for the COMPAS algorithm dataset, which is then used to generate bias and fairness reports using two open-source bias detection tools – Aequitas and What-If Tool.

Concerning the results and conclusions, firstly, the set of different definitions of fairness and their applications proves that one standardized perception of fairness does not exist, and it usually becomes an ethical question on how to interpret it. Secondly, the results of the analysis confirm the presence of racial and gender bias in the COMPAS algorithm. However, the focus is mostly on presenting the capabilities of the chosen tools, which offer a broad range of interesting visualizations and reports, and considerably facilitate detecting and understanding where bias can exist in the data or the machine learning model. Such tools are essential to think about bias detection as the standard process of building models in companies and monitoring their performance afterward. Nonetheless, auditing algorithms is just one step in the process of mitigating and eradicating discrimination from them. Joint efforts and further work are required to make it a common practice.

Contents

List of Figures

Introduction

Artificial Intelligence and data analytics have been entering into almost every sector and every part of our lives. They are becoming prevalent in sensitive industries, such as healthcare, judiciary, or recruitment processes. However, we are often unaware of whether the decisions and suggestions of the Artificial Intelligence algorithms are fair or whether they discriminate against some minority groups, gender, or race. What is more, deep learning and neural networks – with so-called black-box models – are becoming more and more popular, and we understand them even less than the standard models. We are delighted with the models, which optimize our decisions, analyze much more data than any human being is capable of, and make our businesses thrive. Nonetheless, we should pay more attention to the "way of thinking" and the process of automated decision-making of the algorithms, because they can imperceptibly bake in and scale the biases present in human behavior.

The main objective of this thesis is to understand the concepts of bias and fairness of algorithms, especially those used in sensitive areas, and review some open-source tools available for individuals and companies to facilitate and propagate bias detection. The main questions to be answered through this analysis are: "What does the fairness of an algorithm mean?" and "What are the tools available to facilitate bias detection?".

The methodology of the research includes several levels of analysis. The literature review was conducted, first, to understand the importance of bias detection by examining what has been done in this area hitherto and where bias has been detected, second, to understand different approaches to fairness and possible mathematical metrics to be used. Afterward, the dataset used in further analysis – the COMPAS algorithm dataset

(ProPublica, 2016) – was described and an exploratory analysis with visualizations was conducted. Finally, the aforementioned dataset was used to test chosen open-source bias detection tools – “Aequitas” (Saleiro et al., 2019) and “What-If Tool” (People + AI Research (PAIR), 2020). The generated reports and visualizations from these tools were presented and summarized.

This paper consists of four chapters. The first chapter includes a discussion about the relevance of bias detection from different perspectives, and the literature review introducing examples of biased algorithms having a significant impact on human lives. In the second chapter bias and fairness definitions, as well as some tips on how to choose the fairness metric appropriate for a given problem, are provided. The third chapter is a description of the COMPAS algorithm dataset with an exploratory analysis of variables distributions in different groups. Finally, the fourth chapter gives an overview of open-source bias detection tools, and for two of them – Aequitas and What-If Tool – it provides the examples of reports and visualizations generated for the COMPAS dataset with the results discussed.

Chapter 1

Importance of Bias Detection

1.1 The relevance of algorithm audits

The topic of bias in algorithms becomes increasingly important nowadays. We believe that algorithms can make “better” and more accurate decisions than humans because they can analyze much more data and are not characterized by the imperfections of human beings. However, meanwhile, people can be biased in many different ways, and things like mood or fatigue can affect their decisions, we also have to bear in mind that algorithms are based on data generated by humans. Therefore, the bias present in our behavior and decisions can be transferred through algorithms, which learn based on this data. If we want to make more just decisions, we have to mitigate the problem of bias in algorithms. The first step in this direction is to understand where and when bias appears by auditing the algorithms (Silberg & Manyika, 2019).

The task of bias detection can be exceptionally difficult because there are examples of algorithms that, despite indisputably law-abiding, may unwittingly contain bias, for instance, carried by some variables included in the model. Selected examples from the literature of such algorithms are given later in this chapter. By law, some algorithms cannot take into account variables such as race, gender, ethnic group, religion, etc. to predict results of e.g. hospital admission or judicial sentence, because it can directly lead to discrimination against some groups. However, despite legal correctness, bias

may be still present in the underlying data of an algorithm, for example, due to human biases preserved in historical records or the unprivileged position of some groups in the past. It often happens without the awareness of people responsible for the creation of the algorithm. With the rapid spread of Artificial Intelligence, legal systems have not yet been adjusted adequately to handle new types of issues arising from decisions being made by machines. There are not enough regulations, and very few algorithms are audited in terms of responsibility and potential breaches of the law.

However, new laws and regulations are constantly being created, irrefutably they also include regulations of algorithms and Artificial Intelligence, to measure up to rapidly changing reality. Some countries have already taken action in this direction. In the USA in 2019 Algorithmic Accountability Act was introduced, which requires entities to conduct automated system impact assessments and data protection impact assessments (Clarke, 2019). Similar regulations can be expected to arise in different countries as well. Detecting potential problems, which involve discrimination, in algorithms in advance can help companies to avoid serious consequences in the future, and get better prepared for the upcoming regulations.

Finally, this topic is also relevant from an individual perspective of all of us. Such algorithms have an uncontested impact on our lives, for example, when someone applies for a loan in a bank, needs to be admitted to a hospital or is looking for a job. We want to know why a given decision is made, and whether it is fair. If algorithms make unwarranted and biased decisions, the sense of injustice in society may increase, which may lead to further unrest. With a rising number of disciplines using Artificial Intelligence to support the decision-making process and having an influence on more aspects of our life, it becomes an even more significant issue.

1.2 Examples of biased algorithms

We can already find examples confirming that bias has been detected in algorithms of high importance in society. Multiple sources point out and cite similar works, which

have been conducted in this area. They include mainly: the COMPAS algorithm used in the US legal system, the algorithm used in chosen American hospitals determining whether a patient requires extra medical care, Google ads providing different content for black/white or female/male users, and hiring processes leading to biased underlying data and algorithms discriminating particular groups.

Regarding previous research that has been conducted in the field of bias and discrimination of algorithms, in 2016 ProPublica, which is an independent, nonprofit newsroom producing investigative journalism (ProPublica, 2017), analyzed one of the risk assessment algorithms used in the US legal system to assess a likelihood of a criminal becoming a recidivist – an algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) made by Northpointe Inc. (Larson, Mattu, Kirchner, & Angwin, 2016). They focused their research on racial discrimination of this algorithm, and they found out that black criminals were much more often classified incorrectly by the algorithm as being of a high risk of recidivism, while white criminals were more often misclassified as low risk. Additionally, they analyzed the data by gender. The tools used by ProPublica include mainly data visualizations, logistic regression, and contingency tables, and they provide access to their data and notebooks on GitHub (ProPublica, 2016). More details of their analysis are described in Chapter 3.

Moreover, Obermeyer, Powers, Vogeli, and Mullainathan from the University of Chicago and California as well as Brigham and Women's Hospital in Boston analyzed an algorithm used in the above-mentioned hospital to determine whether a patient requires extra medical care to optimize costs and provide the best care for their patients (Obermeyer, Powers, Vogeli, & Mullainathan, 2019). They concluded that according to the algorithm white patients are more likely to be recommended extra care than black patients, even if their health conditions are the same. The authors claim that despite analyzing only one algorithm, similar algorithms are used in different hospitals. Therefore, the problem may be much more prevalent and may affect up to 200 million patients in the US. Additionally, Mullainathan commented that we cannot find a lot of examples of biased algorithms available for the public, because they are often protected by the

companies (Gershgorn, 2019).

On the other hand, Latanya Sweeney – a professor at Harvard University – analyzed types of Google ads appearing for black-sounding and white-sounding names typed in the search engine (Sweeney, 2013). She proved that black-sounding names are more often associated with arrest records than white-sounding names. Although it may seem less relevant at first, in the current era of searching for most of the information online, it can have a negative impact, for instance, on job search as it may lead to confusion with another person or negative, although false, first impression. Furthermore, Datta, Tschantz, and Datta in their analysis of Google ads proved that women received fewer ads of high-paying jobs than men (Datta, Tschantz, & Datta, 2015). Altogether, it demonstrates that search engines can be another field in which discrimination and bias prevail, and it may have a significant impact on our everyday lives.

At the other end of the spectrum, some researchers have conducted experiments on the labor market and hiring processes. They analyzed human decisions that were made to hire somebody or not. Bias usually comes from the underlying data, so if bias in human decisions exists, it may also drive discrimination or bias in the machine learning algorithms used in the hiring process. For example, in their work O'Brien and Kiviat concluded that including credit history when deciding whom to hire discriminates female and black candidates (O'Brien & Kiviat, 2018). A bad credit history reduces the probability of hiring a woman more than a man and reduces the starting salary offered to black candidates compared to white candidates. In another research, Bertrand and Mullainathan established that white-sounding names receive 50% more callbacks for interviews in the recruitment process than black-sounding names and that the racial gap exists among different occupations, industries, and size of the company (Bertrand & Mullainathan, 2003). As evidence of the hypothesis that automated algorithms may also convey bias can be given an experimental Amazon's recruiting tool to screen candidates' resumes and search for exceptional candidates. The algorithm, through particular keywords, was favoring male candidates, and penalizing keywords more prevalent in women's resumes (Dastin, 2018).

The above literature review presents that some algorithms have already been studied, and usually the outcomes pointed out that the bias or discrimination exists in these algorithms. However, there is still a huge spectrum of algorithms to be analyzed, as it is not yet a common practice nowadays. Nevertheless, taking into consideration how much these algorithms can affect our lives, we should pay more attention to fairness and explainability of algorithms in use. Furthermore, it is highly important to encourage companies and institutions to audit their algorithms by providing them with easy-to-use, open-source tools facilitating bias detection.

Chapter 2

Bias and Fairness Definitions

2.1 Bias definition

As Cambridge Dictionary defines, bias is “the action of supporting or opposing a particular person or thing unfairly, because of allowing personal opinions to influence your judgment” (Cambridge Dictionary, 2020). Definitions in different sources are very similar, and all of them point to the notion of fairness. Therefore, it is crucial to define as well what fairness is, meanwhile fairness can be approached from multiple perspectives. Different types of fairness in the context of algorithms include disparate impact, statistical parity, conditional statistical parity, equal opportunity, equalized odds, conditional procedure accuracy equality, predictive parity, calibration, balance for positive/negative class, etc. Only metrics used in the subsequent analysis are going to be explained because explaining every metric would be too extensive.

2.2 Fairness definitions

The mathematical definitions provided below are adjusted to a binary classification problem, where each case can be classified as 1 – positive class – or 0 – negative class. There are also at least 2 different groups of people with a certain characteristic, which can be the basis of discrimination (e.g. female and male). Most of the more advanced

fairness formulas are based on the basic metrics of the confusion matrix:

- True Positive (TP) – both predicted and actual outcomes are in a positive class;
- False Positive (FP) – a case, in reality, belongs to the negative class but was predicted as the positive class;
- False Negative (FN) – a case, in reality, belongs to the positive class but was predicted as the negative class;
- True Negative (TN) – both predicted and actual outcomes are in the negative class.

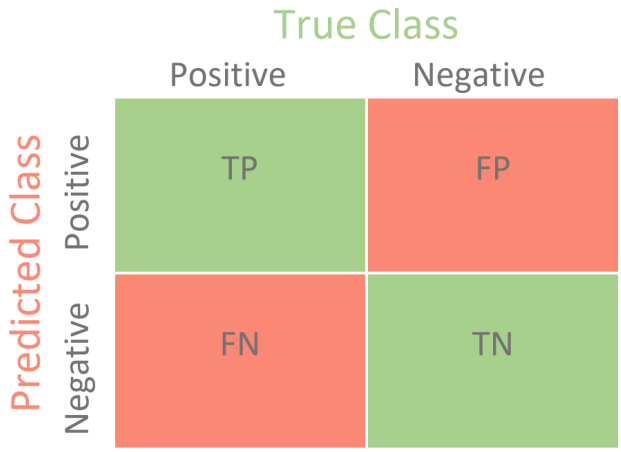

Figure 2.1: Confusion matrix for binary classification.

Source: (Mohajon, 2020)

The definitions provided below are based on the article written by Sahil Verma and Julia Rubin – “Fairness Definitions Explained” (Verma & Rubin, 2018).

