Mitigating Algorithmic Bias in COMPAS Recidivism Algorithm

Matthew Benigni
Marquette University
Milwaukee, Wisconsin
matthew.benigni@marquette.edu

Patrick Burns
Marquette University
Milwaukee, Wisconsin
patrick.m.burns.@marquette.edu

Brad Cooley
Marquette University
Milwaukee, Wisconsin
brad.cooley@marquette.edu

Connor Diggins
Marquette University
Milwaukee, Wisconsin
connor.diggins@marquette.edu

Alexander Franklin
Marquette University
Milwaukee, Wisconsin
alexander.franklin.@marquette.edu

Sam Speake
Marquette University
Milwaukee, Wisconsin
sam.speake@marquette.edu

Javonte Tucker
Marquette University
Milwaukee, Wisconsin
javonte.tucker@marquette.edu

Shion Guha
Marquette University
Milwaukee, Wisconsin
shion.guha@marquette.edu

ABSTRACT

Concepts such as discrimination, fairness, and bias in machine learning have been addressed by many researchers in the past. Each term can be defined in a number of ways. In our analysis, we provide specific definitions as we approach the problem of mitigating bias in the COMPAS risk-assessment algorithm. Algorithms are utilized in many industries ranging from education to mortgage loans, criminal sentencing, and credit approvals. Over the past decade algorithms have continued to advance and impact human decision making across these industries. We will present evidence that the COMPAS algorithm has shown bias towards people based on their protected attributes, such as race and sex. Our goal is to display an improved model, which mitigates bias within the COMPAS algorithm. We will present empirical evidence that our model mitigates bias through the use of statistical methods such as hypothesis z-testing and statistical parity. Our proof of concept will hopefully add to the research of the development of fairer and more efficient sentencing within the criminal justice systems.

PVLDB Reference Format:
Matthew Benigni, Patrick Burns, Brad Cooley, Connor Diggins, Alexander Franklin, Sam Speake, Javonte Tucker, and Shion Guha. Mitigating Algorithmic Bias in COMPAS Recidivism Algorithm . PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at http://vldb.org/pvldb/format_vol14.html.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

1 KEYWORDS

Recidivism, Algorithmic Bias, Machine Learning, Fairness, Criminal Justice System

2 INTRODUCTION

2.1 Background

Criminal sentencing and its effectiveness have always depended on the accuracy of the decision maker and their ability to predict future crimes. Judges have been in control of decisions regarding whether a defendant will receive bail, probation, parole, or the likelihood of a defendant to re-offend. However, empirical evidence shows that judges' decisions are often inaccurate, making them poor predictors for future offending [3]. Biases and discrimination could be found in almost every industry ranging from education, bank loans, credit, mortgages, and the criminal justice system. We define direct *discrimination* as the unfair or unequal treatment of people based on their membership to a category, rather than on individual merit [12, 18]. *Disparate treatment* is another term that is used to describe direct discrimination. *Indirect* or *unintentional discrimination* occurs where biases exists when a selection process has widely different outcomes for different groups, even as it appears to be neutral [9]. Indirect discrimination is also referred to as *disparate impact*. It is unlawful for organizations to discriminate on the grounds of sex, gender, race, ethnicity, religion, and nationality, to name a few. We define individuals who are discriminated against as "protected groups." For example, women can also be described as a protected group when looking at cases such as equal employment in the workforce. In the criminal justice system, African Americans can be described as a protected group in regard to predictive policing.

Predictive policing is an area within the criminal justice system in which law enforcement uses analytical techniques to make statistical predictions about potential criminal activity [4].

2.2 Bias in Law Enforcement

Predictive policing uses 'big data' and information technology that predict areas of future crime locations. The goal of predictive policing is to use the information gathered to reduce crime. There are two main forms of predictive policing: location-based prediction and person-based prediction. In the case of location-based prediction, officers are often deployed to areas where clusters of crimes are. This method acts as a practical way of issuing police in those areas in an effort to intervene and reduce crime in those areas. Person-based prediction may predict individuals or groups most likely to be involved in crimes, either as victims or offenders [4]. For example, data may be used to analyze the social network of gang affiliated members to identify other individuals who may be affiliated and are at higher risk of committing crimes. While the objective of predictive policing is to reduce crime, it is often biased towards certain groups, especially African Americans. The Fourth Amendment of the U.S. Constitution states:

> "The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized."

Police have the right to stop and frisk if they have a search warrant, probable cause to search, or a reasonable suspicion that a crime has been, is being, or is about to be committed by the suspect. Predictive models suggest that certain locations or individuals are at "high-risk" of offending. As a result, police often have a heightened awareness of criminal activity when patrolling those areas (whether or not someone is seen committing a crime) [4]. Police officers are deployed more frequently in communities that are deemed "high-risk", which correlates to higher arrest towards members of that respective region. Consequently, there are higher false-positive rates of arrest in those regions (people who did not commit crimes but were apprehended). The predictive policing model can show bias towards certain groups of people considering that the area where police are deployed the most will always be classified as "high-risk" given the frequent contact between police and members of that respective community.