Statistical Parity (Group Fairness, Demographic Parity, Equal Acceptance Rate) – all groups of people have an equal probability of being assigned to the positive predicted class.

$$\frac{TP + FP}{TP + FP + TN + FN}$$

Example. To provide a feasible example of the use of each fairness metric, let’s assume that a government decided to create a subsidy for a very expensive treatment. An

algorithm was trained, which helps to assess which patients should be qualified for the treatment, and which patients should not. The algorithm assesses patients' eligibility for the treatment subsidy based on their various characteristics, such as age, gender, income, the severity of illness, etc. There are patients of different origins in the group – Asian, American, European, and African – and we want to ascertain that none of these nationality groups is discriminated against. The algorithm satisfies Statistical Parity if a similar proportion of patients from each nationality group is classified as the positive predicted class eligible for the treatment subsidy. It means that the same percentage of each group receives the subsidy.

Overall Accuracy Equality (Accuracy Parity, Equal Accuracy) – all groups of people have an equal proportion of cases correctly classified to positive or negative class.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Example. The algorithm satisfies Overall Accuracy Equality if it correctly classifies the same percentage of patients in each nationality group. Therefore, all nationality groups have the same proportion of patients correctly classified by the algorithm as being eligible or not eligible for the treatment subsidy.

Predictive Parity (Precision Parity) – all groups of people have equal Positive Predictive Value (PPV), where $PPV = \frac{TP}{TP+FP}$. In other words, in both protected and unprotected groups subjects assigned to the positive class have an equal probability of actually belonging to the positive class.

$$\frac{TP}{TP + FP}$$

Example. The algorithm satisfies Predictive Parity, if in all nationality groups patients, who are classified by the algorithm as being eligible for the subsidy, have the same probability of needing the treatment subsidy in reality. However, this metric is more often used for cases, where being classified as the positive predicted class has a punitive character, so it can cause harm to individuals. For instance, such a case would be an

algorithm used to predict whether a suspect of a crime should be temporarily arrested before the trial (for example, because of the risk of not appearing in court or posing a threat to others) or not. Then, being classified as the positive class is harmful to an individual, and we want to ascertain that the majority of individuals in each nationality group placed under temporary arrest is classified correctly.

False Discovery Rate Parity – in both protected and unprotected groups subjects assigned to the positive class have an equal probability of being incorrectly classified as the false positive.

$$\frac{FP}{TP + FP}$$

Example. The algorithm satisfies False Discovery Rate Parity if in all nationality groups patients, who are classified by the algorithm as being eligible for the subsidy, have the same probability of needing the treatment subsidy in reality. The sum of False Discovery Rate and Positive Predictive Value is always equal to 1, so both of them have the same applications. Ergo, this metric is also more often used for cases, where being classified as the positive predicted class has a punitive character, such as the temporary arrest case. The only difference is that we compare the proportion of individuals in each nationality group placed under temporary arrest classified wrongly.

False Positive Rate Parity (False Positive Error Rate Balance, Predictive Equality) – in all groups of people subjects belonging to the negative class have an equal probability of being classified as the false positive.

$$\frac{FP}{TN + FP}$$

Example. The algorithm satisfies False Positive Rate Parity if in all nationality groups patients, who, in reality, should not receive the subsidy, have the same probability of being wrongly classified as being eligible for the treatment. Once again, this metric is more often used for cases, where being classified as the positive predicted class has a punitive character, such as the temporary arrest case. Then, the priority is given to equalizing the proportion of individuals mistakenly placed under temporary arrest.

False Negative Rate Parity (False Negative Error Rate Balance, Equal Opportunity) – in all groups of people subjects belonging to the positive class have an equal probability of being classified as the false negative.

$$\frac{FN}{TP + FN}$$

Example. The algorithm satisfies False Negative Rate Parity, if in all nationality groups patients, who, in reality, should receive the subsidy, have the same probability of being wrongly classified as not being eligible for the treatment. In this case, the priority is to check the proportion of patients in each nationality group who needed the subsidy but did not receive it. It is especially crucial because there is a chance that if they do not receive the subsidy, they may be unable to afford the treatment, which can cause a danger to life.

Equalized Odds (Conditional Procedure Accuracy Equality, Disparate Mistreatment) – combines False Positive Rate Parity and False Negative Rate Parity is such a way that the classifier has to satisfy both of them. Therefore, in all groups of people subjects belonging to the negative class have an equal probability of being classified as the false positive. Simultaneously, in all groups of people subjects belonging to the positive class have an equal probability of being classified as the false negative.

$$\frac{FP}{TN + FP} \quad and \quad \frac{FN}{TP + FN}$$

Example. The algorithm satisfies Equalized Odds, if simultaneously in all nationality groups patients, who, in reality, should not receive the subsidy, have the same probability of being wrongly classified as being eligible for the treatment, and patients, who, in reality, should receive the subsidy, have the same probability of being wrongly classified as not being eligible for the treatment. Therefore, we pay attention to both types of mistakes the algorithm makes – patients that may feel uncredited because of erroneously not being qualified for the subsidy and those mistakenly qualified for it despite not being eligible.

False Omission Rate Parity – in both protected and unprotected groups subjects assigned to the negative class have an equal probability of being incorrectly classified as the false negative.

$$\frac{FN}{TN + FN}$$

Example. The algorithm satisfies False Omission Rate Parity if in all nationality groups patients, who are classified by the algorithm as not being eligible for the subsidy, have the same probability of needing the treatment subsidy in reality. Therefore, the focus is on the proportion of patients among those who were not qualified for the subsidy in each nationality group who should receive it. Once again, in this case, it is exceptionally important because patients wrongly unqualified for the treatment subsidy can feel uncredited, and it may have a negative impact on their health.

Fairness Through Unawareness (Group Unaware) – this type of fairness definition is satisfied if no protected attributes (such as gender, race, ethnic group, etc.) are used for training the algorithm. This type of fairness is often imposed by law.

Example. The gender of applicants cannot be taken into consideration when determining automobile insurance rates in the U.S. state of California (Carrns, 2019). However, the examples of algorithms provided in the previous chapter evidence that it is not sufficient to exclude the protected variables, because the discrimination may still prevail by including other attributes somehow related to the protected groups. Hence the fairness through unawareness does not guarantee real fairness of automated decision-making.

Counterfactual Fairness – this notion assumes that changing the protected attributes of subjects does not impact the output for these subjects, ceteris paribus (Kusner, Loftus, Russel, & Silva, 2020). This definition allows a possibility that the protected attributes can cause changes in other unprotected attributes leading to changes in the output. This definition emphasizes that it is not enough to exclude protected attributes from the model to guarantee fairness.

Example. Let's assume that being a single parent decreases someone's predicted credit score. If women tend to be single parents more often than men, the algorithm

would discriminate women even if we exclude the protected variable, i.e. gender, when training the algorithm. Therefore, including the binary variable of being a single parent in the model makes the algorithm counterfactually unfair.

2.3 Choosing a fairness metric

When it comes to choosing a proper metric to consider, as there are usually errors, which may cause more harm to protected groups than others, the authors of one of the tools facilitating detecting bias in the algorithm – "Aequitas" developed by the Center for Data Science and Public Policy at University of Chicago – propose the "Fairness tree" (Saleiro et al., 2019).

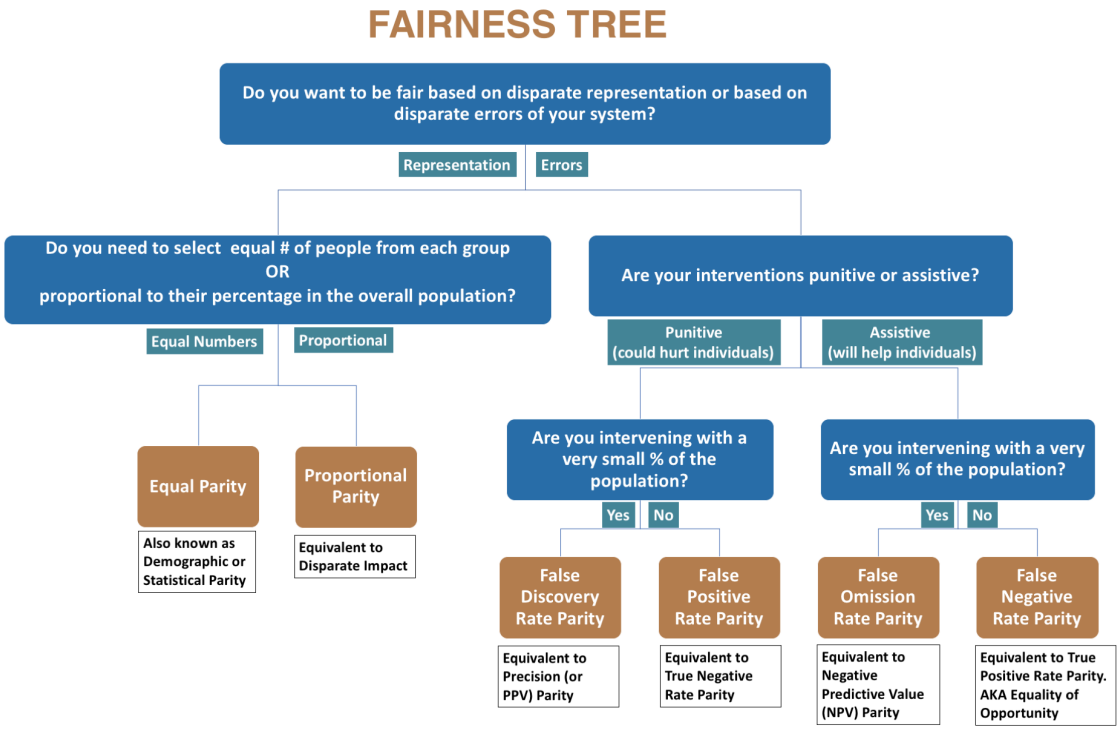

Figure 2.2: Fairness tree from Aequitas facilitating the choice of the right fairness metric suitable for a given problem.

Source: (Center for Data Science and Public Policy, University of Chicago, 2018b)

The "Fairness tree" (see Figure 2.2) allows us to choose the most appropriate metric by answering a few questions concerning the relevant problem. The first question asks whether we care about fair representation of each group or the errors our algorithm

makes. If we follow the path of having a fair representation of each group, then we can choose from having an equal number of people from each group – Statistical Parity – or having an equal proportion of each group – Proportional Parity. Regarding the errors of the algorithm, the interventions related to the algorithm can be punitive – harmful for individuals – or assistive – helpful for individuals. In the case of punitive interventions, we should choose False Discovery Rate Parity for interventions with a small percentage of the population, or False Positive Rate Parity otherwise. In the case of assistive interventions, we should choose False Omission Rate Parity for interventions with a small percentage of the population, or False Negative Rate Parity otherwise.

Example. In the example of treatment subsidy, following the “Fairness tree”, we can decide that we want to focus on the mistakes of our model. Then, the intervention of providing the subsidy is assistive, because it significantly helps patients to fund their treatment. Finally, most probably the subsidy would apply only to a small percentage of patients, so the metric we should mostly focus on is the False Omission Rate Parity.

Chapter 3

COMPAS Algorithm and Dataset

The COMPAS dataset is going to be used in further analysis to test and compare chosen open-source tools of bias detection. The main focus will not be to audit this algorithm, as it has already been proven to be strongly biased by ProPublica, but to review the tools available to the companies and organizations to facilitate and encourage them to audit their algorithms before putting them into production.

The description of the dataset given below includes information necessary to fully understand the data and possible grounds for discrimination. Moreover, ProPublica has also conducted an advanced cleaning of the dataset, but only the main steps are summarized here because it is not necessary to know all the details for a complete understating of subsequent analysis.

3.1 Dataset description

The data used by ProPublica comes from Broward County, Florida, and includes information for over 10,000 criminal defendants (Larson et al., 2016). The defendants usually respond to the COMPAS questionnaire when they are admitted to prison. Then, based on the collected information, the algorithm determines several risk scores, such as “Risk of Recidivism” or “Risk of Violent Recidivism”. To assess the performance of the algorithm, the score given to a defendant is then compared to the real situation, i.e.

whether this defendant recidivated during two years after getting the score.