2.3 COMPAS Algorithm

The Wisconsin state government considered using the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) recidivism algorithm to aid judges during each stage of the criminal justice process. In theory, judges should be able to make more efficient decisions regarding bail or pretrial release given the information provided to them [3]. However, the COMPAS algorithm has proven to be biased toward black defendants. Empirical evidence shows that black offenders are more likely than white offenders to be incorrectly judged a higher risk of recidivism, while white defendants are more likely than black defendants to be incorrectly judged as low risk of recidivism [13]. Research was conducted on the COMPAS algorithm by ProPublica. Their assessment was based on nine risk factors such as criminal history, current case characteristics, and demographic characteristics. The study consisted of 7,000 risk scores that were assigned to defendants in Florida [13]. Researchers were trying to uncover how many people from the sample would re-offend. The algorithm was only able to predict 20 percent of the recurrence of violent crimes. When racial disparity was considered, the algorithm predicted black defendants to be future criminals at twice the rate as white defendants [13], while white defendants were mislabeled as low risk more often than black defendants [13].

Research Questions

(1) What does it mean for an algorithm to be fair?
(2) Is there a data-driven approach towards mitigating bias?
(3) There is limited empirical research on the efficacy of risk-assessment algorithms in producing more fair outcomes. Can we provide empirical evidence to show that aspects of the COPMAS risk-assessment algorithm can be debiased?

2.4 Fairness

We will base our definition of fairness as used in article [15]. Fairness can be categorized in two ways: *individual* and *group fairness*. Group fairness is defined by separating a population into protected attributes (such as gender, age, or ethnicity) and seeks for some statistical measure to be equal across groups. Individual fairness seeks for similar individuals to be treated similarly irrespective of socio-economic factors [14]. Optimizing fairness involves trade-offs between accuracy measure and discrimination, limiting distortion in individual data samples, and preserving utility [5]. Bias can occur during each stage of the criminal justice system. Pre-processing (data collection), in-processing (classification), and post-processing (outcome predictions). The COMPAS data is collected through the answers from a 137-item questionnaire to predict the risk-score of a defendant [20]. The offender's data from the risk-assessment questionnaire is weighted by the offender's criminal history, and basic demographic information, such as age and gender. A score is then generated which can be used to categorize defendants into risk categories ranging from low to high. Algorithms that influence human decision making have received a lot of skepticism due to their lack of transparency. Their decision-making processes are often opaque – which makes it difficult to explain why a certain decision was made [7]. These algorithmic systems are often referred to as a "black-box" because the complex models that output prediction scores are not always publicly available. It is therefore hard to acquire a deeper understanding of model behavior, and in particular how different features influence the model prediction [2]. It is essential to understand this paradigm because their algorithmic systems may impact how judges, prosecutors, and court staff exert their own discretion [6]. Algorithms also have trouble detecting discrimination because other attributes such as personal data, economic and cultural indicators often act as proxies for indirect discrimination. For example, the classic case of redlining. It has been observed that even when the decision-making process explicitly excluded race, but uses zip codes as an attribute, people living in a certain neighborhood frequently get credit denial [2, 19]. Here, race had a disparate impact on the outcome via the zip code, which acts as a proxy towards the people living in that neighborhood who belong to the ethnic minority class [2, 19].

2.5 Project Focus

The goal of our research is precisely to address the problem of discovering discrimination within the COMPAS risk-assessment algorithm based on historical decision records. We derived our data from several datasets which includes columns for violent vs non-violent crimes, defender demographics and time spent in jail for defenders. Given the research prior to our assessment, it has been empirically shown that the COMPAS data is biased towards certain protected groups. In our analysis, we will compare the current COMPAS model to our improved model. Our research is not proof. We will show by means of statistical methods that our model mitigates discrimination towards certain protected groups through statistical evidence.

3 RESEARCH

3.1 How Bias can be Measured

Measuring bias in an algorithm can prove a difficult task. There are many factors to consider, one of the most important parts being investigating the source of the bias. While an algorithm itself can be biased, in many cases the data it is trained on is already inherently biased. For example, our dataset is riddled with biases. It is well known that African Americans are incarcerated at a disproportionate rate in the United States, so it is likely many of our records would not exist at all had the person been Caucasian rather than African American. On top of this, we are trying to predict recidivism of each person. For the same reason, it is more likely that the number of African Americans who "recidivate" will be overstated because they are more likely to be arrested. This means that the data we will be training on is already biased, and any algorithm we create will inherit that bias in some way [10, 16]. Thus, we aim to measure and reduce that bias.

One of the first steps we took towards measuring bias was proving that there is bias in our baseline model. To do so, we conduct a simple Z-test for difference of proportions. We are comparing the proportion of African Americans who are put into the high-risk group, versus the proportion of Caucasians who are put into the high-risk group. Our null hypothesis is that these two proportions should be equal.

$$H_0 : P_c - P_a = 0$$
$$H_a : P_c - P_a \neq 0$$
$$\alpha = 0.01$$

Where P_c is the proportion of Caucasians that are categorized as "high risk" and P_a is the proportion of African Americans categorized as "high risk", and P_{co} will be the combined rate. Our Z-score is given by

$$Z = \frac{P_c - P_a - H_0}{\sigma_{P_c - P_a}}$$

and

$$\sigma_{P_c - P_a} = \sqrt{\frac{P_{co}(1 - P_{co})}{n_c} + \frac{P_{co}(1 - P_{co})}{n_a}}$$

Where n_c and n_a are the total number of people in our Caucasian and African American test sample, respectively. We reference our code output to get these values as well as the proportion of the populations that were categorized high risk. So, building a tabular output based on our code yields Table 1.