The original dataset comprised of 18,610 people scored in 2013 and 2014. However, only entries for defendants scored at the pretrial stage were kept (discarding cases at parole, probation, and other). Each defendant was given a decile score ranging from 1 to 10, and the scores were converted into labels in the following way: score 1-4 – “Low” risk, score 5-7 – “Medium” risk, score 8-10 – “High” risk. The race classification variable included categories African-American, Caucasian, Hispanic, Asian, Native American, Other. The dataset also contains defendants’ personal information such as full name, gender (male and female), date of birth, exact age, and age group (less than 25, 25-45, greater than 45). Furthermore, standardized definitions of recidivism and violent recidivism were defined. The definitions were based on Northpointe’s practitioners’ guide indicating that the recidivism score predicts “a new misdemeanor or felony offense within two years of the COMPAS administration date.” (Northpointe Inc., 2015). Therefore, the two-year follow-up period was taken into consideration. Additionally, Northpointe’s practitioners’ guide states that “Scores in the medium and high range garner more interest from supervision agencies than low scores, as a low score would suggest there is little risk of general recidivism.” (Northpointe Inc., 2015). Therefore, having a “Medium” or “High” score translates to being at risk of recidivism. At the end of the process of cleaning the data and matching defendants’ COMPAS scores with their criminal records, two datasets were created – compas-scores-two-years with 7,214 entries with complete information about overall recidivism risk and compas-scores-two-years-violent with 4,743 entries with complete information about violent recidivism risk.

3.2 Algorithm performance

The algorithm’s performance described below is based on ProPublica’s analysis (Larson et al., 2016). The accuracy of the predictions for “Risk of Recidivism” was 61%, but for “Risk of Violent Recidivism” it was only 20%. In general, the accuracy of

the model regarding the “Risk of Recidivism” indicator was similar for black and white defendants (63% vs. 59%). However, the model makes different mistakes for each group. The False Positive Rate was almost twice as high for black defendants as for white defendants (45% vs. 23%). In other words, black defendants were much more often misclassified as higher risk than white defendants. On the other hand, white defendants were more often misclassified as low risk, while they re-offended, than black defendants. The False Negative Rate amounted to 48% for white defendants, and 28% for black defendants. Similar inequalities were found for the “Risk of Violent Recidivism” indicator.

Despite all the inequalities found in the COMPAS algorithm, Northpointe states that they paid attention to the algorithm’s fairness, and it is racially neutral. But the fairness metric they took into consideration is the Overall Accuracy Equality, and it is in fact similar for both black and white defendants (accuracy around 60%), so according to this metric, the COMPAS algorithm is fair (Angwin & Larson, 2016). However, if we consider different metrics and the types of errors the algorithm makes, it becomes unfair. It shows the paradox that an algorithm can be at the same time fair and unfair. It is usually impossible to satisfy all the metrics at once, and there has to be a compromise between different metrics. The question of which metric should be the priority is, therefore, really tough, and should be well-judged. Arguably, it is also worth to review a wider spectrum of metrics to make sure that there is no extreme discrimination hidden in other categories.

3.3 Exploratory analysis

Because the main aim of the analysis is to explore different bias detection tools and compare them, not to analyze the COMPAS dataset in great detail, only the dataset with the complete information about overall recidivism risk – compas-scores-two-years – will be used in further analysis. The exploratory analysis gives an overview of this dataset to better understand the distributions of different variables.

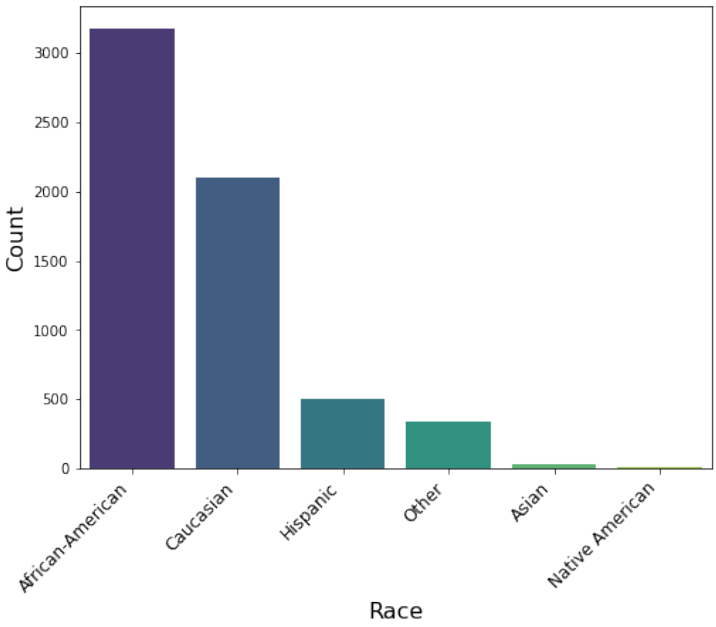

Figure 3.1: Distribution of race.

Source: Own analysis based on the COMPAS algorithm dataset.

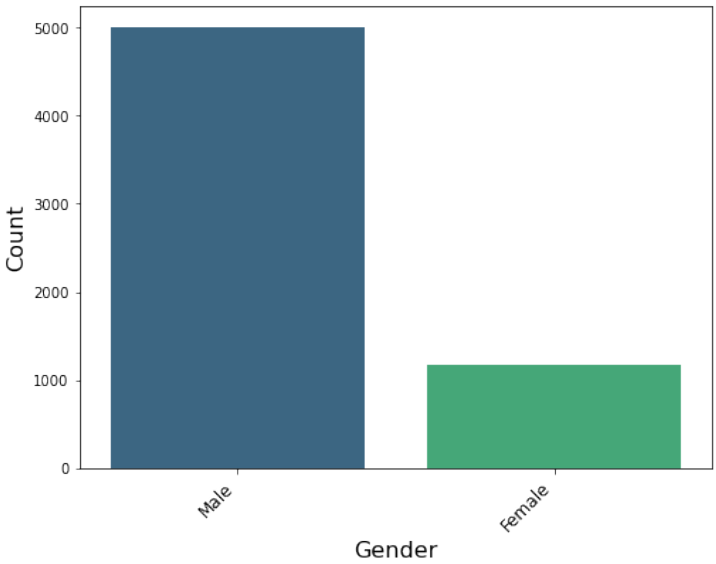

Figure 3.2: Distribution of gender.

Source: Own analysis based on the COMPAS algorithm dataset.

In the dataset, there are 6,172 valid observations. Regarding race, the biggest group (3,175 individuals) of defendants has African-American origins, and the second biggest group (2,103 individuals) is Caucasian (see Figure 3.1). The remaining race groups are rather small (Hispanic – 509, Other – 343, Asian – 31, Native American

– 11), so they are not going to be the center of attention when comparing different fairness metrics, because the results may not be representative. Moreover, most of the defendants are male (4,995 individuals) and there are 1,175 female defendants (see Figure 3.2).

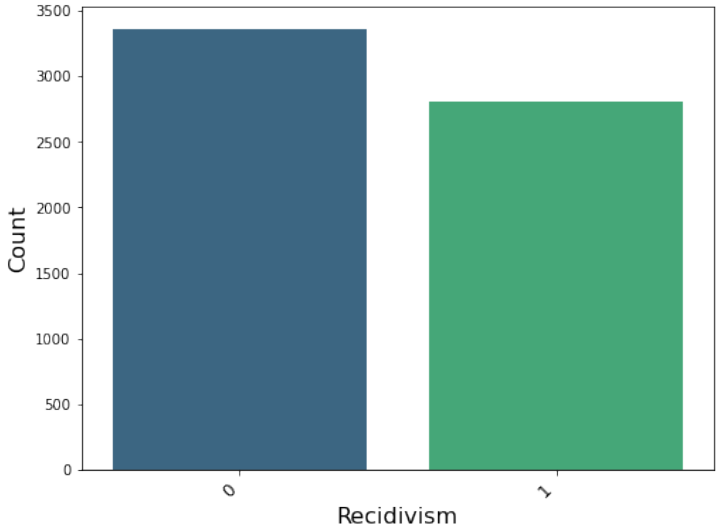

Figure 3.3: Distribution of observed recidivism within two years after receiving the COMPAS score.

Source: Own analysis based on the COMPAS algorithm dataset.

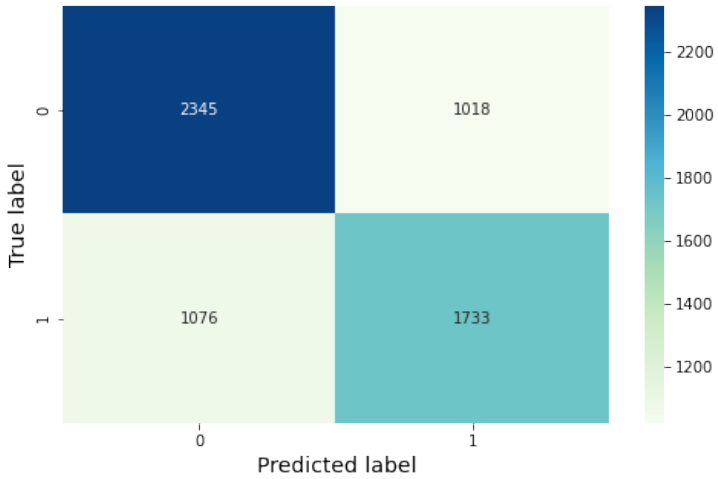

Figure 3.4: Confusion matrix of the risk of recidivism predictions.

Source: Own analysis based on the COMPAS algorithm dataset.

It can be observed that in general, more defendants do not commit a crime again (see Figure 3.3). The confusion matrix in Figure 3.4 indicates that the algorithm makes roughly the same number of mistakes for False Negative and False Positive predictions.

3.3.1 Distribution by race

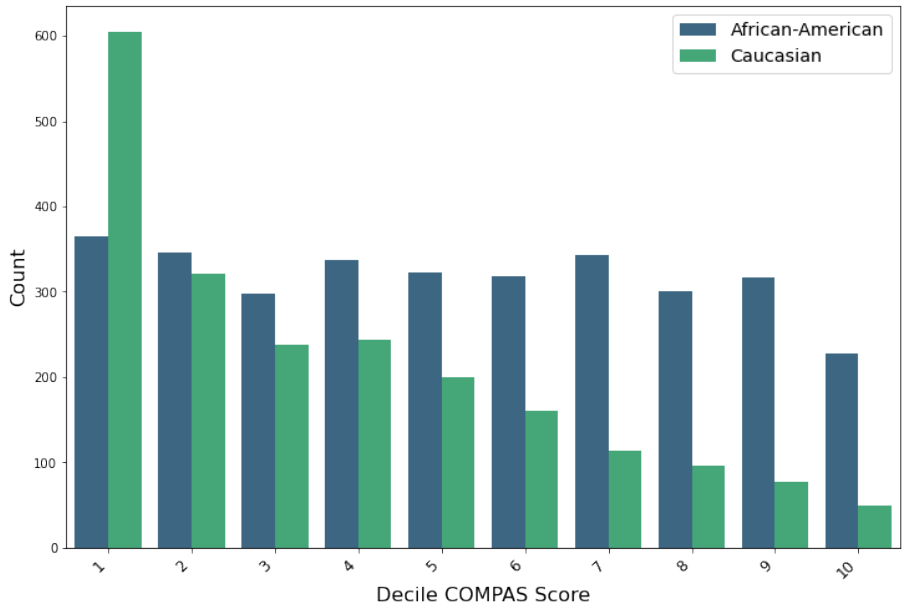

Figure 3.5: Distribution of decile COMPAS scores by race.

Source: Own analysis based on the COMPAS algorithm dataset.

It can be noticed that the decile scores generated by the COMPAS algorithm are perceptibly decreasing for white defendants, and there is only a small group of individuals with the highest scores of risk of recidivism (see Figure 3.5). Meanwhile, for black defendants, the scores are quite evenly distributed, so there are many more defendants classified as “Medium” and “High” risk scores (decile scores from 5 to 10).

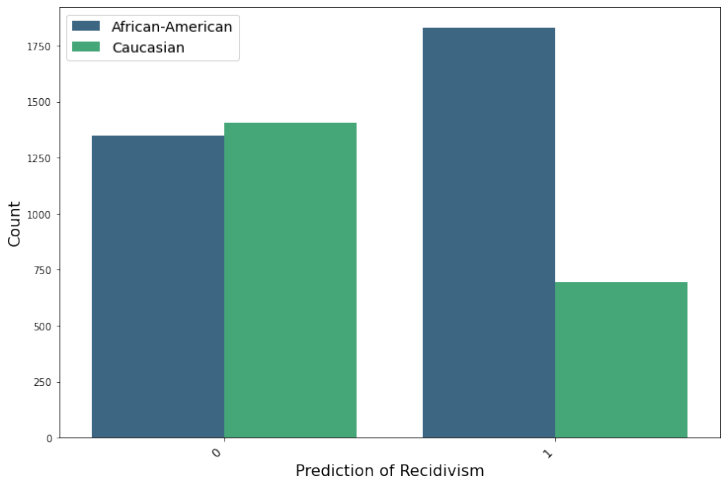

Figure 3.6: Distribution of predicted risk of recidivism by race.

Source: Own analysis based on the COMPAS algorithm dataset.

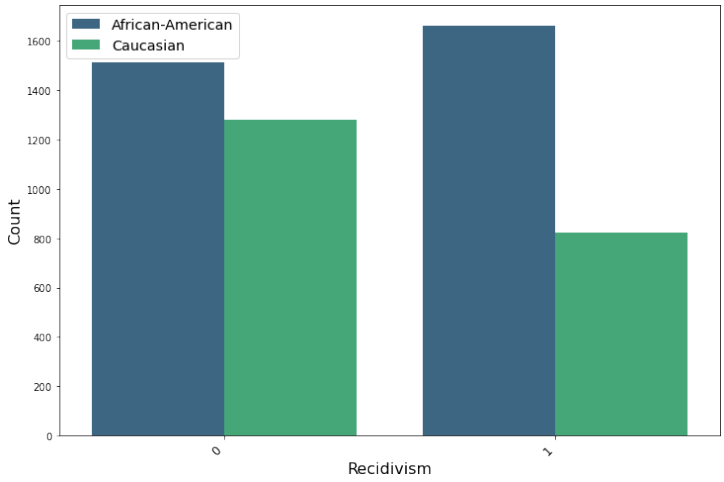

Figure 3.7: Distribution of observed recidivism within two years after receiving the COMPAS score by race.

Source: Own analysis based on the COMPAS algorithm dataset.

According to Northpointe's indications, defendants classified as "Medium" and "High" risk of recidivism are among those suspected of committing a crime again. As can be seen in Figure 3.7 black defendants, in reality, recidivate more often than white defendants, however, the algorithm visibly overshoots the predicted number of black defendants being recidivists. Thus, it can raise doubts about the fairness of this algorithm and it should be tested in subsequent analysis.

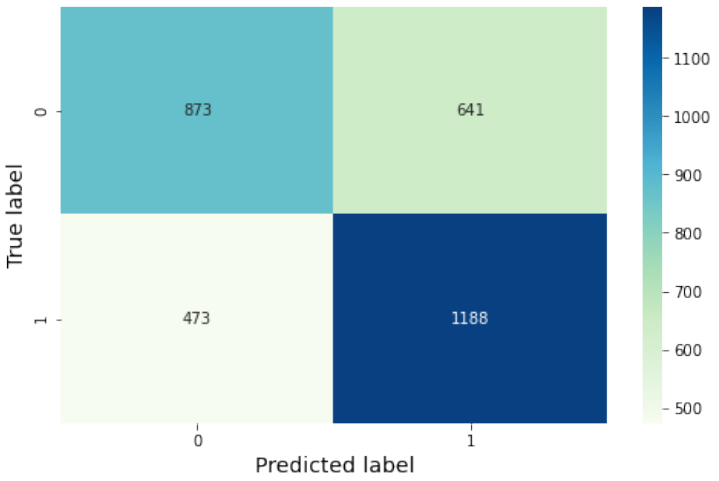

Figure 3.8: Confusion matrix for African-American defendants.

Source: Own analysis based on the COMPAS algorithm dataset.

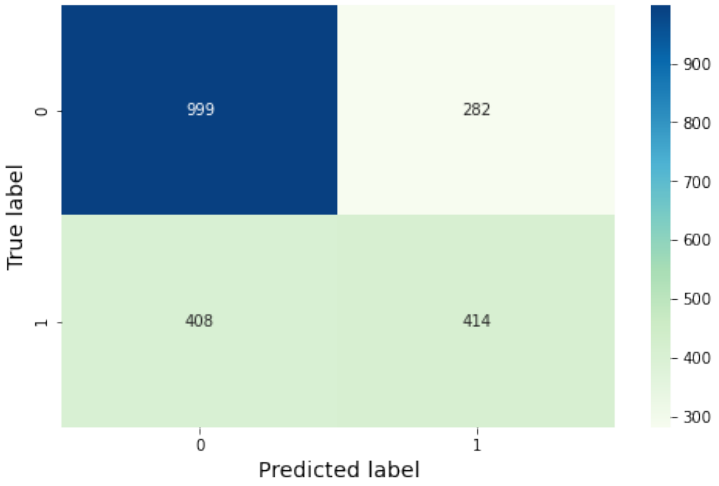

Figure 3.9: Confusion matrix for Caucasian defendants.

Source: Own analysis based on the COMPAS algorithm dataset.

The confusion matrices for African-American and Caucasian defendants confirm the observation that the algorithm makes different types of mistakes for black and white defendants. There are more false positives than false negatives for black defendants (see Figure 3.8) and more false negatives than false positives for white defendants (see Figure 3.9).

3.3.2 Distribution by gender

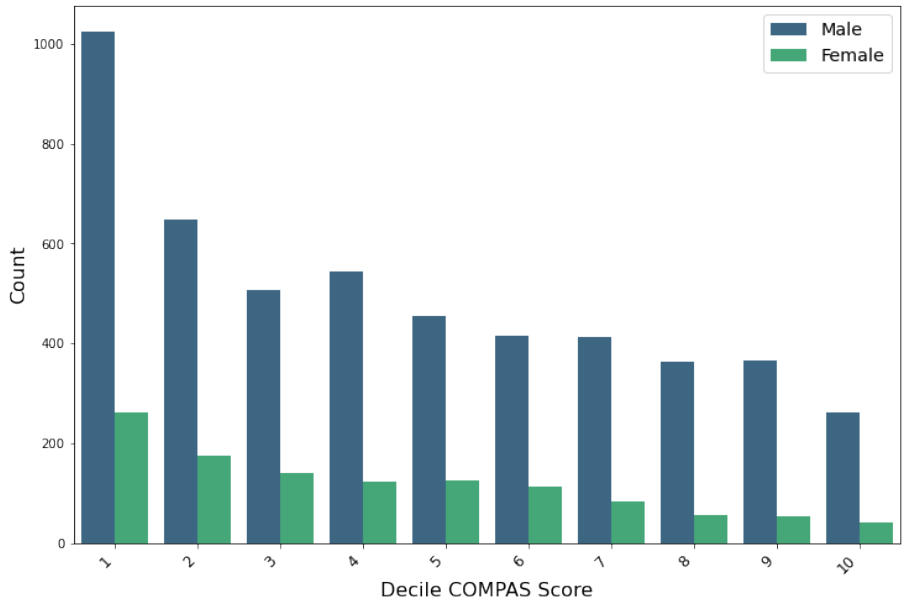

Figure 3.10: Distribution of decile COMPAS scores by gender.

Source: Own analysis based on the COMPAS algorithm dataset.

When it comes to the distributions by gender, the decile scores for the risk of recidivism seem to be more just than for race. Both for men and women the decile scores are decreasing, and there is a relatively small number of people having the highest risk scores (see Figure 3.10).

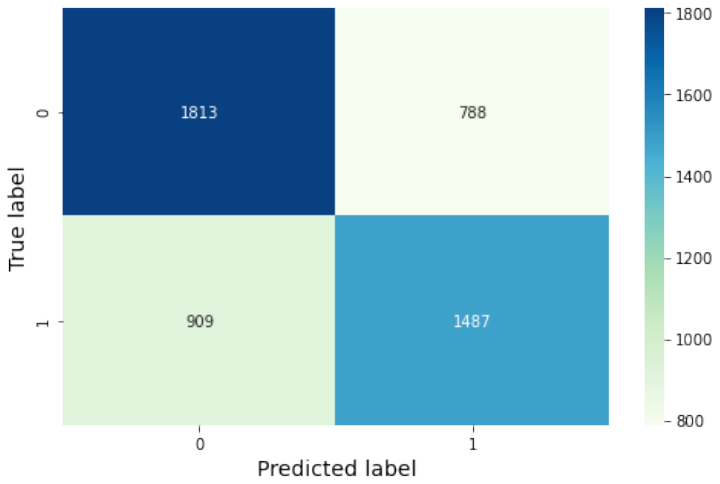

Figure 3.11: Confusion matrix for male defendants.

Source: Own analysis based on the COMPAS algorithm dataset.

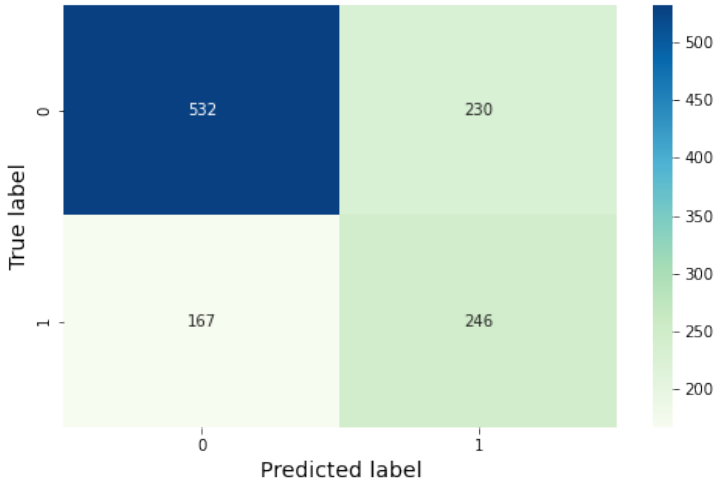

Figure 3.12: Confusion matrix for female defendants.

Source: Own analysis based on the COMPAS algorithm dataset.

In general, there are far fewer women than men in the study group. Deducing from Figures 3.11 and 3.12, the majority of defendants are correctly predicted as not being at risk of recidivism. However, it can be also noticed that for women there are more cases predicted as False Positive – predicted as being at risk of recidivism but did not recidivate, while for men there are more cases predicted as False Negative – predicted as not being at risk of recidivism but in reality recidivated. Therefore, it should be also examined whether the algorithm fairly classifies both genders or not.

The above exploratory analysis, via simple visualizations, gives a clue about the potential areas where discrimination and bias can be present and indicates the focus of the next steps. It also helps to better understand the sample of the population that we are dealing with. It shows that the errors of the model for protected and unprotected groups should be examined in more detail because they seem to arouse the most suspicious. As we already know thanks to ProPublica's analysis, this is where the bias was detected, and now we can visually see it ourselves.

Chapter 4

Bias Detection Tools

There are several tools suggested online, which enable bias detection. Some of them are open-source, while others offer only part of it for free or a demo version. These tools include:

- Aequitas
- What-If Tool
- Pymetrics: audit-AI
- IBM® AI Fairness 360

The Aequitas is an open-source bias audit toolkit, which was created by the Center of Data Science and Public Policy at the University of Chicago (Center for Data Science and Public Policy, University of Chicago, 2020a). It offers a tool to audit machine learning models for discrimination and bias. It provides three different ways to conduct the audits: Web Audit Tool, Python library, and Command line tool. Web Audit Tool can be used directly in the browser without writing any code, which is very quick and convenient. Python library and Command line tool give more options for customization and creating tailor-made reports. Aequitas can automatically generate a bias report for our algorithm, it generates bias and fairness statistics, and enables us to create interactive graphs and dashboards. However, the limitation of the Aequitas tool is that it can only be used for binary classification problems.

The What-If Tool is Google's bias detection and machine learning diagnostic tool. It has been developed by researchers and designers at Google's PAIR – People and AI Research. It allows us to test 5 types of fairness: group unaware, group threshold, demographic parity, equal opportunity, equal accuracy (Weinberger, n.d.). It can be used for binary classification problems, multi-class classification problems, and regression models. This tool can be also used just to analyze a dataset without the model's predictions. It can be used within Jupyter notebooks, Colab notebooks, Google Cloud AI Platform, and on Tensorboard. Additionally, it helps to identify counterfactuals and provides an overview of features distributions (People + AI Research (PAIR), n.d.).

Pymetrics provides an AI platform dedicated to ethical and human-centered hiring and talent acquisition processes (Pymetrics, 2020). They guarantee the gender and ethnic fairness of their algorithms. They open-sourced their bias testing tool called audit-AI. audit-AI is a Python library using pandas and sklearn libraries, which compares differences in algorithm outputs among protected and unprotected groups. However, the audit-AI library lacks structured documentation, so it is difficult to find the complete set of available functions and understand their use.