Classification	Caucasian	African American	Combined
Low/Medium Risk	623	734	1357
High Risk	25	196	221
Total	648	930	1578

Table 1: Dataset classification between two races

Thus

$$P_c = 0.0386$$
$$P_a = 0.2108$$
$$P_{co} = 0.1401$$
$$n_c = 648$$
$$n_a = 930$$

We now calculate

$$\sigma_{P_c - P_a} = \sqrt{\frac{0.14(1 - 0.14)}{648} + \frac{0.14(1 - 0.14)}{930}} = 0.0178$$

and our Z-score is

$$Z = \frac{0.0386 - 0.2108 - 0}{0.0178} = -9.67$$

With a Z-score so extreme, our p-value ≈ 0. Thus, we conclude that there is a statistically significant difference between the rates at which Caucasians and African Americans are being categorized as high-risk.

Moving forward, we need some metric to show bias in each of our models. Racial disparity is the clearest bias in the COMPAS algorithm, so we will focus primarily on group fairness, otherwise known as statistical parity. A model has achieved statistical parity when the proportion of people in the protected group who are classified favorably is equal to the proportion of people not in the protected group that are classified favorably. Also, the proportion of people classified unfavorably should be equal for both classes [24]. In order to measure bias, we conduct a statistical parity difference test, which is simply the difference between the proportion of the protected group and the proportion of the unprotected group that is being classified into a certain outcome. When we conduct this test, we subtract the protected group proportion from the remaining population proportion. The result will be within $(-1, 1)$. An output of zero would indicate that statistical parity has been achieved, and both groups are being classified positively or negatively at the same rate. Otherwise a negative number would indicate bias towards the protected group, a positive number indicates bias towards the rest of the population. If the metric falls within $(-0.1, 0.1)$ we consider that to be "fair" [1]. Still, if we approach zero as we develop our model that shows that we have reduced bias by improving group fairness.

3.2 How Accuracy can be Measured

In the context of our project, we measure accuracy by looking at true positives and false negatives. We determine positives and negatives by looking at whether or not a person ended up actually recidivising after their initial arrest. In order to do this, we had to filter out any records where there was no information on whether or not a person recidivated. Further, both COMPAS' model and the models we developed have ordinal outputs of three risk categories, but we are looking at the binary outcome of whether or not a person recidivated. To address this, we decided that we would not be looking at people categorized as medium-risk. Medium does not give us a clear outcome, (yes, they will recidivate, or no, they will not), so it would not make sense to try and measure accuracy by those instances anyway.

Instead, we focus on the two extremes, high-risk and low-risk. We expect people categorized high-risk to recidivate frequently, while on the opposite side we expect people who were categorized low-risk to not recidivate. As such, we calculate accuracy as the percentage of high-risk people who recidivate, and as the percentage of low-risk people who do not recidivate. This was simple to implement. Count the number of records that the model ranked high-risk (low-risk) and recidivated (did not recidivate), and divide that by a count of the number of rows that are ranked high-risk (low-risk). A high percentage would indicate high accuracy for both measures. This will allow us to compare accuracy between models, as we aim to create a model with similar (or better) accuracy than the original COMPAS model.

3.3 Debiasing Without Losing Accuracy

One of the greatest challenges of reducing bias is maintaining accuracy. If there was no concern for accuracy, we could simply suppress or leave out all potentially biased attributes from our algorithm. In theory this would be a simple way to create an unbiased model, but we would need to sacrifice some of the most important predictors. As discussed prior, incarceration rates are not equal for all groups. So this would mean suppressing factors such as prior arrest record. Clearly, the arrest record of a person is one of the best predictors as to whether or not they will commit a crime. Thus, we need to find some balance between debiasing our algorithm and accuracy of the model.

There are some ways that we could both reduce bias and maintain accuracy, although most of the time this is ethically wrong and will simply create a different type of bias. For example, we could approach the problem by simply giving instances of the unprivileged group more favorable labels. This of course would be illegal and is not an option [10]. One option that would be fair, reduce bias, and maintain accuracy would be creating separate models for each protected group, so in this case that would be a separate model for each race [16]. Unfortunately that raises both ethical and legal questions, and would not be usable in courts as COMPAS is. Something we will need to be mindful of is that in many cases it is easy to achieve statistical parity at the cost of individual fairness [8]. As such we will avoid making radical changes to our data and algorithms, being mindful of how other scopes of fairness could be affected.

Ultimately, it is unlikely that we will debias an algorithm without losing accuracy, so our work becomes finding some threshold of accuracy we are willing to sacrifice for a fairer model. Since we do not have access to COMPAS' model, we will be creating our own model and can aim to improve accuracy at the baseline. Then hopefully any reduction in accuracy from our bias mitigation will still be comparable to, or better than COMPAS.