IBM® AI Fairness 360 helps to examine, report, and mitigate bias and discrimination in algorithms (IBM Research, 2018). It offers five different bias metrics: statistical parity difference, equal opportunity difference, average odds difference, disparate impact, and Theil index. It compares each of these metrics for the privileged group and unprivileged groups showing whether they are in an acceptable range or not. Additionally, this tool has not only been implemented to detect bias but also to mitigate it. It offers four different methods of bias mitigation and at the end allows us to compare the models before and after mitigation. However, the mitigation of bias in machine learning models is out of the scope of this thesis, so this tool is not going to be presented in detail, as it offers similar possibilities for bias detection as the tools mentioned above.

4.1 Aequitas

First of all, a web report was generated for the COMPAS algorithm dataset using the Aequitas bias detection tool. The audit was conducted to detect potential discrimination against race and gender. The process of conducting the Bias Audit consists of 4 steps presented in Figure 4.1.

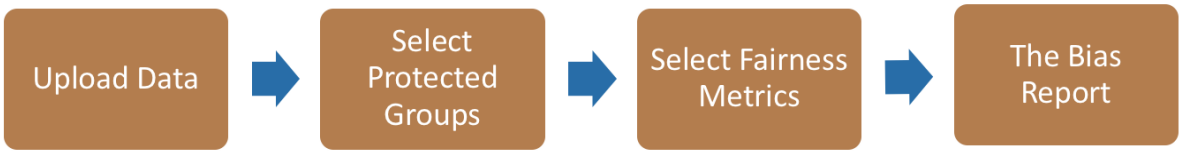

Figure 4.1: The flow of the Bias Audit by Aequitas.

Source: (Center for Data Science and Public Policy, University of Chicago, 2018a)

The first step is to upload the data. The dataset should be uploaded in a CSV format, and should contain the following columns:

- score – the binary (0 and 1) assessment of the predictive model, where 1 indicates the individuals selected for the intervention;
- label_value – the true binary outcomes (0 and 1) for each individual, necessary to examine the errors of the model;
- attributes – the attributes that are going to be audited for bias containing the information about protected and unprotected groups, such as gender, race, age, ethnic group, etc.; these columns can be categorical or continuous.

Afterward, the reference groups should be selected for the audited attributes. Fairness metrics of other groups are then compared to the reference group we select. We can manually choose the reference groups, for example, Male for gender attribute and Caucasian for race attribute in the COMPAS algorithm case. However, the reference group can be also chosen automatically by selecting the majority groups for each attribute or the group with the lowest bias metric. After choosing the reference groups, the fairness metrics should be chosen as well. The Aequitas bias detection tool offers six

different metrics: Equal Parity, Proportional Parity, False Positive Rate Parity, False Discovery Rate Parity, False Negative Rate Parity, and False Omission Rate Parity. Additionally, there is the “Fairness tree” (see Figure 2.2) included facilitating the choice of the right metric. According to the “Fairness tree”, the best fairness metric for the COMPAS algorithm audit is the False Positive Rate Parity because the interventions have a punitive character and involve a substantial proportion of individuals. However, for illustrative purposes also other metrics are going to be examined. The disparity intolerance in percentages should also be entered. It is a specific threshold used for determining whether an algorithm should pass or fail the audit. By default, it is set to 80%, and in our case, it will be left as default. The audit will pass when bias metrics for other groups remain within the specified percentage of the reference group, i.e. in the case of the threshold of 80%, the metrics should be withing 80% and 125% to pass the audit.

Then, the audit is ready to be generated. It contains the following information: summary, details by fairness measures, details by protected attributes, bias metrics values, and base metrics calculated for each group. The complete bias report can be found in Appendix A.

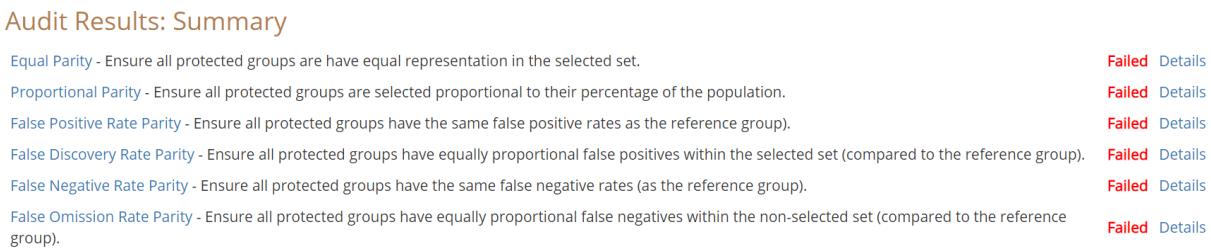

Figure 4.2: Aequitas - The Bias Report - Audit Results: Summary.

Source: (Center for Data Science and Public Policy, University of Chicago, 2020b)

In the summary of audit results, the general information can be found whether the audit passed according to each metric and a short definition of each metric. As can be seen in Figure 4.2, the COMPAS algorithm failed the audit for all of the metrics. Then, the details of each metric are presented.

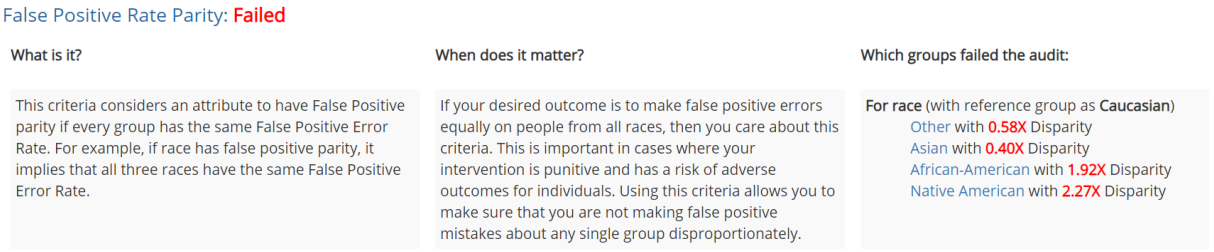

Figure 4.3: Aequitas - The Bias Report - Audit Results: Details by Fairness Measures - False Positive Rate Parity.

Source: (Center for Data Science and Public Policy, University of Chicago, 2020b)

As the priority metric in our case is the False Positive Rate Parity, the details of this metric are presented in Figure 4.3. In the details, we can see the more comprehensive definition of the metric, when this metric matters the most, and which groups failed the audit. In our case, in reference to the Caucasian group, African-American, Native American, Asian, and Other groups failed the audit. The African-American group is the most important for us because the rest of the groups are small in number, so they may not be representative. When it comes to the comparison of African-American and Caucasian defendants, black individuals are almost twice as likely to be misclassified by the algorithm as False Positives. On the other hand, according to this metric, the algorithm passes the audit for gender groups because they are not specified in the table. The rest of the results for all metrics with their comprehensive definitions can be found in Appendix A.

Attribute Value	Predicted Positive Rate Disparity	Predicted Positive Group Rate Disparity	False Discovery Rate Disparity	False Positive Rate Disparity	False Omission Rate Disparity	False Negative Rate Disparity
African-American	2.63	1.74	0.86	1.92	1.21	0.57
Asian	0.01	0.68	0.71	0.4	0.43	0.76
Caucasian	1.0	1.0	1.0	1.0	1.0	1.0
Hispanic	0.2	0.84	1.09	0.88	1.03	1.17
Native American	0.01	2.2	0.93	2.27	0.0	0.0
Other	0.1	0.62	0.99	0.58	1.04	1.33

Figure 4.4: Aequitas - The Bias Report - Audit Results: Bias Metrics Values - race.

Source: (Center for Data Science and Public Policy, University of Chicago, 2020b)

In the Bias Metrics Values section, we can find the ratio of each bias metric for every group compared to the bias metric for the reference group – Caucasian (see Figure

4.4). All values between 0.8 and 1.25 are marked green, which means they are within the acceptable range to pass the audit. If the values are outside this range, they are marked red and lead to the failed audit. Looking at the African-American defendants, only two metrics are within the acceptable range – False Discovery Rate Parity and False Omission Rate Parity. The most balanced group in comparison with the Caucasian is the Hispanic group, which passes the audit for almost all metrics.

Attribute Value	Predicted Positive Rate Disparity	Predicted Positive Group Rate Disparity	False Discovery Rate Disparity	False Positive Rate Disparity	False Omission Rate Disparity	False Negative Rate Disparity
Female	0.21	0.89	1.4	1.0	0.72	1.07
Male	1.0	1.0	1.0	1.0	1.0	1.0

Figure 4.5: Aequitas - The Bias Report - Audit Results: Bias Metrics Values - gender. Source: (Center for Data Science and Public Policy, University of Chicago, 2020b)

As far as gender is concerned, it can be seen that the algorithm passes the audit for half of the metrics (see Figure 4.5). The audit is failed for Predicted Positive Rate Parity, False Discovery Rate Parity, and False Omission Rate Parity. However, they are not the priority metrics in this case and the ratios are less extreme than for race groups, so the algorithm is more equitable for gender than for race.

Attribute Value	Group Size Ratio	Predicted Positive Rate	Predicted Positive Group Rate	False Discovery Rate	False Positive Rate	False Omission Rate	False Negative Rate
African-American	0.51	0.66	0.58	0.35	0.42	0.35	0.28
Asian	0.01	0.0	0.23	0.29	0.09	0.12	0.38
Caucasian	0.34	0.25	0.33	0.41	0.22	0.29	0.5
Hispanic	0.08	0.05	0.28	0.44	0.19	0.3	0.58
Native American	0	0.0	0.73	0.38	0.5	0.0	0.0
Other	0.06	0.03	0.2	0.4	0.13	0.3	0.66

Figure 4.6: Aequitas - The Bias Report - Audit Results: Group Metrics Values - race. Source: (Center for Data Science and Public Policy, University of Chicago, 2020b)

Attribute Value	Group Size Ratio	Predicted Positive Rate	Predicted Positive Group Rate	False Discovery Rate	False Positive Rate	False Omission Rate	False Negative Rate
Female	0.19	0.17	0.41	0.48	0.3	0.24	0.4
Male	0.81	0.83	0.46	0.35	0.3	0.33	0.38

Figure 4.7: Aequitas - The Bias Report - Audit Results: Group Metrics Values - gender. Source: (Center for Data Science and Public Policy, University of Chicago, 2020b)

In the Group Metrics Values section, the exact values (not the ratio) of all metrics are presented for all groups. Again, these values can be marked green or red to indicate whether they pass or fail the audit. What is important, in this instance the performance of the model is also visible, not only the bias itself. The error rates and group rates can be interpreted. It can be seen that Asian, Hispanic, Native American, and Other groups represent a very small percentage of the population of defendants (see Figure 4.6), as well as female defendants representing only 19% of the population (see Figure 4.7). Regarding the False Positive Rate for different race groups, it ranges from 0.09 for Asian defendants to 0.5 for Native Americans, which means that in the best-case scenario only 9% of Asian non-recidivists were wrongly classified by the algorithm as recidivists, while in the worst-case scenario as much as 50% of Native American non-recidivists were mistakenly classified by the algorithm as recidivists. This metric is also very high for African-Americans, for which the algorithm mistakenly classifies 42% of non-recidivists as recidivists.

The Aequitas tool provides also the Python library which also allows us to quickly calculate all the group and bias metrics presented above. Furthermore, it offers the possibility to plot different metrics in a visually attractive way. The graphs convey similar information as the tables presented above and they can be found in Appendix B.

4.2 What-If Tool

As distinct from the Aequitas bias detection tool, What-If Tool requires not only the binary predictions of the algorithm, but also the model itself, and the accurate model's predictions, i.e. the probability of belonging to a specific class. Additionally, What-If Tool can be used on a dataset without a model to conduct an exploratory analysis of variables distributions and basic statistics. The What-If Tool is very elaborate, so only chosen parts are going to be presented.

A model used to predict the recidivism score is necessary to conduct the audit with What-If Tool, so, for ease of implementation, an algorithm created by Google to imitate

the one used by Northpointe is going to be used (Google LLC, 2019). It is important to bear in mind that it is not the model originally used by Northpointe, nevertheless, it gives a good overview of how the report could look like for the original model.

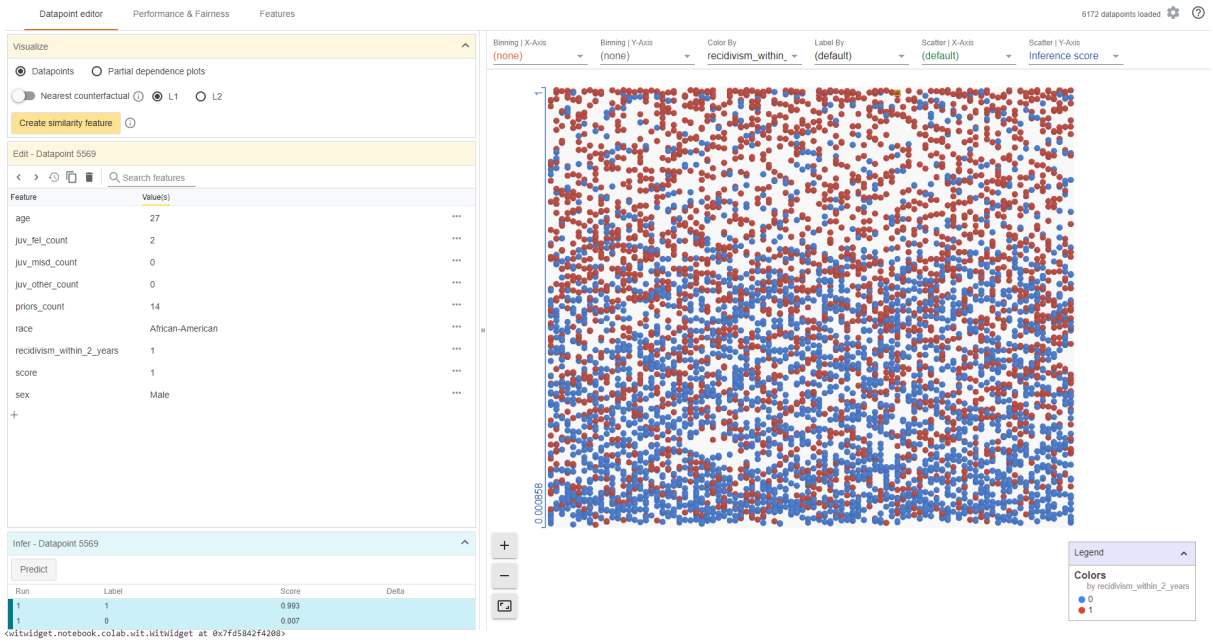

Figure 4.8: What-If Tool - Datapoint editor - Overview and the scatterplot of inference score (the probability of belonging to the positive class) and the observed recidivism within two years.

Source: Own analysis with What-If Tool Python library based on the COMPAS algorithm dataset and the model developed by Google imitating the original algorithm (Google LLC, 2019).

As can be seen in Figure 4.8, the What-If Tool provides an overview of all data points in the dataset and generates a scatterplot which can be individually adjusted to represent different variables. By clicking on a chosen data point, it provides detailed information about the selected individual and the model's prediction for this individual. The scatterplot in Figure 4.8 shows that the predictions of the model are not very accurate because the red points (who recidivated within two years) should be placed at the top of the graph (high inference score, i.e. the probability of belonging to the positive class), while the blue points should be mostly located at the bottom of the graph. Meanwhile, the different points are rather mixed.

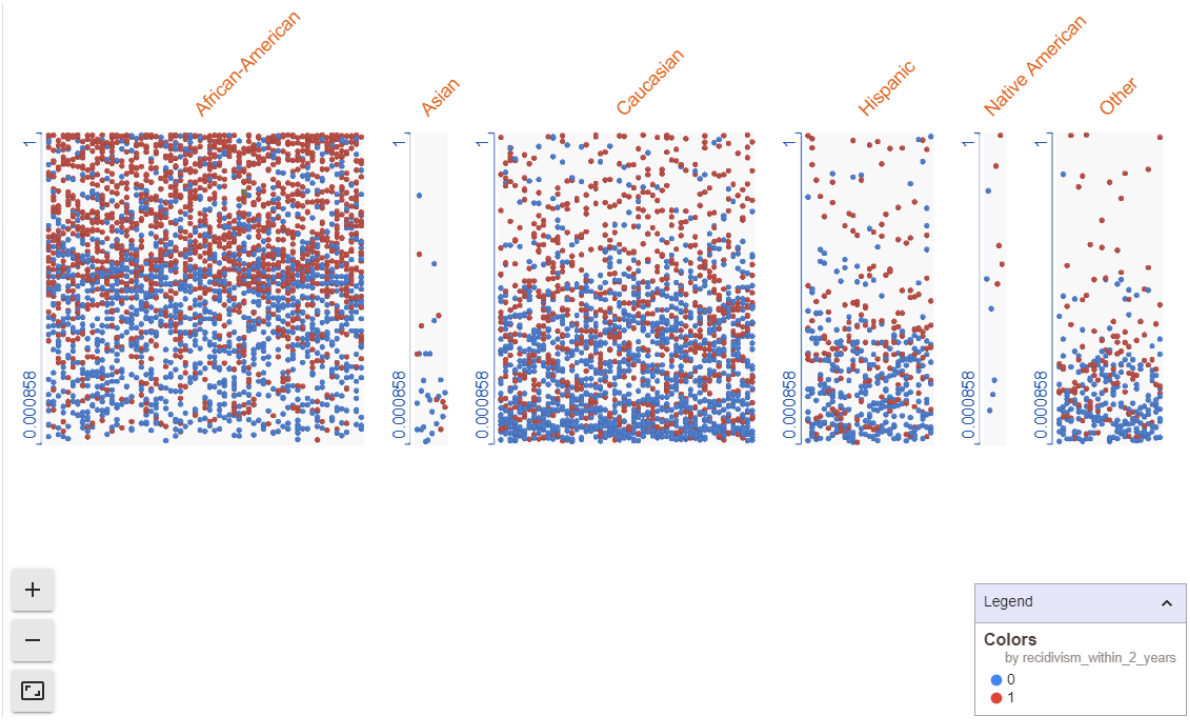

Figure 4.9: What-If Tool - the scatterplot of inference score and the observed recidivism within two years by race.

Source: Own analysis with What-If Tool Python library based on the COMPAS algorithm dataset and the model developed by Google imitating the original algorithm (Google LLC, 2019).

As shown in Figure 4.9, the same scatterplot as in the previous example is presented separately for each race group. It simultaneously allows us to assess the size of each group and the predictions of the model compared to the true values. It can be observed that in the African-American group there are relatively many points located in the upper part of the graph with a higher probability of recidivating. Meanwhile, for the Caucasian, Hispanic, and Other groups the majority of points are located in the lower part of the graph.

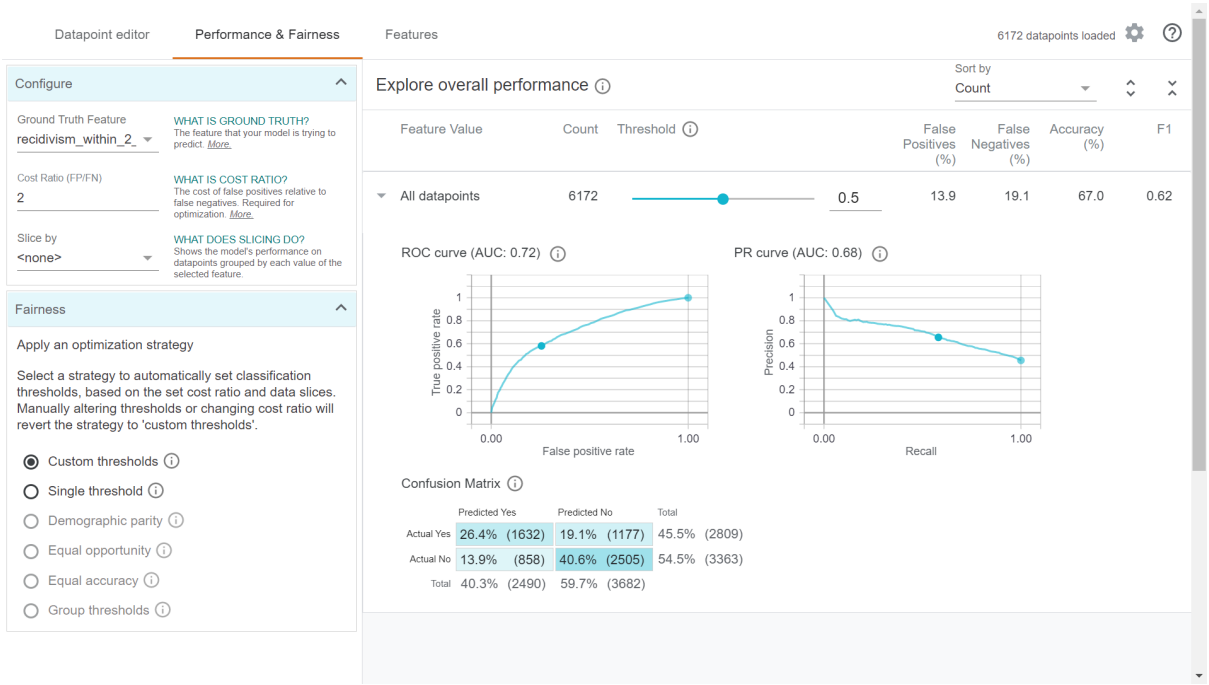

Figure 4.10: What-If Tool - Performance & Fairness - the overall performance of the model and configuration settings.

Source: Own analysis with What-If Tool Python library based on the COMPAS algorithm dataset and the model developed by Google imitating the original algorithm (Google LLC, 2019).

The Performance & Fairness tab of the What-If Tool includes information about chosen fairness and performance metrics (see Figure 4.10). The ground truth feature has to be chosen, which shows the real observed values of the predicted variable. Moreover, the cost ratio (which is the ratio of the cost associated with False Positive to False Negative) can be chosen to optimize the model's parameters. It is especially useful because, for instance, in our case, mistakenly allocating an individual as the false positive can be more harmful to them than being classified as the false negative, therefore, the cost ratio should be set to more than 1. It would be the opposite (the cost ratio lower than 1) in the case of algorithms with assistive interventions, such as the example of treatment subsidy. Additionally, graphs of the ROC curve and PR curve, as well as the confusion matrix, are displayed for the model.

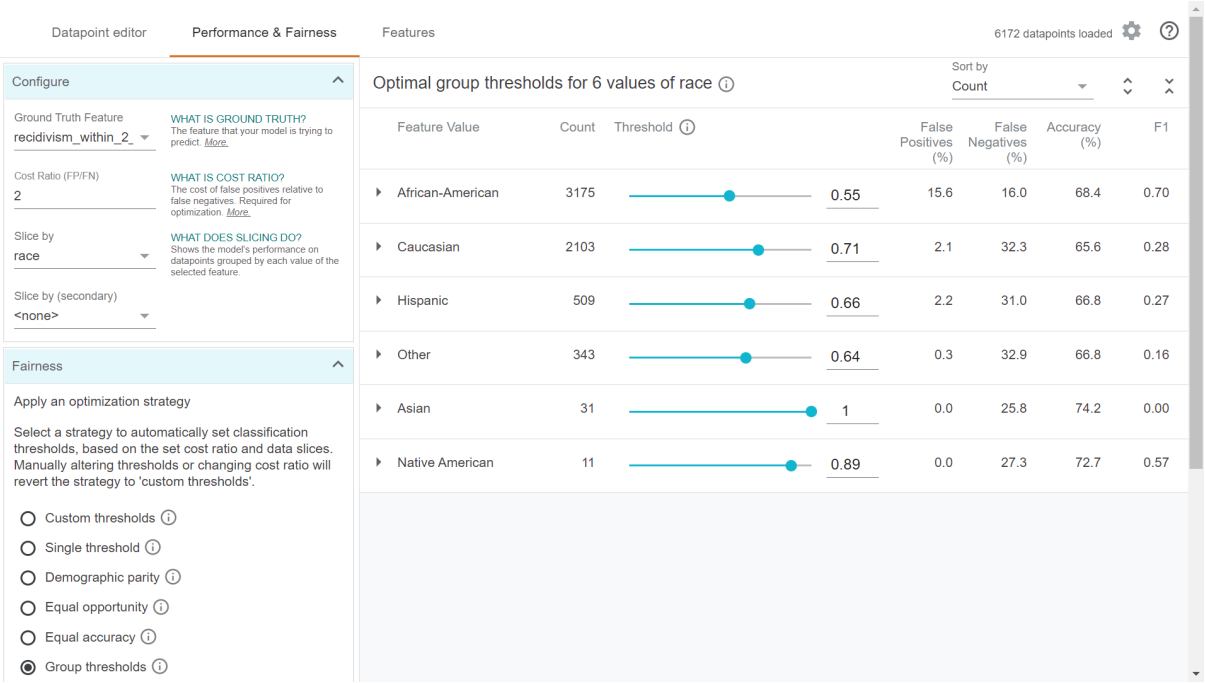

Figure 4.11: What-If Tool - Performance & Fairness - fairness metrics by race groups for Group thresholds optimization strategy.