3.4 Prior Work

Through our research, we encountered numerous existing attempts to debias algorithms, not all in the context of prison sentencing. Bias mitigation has been attempted to address algorithmic problems in areas such as loan approval, online recruitment and hiring, and most other machine learning algorithms. Since machine learning algorithms often analyze trends in human activity and data, it is easy for human biases to be perpetuated through these algorithms. This is the main goal that many of the researched, debiasing algorithms share: to develop algorithms that do not exhibit the same biases that humans do. Articles that we have looked at specifically touch on several different topics in reference to debiasing algorithms. One of these topics is the attempt to increase both individual and group fairness. Group fairness is defined by splitting a population into protected attributes (such as gender) and seeks for some statistical measure to be equal across groups. Individual fairness seeks for similar individuals to be treated similarly [15]. Obviously, the best case scenario is to develop an increase in both individual and group fairness, but sometimes these cannot both happen. This leads to the problem of having to decide between the two measures. Another topic commonly referenced in relevant articles is the topic of false-positives or false-negatives. Although a small degree of false-positives or false-negatives in an algorithm is nearly unavoidable, they still show a need to be evaluated. One concern is the distribution of these false results. Results that disproportionately affect one population subgroup over another (e.g., black males are falsely predicted to recidivate more often than white males are falsely predicted) [17]. Some of the analyzed articles point their focus toward problems like these. A third topic commonly referenced is the measures and relationship between bias and accuracy, as discussed in the previous section. These three topics, among others, all have connected bias mitigation methods that attempt to limit the challenges present in algorithms.

3.5 Bias Mitigation Methods

Bias mitigation methods can generally be broken up into two categories: pre-processing and post-processing. As can likely be assumed, pre-processing methods involve transformation of the raw data before being fed into an algorithm, whereas post-processing is after the algorithm.

3.6 Pre-Processing Methods

A pre-processing method that occurs in almost every project is data cleaning. There are some basic steps to data cleaning. First is to remove duplicate or irrelevant observations. Another step is to filter out outliers. Finally, handling incomplete data or null values is important as well. A second pre-processing technique that was examined is "reweighting". The reweighting algorithm creates weights for examples in each combination of protected characteristics in order to ensure fairness, before using the algorithm. It is

essentially giving weights so the model can learn what to be careful of.

If the dataset D is unbiased, i.e., S and $Class$ are statistically independent, the expected probability $P_{exp}(S = b \wedge Class = +)$ would be:

$$P_{exp}(S = b \wedge Class = +) := \frac{|\{X \in D | X(S) = b\}|}{|D|} \times \frac{|\{X \in D | X(Class) = +\}|}{|D|}$$

[11]
However, the method `reweight()` under R-package `fairmodels` is much easier to implement.

3.7 Post-Processing Methods

Post-processing methods were more commonly found in our research, in a number of different ways. One post processing method is *Reject Option-based Classification* (ROC). We know discrimination can occur in three places; the most discrimination occurs around the decision boundary (classification threshold). The method used here uses the low confidence region of a classifier for discrimination reduction and reject its predictions. Through this process the hope is to reduce the bias in model predictions. One advantage this method has over other methods is that the final predictions can be manipulated easily. Simply, if a subgroup is privileged, this algorithm will change their probabilities to the other side of the bias cutoff, which essentially tries to even out the difference between privileged and unprivileged subgroups. This concept is clearly represented by the method, `roc_pivot()`, under R-package `fairmodels`.

A second solution, called *Discrimination-Aware Ensemble* (DAE), exploits the disagreement region of a classifier ensemble to relabel deprived and favored group instances for reduced discrimination. If classifiers predict the deprived group, they are assigned the C+ label, and instances that belong to the favored group are given the C- label [12]. DAE works well compared to ROC when the dataset is not limited to probabilistic classifiers, which is what ROC fits well with. There are several advantages to DAE. Solutions are not restricted to a particular classifier. Solutions require neither modification of learning algorithms nor preprocessing of historical data. Solutions give better control and interpretability of discrimination-aware classification to decision makers.

A third post-processing solution is *equalized odds*. Equalized Odds (also referred to as Disparate Mistreatment) is a classification algorithm which aims to ensure that no error type (false-positive or false-negative) disproportionately affects any population subgroup. It contains a method called Relaxed Equalized Odds with Calibration, which analyses the trade-offs between false-positive and false-negative rates. The algorithm provided achieves the calibrated Equalized Odds relaxation by post-processing existing calibrated classifiers [17].

4 THE BASELINE MODEL

4.1 Ideation

Seeing as the COMPAS sentencing algorithm is proprietary, we could not gain an understanding of precisely how it classified individuals as low, medium, or high risk for recidivism. Because of this, all we had to draw an understanding from was a data file called `compas-scores.csv`, for which each observation was an individual who had gone through the Wisconsin criminal justice system. The attributes were various factors about the individual, ranging from protected attributes such as race and sex to specifics about their criminal history such as what crime they had committed and whether or not they had previous felonies or misdemeanors. Finally, for each individual there was an attribute `score-text`, which is the output of the COMPAS sentencing algorithm on their data, as well as for some there was a boolean variable as to whether or not they had recidivated. With this information, we were able to begin to piece together a strategy to replicate their algorithm, measure the bias in it, and then find a way to create a similarly accurate model which would be quantifiable less biased based on an individual's protected characteristics. As we could not directly recreate the COMPAS algorithm, we set our sights on creating our baseline model, an ordinal regression with no anti-bias measures in place on which to measure our future endeavors against.

The first step was to understand the data we were working with in order to be able to do it justice and use it correctly. To achieve this understanding we created a master file translating each attribute name to an English sentence description of what it represented, along with the possible values it could hold. This would prove to be a useful reference moving forward with development and analysis. Furthermore, this deep dive into the nature of the attributes we had to work with provided us with the ability to clean our data, as well as derive useful attributes from less useful ones.

As far as cleaning, we were able to remove any rows for which the COMPAS algorithm had not given the individual a score, as without that we had no basis to compare against for that individual. We removed rows for which we could not know for sure whether or not they were actually being charged with the crime that was recorded based on the days between their screening and arrest. Finally, we removed any rows for which we had no data on whether or not the individual has recidivated, as without that we had no ability to measure accuracy, which is important in order to make sure that we do not trade accuracy in favor of less bias, rather improving both.

Following the cleaning of data, we set out to squeeze useful information out of the attributes which were lacking in usability. The first example of this was the variables representing when the individual went to jail, and when they were released from jail. On their own, an ordinal regression model has a difficult time accurately deriving use from these, but when subtracted to create a variable representing the duration of an individual's time spent in jail we now had a useful representation. Another example of this is the category which gave a detailed description of the charge the individual was being tried for. The charges descriptions were not simple, repeated categories but police descriptions of the arrest. For example, rather than something similar to "minor drug distribution," we had instances of the attribute such as "selling cocaine within

twenty feet of a church." Due to this inconsistency, there were 655 different values for the, technically, categorical variable, making the use of this attribute entirely insignificant in our model. To combat this, with a combination of a python script and hand combing, we created a new attribute categorizing offender's crimes into being either violent or non-violent, once again creating an extremely useful attribute from a useless one.

Finally, we randomly selected 80% of the dataset to be used as training data, leaving 20% of the dataset for validation. This would allow us to quickly and efficiently test iterations of our model for accuracy, the most important factor as we needed our baseline model to be as accurate as possible before we attempted to reduce the present bias within it. Now, having our data cleaned, prepared, and split, as well as a clear vision of what we needed to accomplish to begin working on our proof of concept, it was time to begin the development of our baseline ordinal regression.

4.2 Creation

To create this model, it seemed most appropriate to work in R, a language and environment for statistical computing and graphics, as this would provide a wide range of useful libraries, as well as allow for easy porting of any relevant outputs, in the form of statistical results and/or graphs, to the R Shiny dashboard that would eventually be created as a visual aid to our proof of concept we were working towards [21, 23]. Within the packages available, we found the `polr` function in the `MASS` package, a function which would fit a proportional odds logistic regression of the form `response ~predictors` resulting in an ordered factor, interpreted as an ordinal response. Hence, our ordinal response being the ranking of an individual's risk of recidivism as low, medium, or high.

4.3 Figures

Seeing as the model being created was relatively straightforward in theory, and a robust library of tools had been selected, it was simple to start by throwing all available attributes at the issue. This, of course, resulted in a model which basically just predicted an individual's classification to a category based on their race, which makes sense seeing as there was absolutely no effort in play to remove the embedded bias from historical criminal justice data. In fact, effort had basically been done in the opposite direction by including protected attributes, as well as proxies for protected attributes. With this result, we began the process of removing attributes which perpetuated biases wherever possible, trying to get the model to be as similar as possible to the COMPAS sentencing algorithm in terms of raw results. The first steps were to take out protected attributes, such as race and sex. This saw slight to no improvements, seeing as proxies still existed. Hence, removing factors which serve as a proxy for protected attributes, such as zip code for race, saw further progress towards this goal we were aiming for. Finally, after further sifting of attributes, we had what, at first look, seemed to be a race-less model which gave results relatively resembling that of the COMPAS sentencing algorithm.

4.4 Results

In order for this baseline model to be useful, there needed to be some way for its results to be compared to that of future iterations of the model in terms of bias and accuracy. At this point, it was decided to put off the creation of a standardized accuracy measure in favor of focusing solely on the reduction of embedded biases, due to the relatively short time-frame in which the project was set to be accomplished. To test for bias, we decided the method which fit our model and outlook the most was that of statistical parity. Statistical parity provides a numeric measure of bias, and a model achieves statistical parity if it always categorizes the general population the same way that it categorizes the protected class. In our use case, this meant that our improved models would achieve true statistical parity if they treated people of all races, specifically African Americans, the same way that other races, typically Caucasians, when generating sentencing recommendations.

Numerically, a statistical parity output between $(-0.1, 0.1)$ implies that the model in question is not statistically significantly biased. When applied to the baseline model, the output was -0.17, a result substantially outside of the perfect range.