Source: Own analysis with What-If Tool Python library based on the COMPAS algorithm dataset and the model developed by Google imitating the original algorithm (Google LLC, 2019).

Finally, if we slice our data by a selected group, i.e. race, fairness metrics for all groups are presented (see Figure 4.11). It is also possible to choose an optimization strategy, and then automatically the thresholds of classifying a case to the positive class are changed for each group. Some of the previously described fairness metrics are available to be chosen for the optimization, such as demographic parity, equal opportunity, or equal accuracy. Furthermore, custom thresholds can be applied or a single threshold for all groups, the group thresholds strategy optimizes the thresholds according to the chosen cost ratio. This section of the What-If Tool tool enables us to experiment with different thresholds and see instantly how specific metrics are changing. What can be inferred from such experiments with the thresholds is that it is impossible to satisfy all of the metrics at once. While some of them are being equalized, the others become significantly different for some groups. It emphasizes the difficulty of choosing the right metrics to optimize a machine learning model because there is a tradeoff between them.

Conclusions

This thesis aimed to understand the concept of bias and fairness of algorithms, as well as, to review open-source tools available online to facilitate bias detection and encourage companies and individuals to audit their algorithms. As it turned out, the fairness of algorithms can be defined in many different ways, and usually, an algorithm cannot satisfy all of the fairness metrics simultaneously. Therefore, the concept of fairness does not come down to the mathematical formulas, but rather to an ethical question. There are some guidelines and suggestions when which metrics should be prioritized, however, these simple mathematical metrics always involve the fate of human beings, so I hope that by studying the provided examples readers will understand the importance of, firstly, auditing algorithms for bias, and secondly the in-depth discussion about choosing a fairness metric and how it impacts human lives.

Moreover, the analysis of various open-source audit tools based on the COMPAS algorithm shows that there are already available user-friendly tools that significantly help in auditing the machine learning models. Providing such tools is crucial to encourage companies and individuals to conduct such audits because very few companies want to spend a huge amount of time and effort on additional unnecessary audits. Meanwhile, the tools presented in this paper quickly provide the results and facilitate the understanding of all different metrics. The examples and reports generated in Chapter 4, as well as the exemplary notebooks and web reports shared online by the authors of these tools, should hopefully encourage more and more people to try and use them in their everyday work and to contribute to the promotion of auditing machine learning models for bias and discrimination.

Naturally, the analysis conducted in this thesis is not exhaustive. The first intention for this research was to conduct a bias audit on a new dataset that has not been audited yet. However, access to such datasets and algorithms, which often come from sensitive industries, is very limited and proprietary. On this account, because of data unavailability, the open-access COMPAS algorithm dataset was used. Nonetheless, the analysis could be enriched by not only experimenting with different tools based on a well-known example but also auditing a new algorithm and increasing the transparency of algorithms that are commonly used. Furthermore, detecting bias is just the first step in the whole process of eliminating bias from algorithms. The extent of this thesis allowed me to focus only on this step, but the concept of mitigating bias, once detected, is another extremely important task. Some of the bias detection tools also contain a section dedicated to the task of bias mitigation, which has not been explored in this work. Additionally, to make the bias detection a matter of routine, a change in our way of thinking is required, as well as setting it as one of the core values within organizations. People need to understand the importance and the benefits of model monitoring and mitigation of bias, and it requires joint efforts of many people at different levels of organizations.

References

Angwin, J., & Larson, J. (2016, December 30). *Bias in criminal risk scores is mathematically inevitable, researchers say.* Retrieved September 21, 2020, from https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

Bertrand, M., & Mullainathan, S. (2003). Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *SSRN Electronic Journal*. doi: 10.2139/ssrn.422902

Cambridge Dictionary. (2020, September 2). *BIAS: Meaning in the Cambridge English Dictionary.* Retrieved September 8, 2020, from https://dictionary.cambridge.org/dictionary/english/bias

Carrns, A. (2019, January 18). *In California, gender can no longer be considered in setting car insurance rates.* Retrieved September 15, 2020, from https://www.nytimes.com/2019/01/18/your-money/car-insurance-gender-california.html

Center for Data Science and Public Policy, University of Chicago. (2018a). *Bias and fairness audit toolkit.* Retrieved October 1, 2020, from http://aequitas.dssg.io/

Center for Data Science and Public Policy, University of Chicago. (2018b). *Metrictree.png [Digital image].* Retrieved September 15, 2020, from http://aequitas.dssg.io/static/images/metrictree.png

Center for Data Science and Public Policy, University of Chicago. (2020a, February 12). *Aequitas.* Retrieved September 29, 2020, from http://www.datasciencepublicpolicy.org/projects/aequitas/

Center for Data Science and Public Policy, University of Chicago. (2020b, September 30). *Aequitas - the bias report.* Retrieved September 30, 2020, from http://aequitas.dssg.io/audit/v4_j1e_4/recid_aequitas/report-1.html#equal-parity-span-red-initfailedspan-red-end

Clarke, Y. D. (2019, April 11). *All information (except text) for H.R.2231 - Algorithmic Accountability Act of 2019.* Retrieved May 15, 2020, from https://www.congress.gov/bill/116th-congress/house-bill/2231/all-info

Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women.* Retrieved May 15, 2020, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G?utm_campaign=the_algorithm.unpaid.engagement&utm_source=hs_email&utm_medium=email&_hsenc=p2ANqtz-___QLmnG4HQ1A-IfP95UcTpIXuMGTCsRP6yF2OjyXHH-66cuuwpXO5teWKx1dOdk-xB0b9

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, *2015*(1), 92–112. doi: 10.1515/popets-2015-0007

Gershgorn, D. (2019, October 24). *Hospital algorithms are biased against black patients, new research shows.* Retrieved July 24, 2020, from https://onezero.medium.com/hospital-algorithms-are-biased-against-black-patients-new-research-shows-7ab4cc896fb3

Google LLC. (2019). *What-If Tool on COMPAS.* Retrieved October 1, 2020, from http://aequitas.dssg.io/

IBM Research. (2018). *AI Fairness 360.* Retrieved October 1, 2020, from https://aif360.mybluemix.net/

Kusner, M. J., Loftus, J. R., Russel, C., & Silva, R. (2020, March 8). *Counterfactual fairness.* Retrieved September 15, 2020, from https://arxiv.org/abs/1703.06856

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). *How we analyzed the COMPAS recidivism algorithm.* Retrieved July 24, 2020, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Mohajon, J. (2020, May 29). *Confusion matrix for binary classification. [Digital image].* Retrieved September 23, 2020, from https://miro.medium.com/max/1577/1*fxiTNIgOyvAombPJx5KGeA.png

Northpointe Inc. (2015, March 19). *Practitioner's guide to COMPAS core.* Retrieved September 16, 2020, from http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. doi: 10.1126/science.aax2342

O'Brien, R. L., & Kiviat, B. (2018). Disparate impact? Race, sex, and credit reports in hiring. *Socius: Sociological Research for a Dynamic World*, *4*. doi: 10.1177/2378023118770069

People + AI Research (PAIR). (n.d.). *What-If Tool - get started.* Retrieved September 29, 2020, from https://pair-code.github.io/what-if-tool/get-started/

People + AI Research (PAIR). (2020). *What-If Tool.* Retrieved September 27, 2020, from https://pair-code.github.io/what-if-tool/

ProPublica. (2016, May 22). *Propublica/compas-analysis.* Retrieved September 16, 2020, from https://github.com/propublica/compas-analysis

ProPublica. (2017, January 31). *Propublica - about us.* Retrieved May 15, 2020, from https://www.propublica.org/about/

Pymetrics. (2020, September 18). *pymetrics - solutions.* Retrieved September 29, 2020, from https://www.pymetrics.ai/solutions

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., … Ghani, R. (2019, April 29). *Aequitas: A bias and fairness audit toolkit.* Retrieved September 14, 2020, from https://arxiv.org/abs/1811.05577

Silberg, J., & Manyika, J. (2019, June 6). *Tackling bias in artificial intelligence (and in humans).* Retrieved May 15, 2020, from https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, *11*(3), 10–29. doi: 10.1145/2460276.2460278

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness - FairWare '18*. doi: 10.1145/3194770.3194776

Weinberger, D. (n.d.). *Playing with AI fairness.* Retrieved September 21, 2020, from https://pair-code.github.io/what-if-tool/ai-fairness.html

Appendix A

Aequitas - The Bias Report

Back to Customize (/audit/v4_j1e_4/recid_aequitas/) Home (/)

About (http://dsapp.uchicago.edu/aequitas)

Aequitas (/)
Bias & Fairness Audit

The Bias Report

Audit Date:	30 Sep 2020
Data Audited:	6172 rows
Attributes Audited:	race, gender
Audit Goal(s):	Equal Parity - Ensure all protected groups are have equal representation in the selected set. Proportional Parity - Ensure all protected groups are selected proportional to their percentage of the population. False Positive Rate Parity - Ensure all protected groups have the same false positive rates as the reference group). False Discovery Rate Parity - Ensure all protected groups have equally proportional false positives within the selected set (compared to the reference group). False Negative Rate Parity - Ensure all protected groups have the same false negative rates (as the reference group). False Omission Rate Parity - Ensure all protected groups have equally proportional false negatives within the non-selected set (compared to the reference group).
Reference Groups:	Custom group - The reference groups you selected for each attribute will be used to calculate relative disparities in this audit.
Fairness Threshold:	80%. If disparity for a group is within 80% and 125% of the value of the reference group on a group metric (e.g. False Positive Rate), this audit will pass.

Audit Results:

1. Summary
2. Details by Fairness Measures
3. Details by Protected Attributes

4. Bias Metrics Values
5. Base Metrics Calculated for Each Group

Audit Results: Summary

Equal Parity - Ensure all protected groups are have equal representation in the selected set.	**Failed**	Details
Proportional Parity - Ensure all protected groups are selected proportional to their percentage of the population.	**Failed**	Details
False Positive Rate Parity - Ensure all protected groups have the same false positive rates as the reference group).	**Failed**	Details
False Discovery Rate Parity - Ensure all protected groups have equally proportional false positives within the selected set (compared to the reference group).	**Failed**	Details
False Negative Rate Parity - Ensure all protected groups have the same false negative rates (as the reference group).	**Failed**	Details
False Omission Rate Parity - Ensure all protected groups have equally proportional false negatives within the non-selected set (compared to the reference group).	**Failed**	Details

---

Audit Results: Details by Fairness Measures

Equal Parity: **Failed**

What is it?	When does it matter?	Which groups failed the audit:

What is it?	When does it matter?	Which groups failed the audit:
This criteria considers an attribute to have equal parity is every group is equally represented in the selected set. For example, if race (with possible values of white, black, other) has equal parity, it implies that all three races are equally represented (33% each)in the selected/intervention set.	If your desired outcome is to intervene equally on people from all races, then you care about this criteria.	**For race** (with reference group as **Caucasian**) Other with **0.10X** Disparity Asian with **0.01X** Disparity African-American with **2.63X** Disparity Native American with **0.01X** Disparity Hispanic with **0.20X** Disparity **For gender** (with reference group as **Male**) Female with **0.21X** Disparity

Go to Top

Proportional Parity: **Failed**

What is it?	When does it matter?	Which groups failed the audit:
This criteria considers an attribute to have proportional parity if every group is represented proportionally to their share of the population. For example, if race with possible values of white, black, other being 50%, 30%, 20% of the population respectively) has proportional parity, it implies that all three races are represented in the same proportions (50%, 30%, 20%) in the selected set.	If your desired outcome is to intervene proportionally on people from all races, then you care about this criteria.	**For race** (with reference group as **Caucasian**) Native American with **2.20X** Disparity Asian with **0.68X** Disparity Other with **0.62X** Disparity African-American with **1.74X** Disparity

Go to Top

False Positive Rate Parity: **Failed**

What is it?	When does it matter?	Which groups failed the audit:
This criteria considers an attribute to have False Positive parity if every group has the same False Positive Error Rate. For example, if race has false positive parity, it implies that all three races have the same False Positive Error Rate.	If your desired outcome is to make false positive errors equally on people from all races, then you care about this criteria. This is important in cases where your intervention is punitive and has a risk of adverse outcomes for individuals. Using this criteria allows you to make sure that you are not making false positive mistakes about any single group disproportionately.	**For race** (with reference group as **Caucasian**) Other with **0.58X** Disparity Asian with **0.40X** Disparity African-American with **1.92X** Disparity Native American with **2.27X** Disparity

Go to Top

False Discovery Rate Parity: **Failed**

What is it?	When does it matter?	Which groups failed the audit:

What is it?	When does it matter?	Which groups failed the audit:
This criteria considers an attribute to have False Discovery Rate parity if every group has the same False Discovery Error Rate. For example, if race has false discovery parity, it implies that all three races have the same False Discvery Error Rate.	If your desired outcome is to make false positive errors equally on people from all races, then you care about this criteria. This is important in cases where your intervention is punitive and can hurt individuals and where you are selecting a very small group for interventions.	For race (with reference group as Caucasian) Asian with 0.71X Disparity For gender (with reference group as Male) Female with 1.40X Disparity

Go to Top

False Negative Rate Parity: Failed

What is it?	When does it matter?	Which groups failed the audit:
This criteria considers an attribute to have False Negative parity if every group has the same False Negative Error Rate. For example, if race has false negative parity, it implies that all three races have the same False Negative Error Rate.	If your desired outcome is to make false negative errors equally on people from all races, then you care about this criteria. This is important in cases where your intervention is assistive (providing helpful social services for example) and missing an individual could lead to adverse outcomes for them. Using this criteria allows you to make sure that you're not missing people from certain groups disproportionately.	For race (with reference group as Caucasian) Native American with 0.00X Disparity African-American with 0.57X Disparity Asian with 0.76X Disparity Other with 1.33X Disparity

Go to Top

False Omission Rate Parity: **Failed**

What is it?	When does it matter?	Which groups failed the audit:
This criteria considers an attribute to have False Omission Rate parity if every group has the same False Omission Error Rate. For example, if race has false omission parity, it implies that all three races have the same False Omission Error Rate.	If your desired outcome is to make false negative errors equally on people from all races, then you care about this criteria. This is important in cases where your intervention is assistive (providing help social services for example) and missing an individual could lead to adverse outcomes for them , and where you are selecting a very small group for interventions. Using this criteria allows you to make sure that you're not missing people from certain groups disproportionately.	**For race** (with reference group as **Caucasian**) Asian with **0.43X** Disparity Native American with **0.00X** Disparity **For gender** (with reference group as **Male**) Female with **0.72X** Disparity

Go to Top

Audit Results: Details by Protected Attributes

race

Attribute Value	Equal Parity	Proportional Parity	False Discovery Rate Parity	False Positive Rate Parity	False Omission Rate Parity	False Negative Rate Parity
African-American	African-American	African-American	African-American	African-American	African-American	African-American
Asian	Asian	Asian	Asian	Asian	Asian	Asian
Caucasian	Ref	Ref	Ref	Ref	Ref	Ref

Attribute Value	Equal Parity	Proportional Parity	False Discovery Rate Parity	False Positive Rate Parity	False Omission Rate Parity	False Negative Rate Parity
Hispanic	Hispanic	Hispanic	Hispanic	Hispanic	Hispanic	Hispanic
Native American	Native American	Native American	Native American	Native American	Native American	Native American
Other	Other	Other	Other	Other	Other	Other

Go to Top

gender

Attribute Value	Equal Parity	Proportional Parity	False Discovery Rate Parity	False Positive Rate Parity	False Omission Rate Parity	False Negative Rate Parity
Female	Female	Female	Female	Female	Female	Female
Male	Ref	Ref	Ref	Ref	Ref	Ref

Go to Top

Audit Results: Bias Metrics Values

race

Attribute Value	Predicted Positive Rate Disparity	Predicted Positive Group Rate Disparity	False Discovery Rate Disparity	False Positive Rate Disparity	False Omission Rate Disparity	False Negative Rate Disparity
African-American	2.63	1.74	0.86	1.92	1.21	0.57
Asian	0.01	0.68	0.71	0.4	0.43	0.76
Caucasian	1.0	1.0	1.0	1.0	1.0	1.0
Hispanic	0.2	0.84	1.09	0.88	1.03	1.17
Native American	0.01	2.2	0.93	2.27	0.0	0.0

Attribute Value	Predicted Positive Rate Disparity	Predicted Positive Group Rate Disparity	False Discovery Rate Disparity	False Positive Rate Disparity	False Omission Rate Disparity	False Negative Rate Disparity
Other	0.1	0.62	0.99	0.58	1.04	1.33

Go to Previous

Go to Top

gender

Attribute Value	Predicted Positive Rate Disparity	Predicted Positive Group Rate Disparity	False Discovery Rate Disparity	False Positive Rate Disparity	False Omission Rate Disparity	False Negative Rate Disparity
Female	0.21	0.89	1.4	1.0	0.72	1.07
Male	1.0	1.0	1.0	1.0	1.0	1.0

Go to Previous

Go to Top

Audit Results: Group Metrics Values

race

Attribute Value	Group Size Ratio	Predicted Positive Rate	Predicted Positive Group Rate	False Discovery Rate	False Positive Rate	False Omission Rate	False Negative Rate
African-American	0.51	0.66	0.58	0.35	0.42	0.35	0.28
Asian	0.01	0.0	0.23	0.29	0.09	0.12	0.38
Caucasian	0.34	0.25	0.33	0.41	0.22	0.29	0.5
Hispanic	0.08	0.05	0.28	0.44	0.19	0.3	0.58

Attribute Value	Group Size Ratio	Predicted Positive Rate	Predicted Positive Group Rate	False Discovery Rate	False Positive Rate	False Omission Rate	False Negative Rate
Native American	0	0.0	0.73	0.38	0.5	0.0	0.0
Other	0.06	0.03	0.2	0.4	0.13	0.3	0.66

Go to Previous

Go to Top

gender

Attribute Value	Group Size Ratio	Predicted Positive Rate	Predicted Positive Group Rate	False Discovery Rate	False Positive Rate	False Omission Rate	False Negative Rate
Female	0.19	0.17	0.41	0.48	0.3	0.24	0.4
Male	0.81	0.83	0.46	0.35	0.3	0.33	0.38

Go to Previous

Go to Top

Appendix B

Aequitas - Additional Visualizations

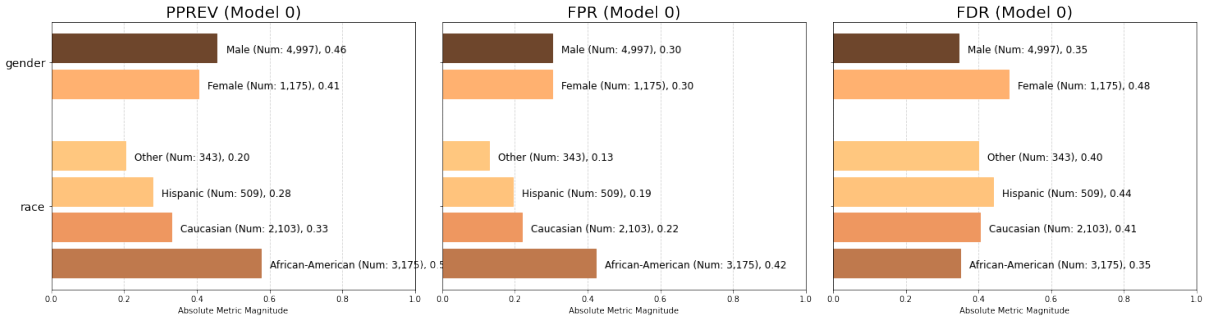

Figure B.1: Comparison of group metrics by gender and race: PPREV – Predicted Positive Group Rate, FPR – False Positive Rate, FDR – False Discovery Rate.

Source: Own analysis with Aequitas Python library based on the COMPAS algorithm dataset.

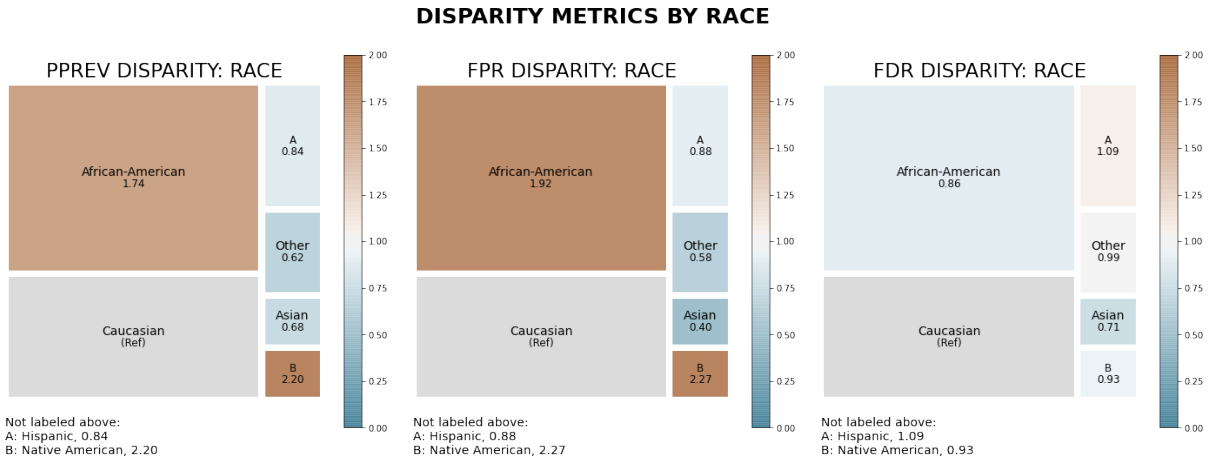

Figure B.2: Comparison of disparity ratios by race: PPREV – Predicted Positive Group Rate, FPR – False Positive Rate, FDR – False Discovery Rate.

Source: Own analysis with Aequitas Python library based on the COMPAS algorithm dataset.

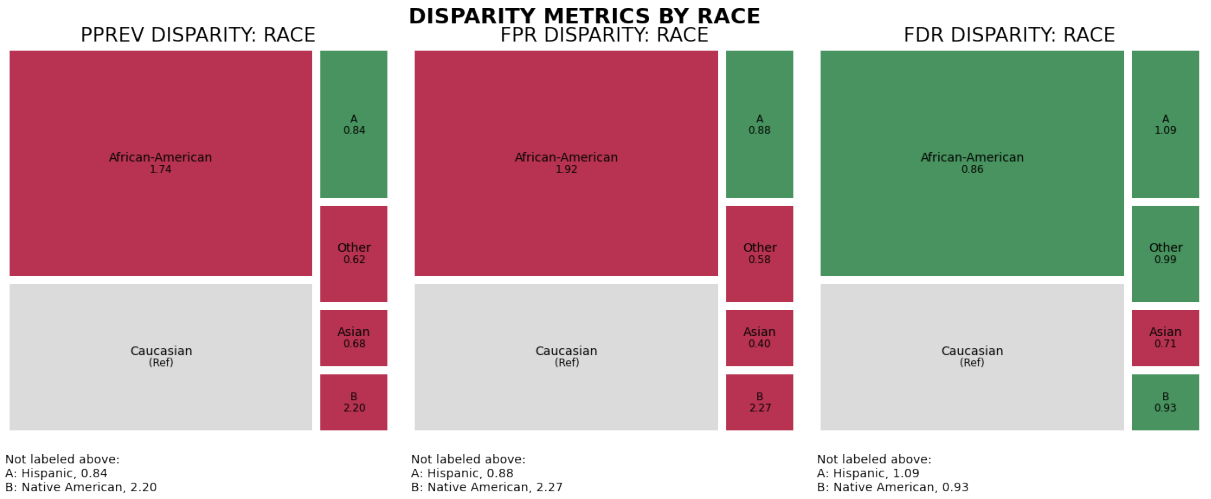

Figure B.3: Comparison of fairness audit results and disparity ratios by race: PPREV – Predicted Positive Group Rate, FPR – False Positive Rate, FDR – False Discovery Rate.

Source: Own analysis with Aequitas Python library based on the COMPAS algorithm dataset.