This level of bias is visible when looking at individual outputs with the individual's race reattached after their evaluation for recidivism. For example, the baseline model was less likely to give a Caucasian with a prior charge a "high risk" ranking than an African American without a prior charge, as well as a Caucasian charged with a felony was also less likely to be categorized as 'high risk' than an African American charged with a misdemeanor. This quantifiable evidence of a high level of bias in a baseline criminal sentencing algorithm became the basis for building a proof of concept that bias can be reduced significantly, ideally without a significant loss of accuracy.

4.5 Future Discussion

The prior strategy comes with the caveat that we are basing our comparison on a model that we created rather than the actual algorithm used by COMPAS. This sparked the beginning of discussions on how to make sure that we are not just showing an improvement from our own, possibly sub-par, model, but an improvement from that of the COMPAS sentencing algorithm. This is shown in later results, as statistical analysis for bias and accuracy can still be run on the COMPAS outputs available to us via the provided dataset.

With that caveat reasonably resolved, the focus shifts to deciding what the next steps are to create an applicable proof of concept. Would we be making changes to the model itself, in pre- or post-processing, or changing the attributes further, or follow a plethora of other methodologies uncovered during the exploratory phase of our research? These are all interesting possibilities to attempt in order to accomplish the goal of creating an algorithmic sentencing model which would produce less biased, accurate results, proving the feasibility of the unbiased use of algorithms throughout the criminal justice system.

5 THE IMPROVED MODEL

5.1 Ideation

One of the first ideas that we wanted to try out was reweighing our dataset. Reweighing weighs the examples in each (group, label) combination differently to ensure fairness before classification. (pre-process) Pre-processing is one of the three areas of focus when considering mitigating bias in algorithms. As a group we discussed

several approaches (such as disparate impact in relation to group fairness, individual fairness, and equal error rates) towards making our algorithmic system fairer. This then proposes an optimization formula comprised of three goals:

- controlling discrimination (trade-offs)
- limiting distortion in individual data samples
- preserving utility

The distortion constraint included in this model distinguishes it from the previous approaches. Our goal is to determine a randomized mapping that transforms the given dataset into a new dataset, which may be used to train a model, and similarly transforms data to which the model is applied so that we can test the data. The optimization in this study utilizes a probabilistic framework for discrimination-prevention pre-processing in specific learning.

Once we understood how the data wanted to present itself in the way we wanted we could then use our data correctly. We wanted to apply a learning algorithm for fair classification that accommodates for both individual fairness and group fairness. Group fairness is defined as people of protected variable to have similar proportion to total population. Individual fairness is defined by people of similar qualifications and will be rated similarly.

Secondly, a key method that was used in our project is statistical parity. We can utilize statistical parity in our project to promote group and individual fairness for race, gender, and age. Statistical parity is defined as the probability of someone in a protected group being approved or the probability of anyone being approved. Many of its metrics are similar to disparate impact (group fairness).

Lastly, we wanted to implement statistical methods that maintain the high accuracy of these learning algorithms while reducing the degree to which they discriminate against individuals because of their membership in a protected group. In our case, if the protected attribute is race or gender – the classifier should not correlate someone's race or gender to the likelihood of them getting a higher risk score due to their membership of a particular group. Which then leads into adversarial debiasing and we found that it maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute (race, gender) from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can determine.

5.2 Creation

As we mentioned in our Baseline Model we also preferred to use RStudio when using the Improved Model [22]. It was appropriate to use RStudio as it is used for our graphics and we will then be able to see the outputs that are relevant to our needs [22]. We would also be able to create our statistical results and easily create visible graphs to the R Shiny dashboard to show our results that we are working towards [21].

While working within RStudio and our Improved Method we came across different tools to implement our dataset [22]. We tried to use the `AI Fairness` package, but in the end we were not successful. Instead we found a package that was successful with the fairness package. The name of the package that we wanted to implement was called `aif360`. This specific package is in a specific library that was going to be used to mitigate bias in machine learning models throughout the dataset. In order to use their methods, we created a new generalized linear model (GLM) model that allows for response variables that have error distribution models other than a normal distribution. This let us apply a reweighing method as well as a reject option classification method. Unfortunately, the reject option model's predict function does not work for that package but in the end it was still beneficial for us. Since it was still reversing the bias to an extreme that would have been worse than the visualization. In the end we found that the reweighed GLM model was a great improvement from COMPAS and our baseline model.

5.3 Results

The specific method used for mitigating bias was covered under Section section 3: Research. Shown by this method, the results from our improved model turned out to be overwhelmingly positive, especially in terms of bias mitigation.

5.4 Bias Results

Statistical Parity values were measured for each of the three models that we worked with (baseline model, COMPAS model, generalized linear model). Within each model, the values were measured for the rates in which individuals were considered high risk, as well as rates for low risk. Within the GLM, values were also measured before and after the implementation of the bias mitigation methods, covered previously. The Statistical Parity results can be seen in Table 2.

Model Type	Low Risk	High Risk
Baseline Model	0.3858	-0.1738
COMPAS Model	0.2959	-0.1899
GML Model (*before* bias mitigation)	0.2230	-0.1650
GML Model (*after* bias mitigation)	0.2249	-0.1510

Table 2: Statistical Parity results

From Table 2 we calculate the difference between our improved GLM Model and the COMPAS model:

$$\frac{0.1899 - 0.1510}{0.1899 - 0.1} = 0.433$$

This indicates that our model is 43.3% *less* biased when looking at the "High Risk" group. We can do the same calculation for the "Low Risk" group as well:

$$\frac{0.2959 - 0.2249}{0.2959 - 0.1} = 0.362$$

This indicates that our model is 36.2% *less* biased when looking at the "Low Risk" group.

5.5 Accuracy Results

A decrease in bias only holds significance when it is not accompanied with a similar decrease in accuracy. Accuracy is determined by measuring the percentage of high ranked individuals that actually did recidivate and the percentage of low ranked individuals that did not recidivate. The accuracy results can be seen in Table 3.

Model Type	Low Risk	High Risk
Baseline Model	0.736	0.571
COMPAS Model	0.767	0.534
GML Model (*before* bias mitigation)	0.753	0.580
GML Model (*after* bias mitigation)	0.760	0.570

Table 3: Accuracy results

5.6 Results Conclusions

Accuracy improved in the high risk category, from the baseline and COMPAS models to the GLM model (before bias mitigation). The bias mitigation caused a 1% decrease in accuracy between GLM models (in the "High Risk" group). However, a decrease of this insignificant magnitude of 1% is acceptable when partnered with the results of the bias testing. Among the more scrutinized high risk group, a model that is 43% less biased, while maintaining a significant amount of accuracy is exactly the type of result that this project set out to achieve. In a project with the finite timeline of one semester, the above results, achieved by the techniques outlined in previous sections, are definitely noteworthy and have the potential to be improved upon even more.

6 DASHBOARD

6.1 Ideation

Having found a way to reduce bias in algorithmic sentencing algorithms, we needed to find a way to share and display our findings in a way which would be accessible, yet informative, to all of our audiences. With the goal being to present our findings to the Wisconsin state government, it was apparent that the medium in which we chose to communicate our results would need to be meaningful and eye opening to both technical and non-technical audiences, a feat which is often easier discussed than executed. Furthermore, we did not want to alienate any members of our audience based on whether or not they were familiar with the COMPAS dataset and algorithm, ideally in a way which would allow those unfamiliar to see the impact of our debiasing as well as impress and drive the impact home to those who are familiar. With these requirements in mind, it quickly became apparent that an interactive dashboard would be our best option, allowing the user to drill down into the results based on their own level of technical training.

Naturally, the next step after deciding to create a dashboard is deciding what views we want to create within it. The work done by data scientists is often only as valuable as their ability to communicate findings to management, law makers, what have you. This means that we needed to make our findings speak for themselves as we would be handing off a dashboard rather than giving a formal presentation. Hence, three main components were envisioned which would bring our lofty goals to fruition. The first of which is a way for individuals to be able to play with the parameters of the algorithm and see results in real time. The second is a visual representation of our debiasing coupled with raw text output, allowing viewers to inspect graphs further, compare multiple graphs simultaneously, and export visuals for later use. Finally, the third is an overall set of interactions in which users can explore our work and findings, and compare what we have accomplished with the original, extremely biased results. The third idea was brought about in a relatively general manner that way, ideally, it could be seen as an overall result of the two more tangible components.

6.2 Creation

Having a solid idea set out for what needed to be created, it was time to decide what tools to utilize in order to actually build it. Discussion started within a cross section we found between data science and computer science academia and industry standard development. This cross section being the vanilla development practices of academia, think plain Python, R, and Java, in comparison with the vast libraries and frameworks which allow for much faster, more robust development of full fledged applications and products [23]. A happy medium we found was that of R Shiny, a very powerful statistical analysis tool with a large library of third party packages which does not take away from the value of native R [21, 23]. While similar tools exist for Python and JavaScript, R was decided on as many members of the team had experience with the R Shiny framework, making it much faster to start development and get mock-ups and wireframes spun up [21, 23]. Furthermore, R Shiny can be used and hosted completely for free even without a student license or academic contract, an invaluable aspect solely for the sake of avoiding headaches when handing off our findings and what we have created, as the entire team is graduating at the culmination of the project [21]. Finally, having a framework selected and a goal defined, the development lifecycle became relatively straightforward, with the largest workload being in spinning up different types of graphs to analyze which visualizations gave the best representation of our findings. To speed up this process, the `Plotly` package was used for all of the graphing as it includes a plethora of built in features such as zooming, exporting, and data point comparison. This allowed for quick comparisons between types of graphs without the hassle of redeveloping many necessary features.

6.3 Results

The aforementioned ideation and creation processes resulted in a fully functioning dashboard hosted on GitHub, allowing for free and easy access. The dashboard shows the differences between our debiased model and the original model, both in statistical output and visual output, as seen in Figure 1. The statistical output allows for the statistically inclined to gain an understanding of exactly how much improvement we were able to garner from the methods implemented, while the graphical output allows for quickly gaining a personalizable level of understanding of tangible differences between the models. These outputs allow the user to easily read and interpret the results.

We wanted the dashboard to be interactive, though, so additional features were added to allow the individual to do a deeper dive into the results. One of these features allows the user to choose which predictors they want to run the original model using, then re-runs the model and displays how much each predictor impacts the outcome. Another feature allows the user to compare these predictors between the original model and the improved model. The transparency that these features provide is important for the sake of algorithmic integrity and fairness, two pillars which are

extremely important in the criminal justice field as people's lives are on the line.

7 CONCLUSION

As shown, there is an empirically evident need for improved decision making in the criminal justice system in order to remove embedded systemic biases from the processes which decide the fate of those at the hands of the system. While many trainings, procedures, and laws have been put in place to attempt to remove this bias, they have shown themselves to be far from perfect solutions due to the continuation of injustice. With the rise of big data and data science it makes sense to attempt to mesh the fields, although no company or individual has yet to find a way to create the needed algorithms without embedding systemic bias, typically via the use of historical data. As mentioned, though, the solution can not be to simply remove the use of historical data as without data we have no model. Hence, the need to find statistical methods to remove bias was introduced, in order to prove the viability of algorithms in criminal justice.

The impact of a more fair system is likely farther reaching than one could imagine, although we can make an attempt regardless. Specifically in terms of more accurate, more fair sentencing algorithms based on the likelihood of an individual to ricidicate we would likely see smaller prison populations, a reduction in the race tensions created by law enforcement, and less lives ruined by the difficulties associated with reentering society post incarceration. These changes would have far reaching positive effects on communities, families, and individuals throughout the United States. Hence, it is an extremely worthy investment of time and resources to continue to research and develop ways to debias the algorithms which could be used in the criminal justice system.

Our group of Data Science students here at Marquette University attempted to take on this challenge with less than four full months to work on it, while each of us had a full schedule of courses, work, extra-curricular activities, as well as limited resources. Despite these limiting factors, we still were able to create an improved model with which we saw massive improvements in the accuracy and debiasing compared to the outcomes of the COMPAS model, a model created and sold to the Wisconsin state government. This is why we believe that it is a worthy investment for the Wisconsin state government to invest in data science based research into the creation of algorithms which have statistically insignificant levels of bias, which can then be employed to help judges throughout the state make accurate and fair sentencing decisions so that the individuals, communities, and families of Wisconsin can all see the massively positive effects that less biased criminal justice systems would usher in.

REFERENCES

[1] IBM [n.d.]. *AI Fairness 360.* IBM. Retrieved May 1, 2021 from https://aif360.mybluemix.net/

[2] Philip Adler, Casey Falk, Sorelle A. Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. Auditing Black-box Models for Indirect Influence. (2016), 1–10. http://sorelle.friedler.net/papers/auditing_icdm_2016.pdf

[3] Mirko Bagaric, Dan Hunter, and Nigel Stobbs. 2020. Erasing the Bias Against Using Artificial Intelligence to Predict Future Criminality: Algorithms are Color Blind and Never Tire. *University of Cincinnati Law Review* 88, 4, Article 3 (May 2020). https://scholarship.law.uc.edu/cgi/viewcontent.cgi?article=1365&context=uclr

[4] Sarah Brayne, Alex Rosenblat, and Danah Boyd. 2015. Predictive Policing. *Data Civil Rights: A New Era of Policing and Justice* (October 2015), 1–11. http://www.datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf

[5] Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized Data Pre-Processing for Discrimination Prevention. (April 2017), 1–18. https://arxiv.org/pdf/1704.03354.pdf

[6] Angèle Christin, Alex Rosenblat, and Danah Boyd. 2015. Courts and Predictive Algorithms. *Data Civil Rights: A New Era of Policing and Justice* (October 2015), 1–13. http://www.datacivilrights.org/pubs/2015-1027/Courts_and_Predictive_Algorithms.pdf

[7] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. (2016), 1–20. https://www.andrew.cmu.edu/user/danupam/datta-sen-zick-oakland16.pdf

[8] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2011. Fairness Through Awareness. (November 2011), 1–24. https://arxiv.org/pdf/1104.3913.pdf

[9] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. (July 2015), 1–28. https://arxiv.org/pdf/1412.3756.pdf

[10] Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. 2016. A Confidence-Based Approach for Balancing Fairness and Accuracy. (January 2016), 1–10. https://arxiv.org/pdf/1601.05764.pdf

[11] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge Information Systems* 33 (2011), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[12] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision Theory for Discrimination-aware Classification. *2012 IEEE 12th International Conference on Data Mining* 12 (2012). https://doi.org/10.1109/ICDM.2012.45

[13] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. *How We Analyzed the COMPAS Recidivism Algorithm.* ProPublica. Retrieved May 1, 2021 from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[14] Pranay Lohia. 2021. Priority-Based Post-Processing Bias Mitigation for Individual and Group Fairness. (January 2021), 1–5. https://arxiv.org/pdf/2102.00417.pdf

[15] Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, and Ruchir Puri. 2018. Bias Mitigation Post-Processing for Individual and Group Fairness. (December 2018), 1–5. https://arxiv.org/pdf/1812.06135.pdf

[16] Sarah Picard, Matt Watkins, Michael Rempel, and Ashmini Kerodal. 2019. Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness. (June 2019), 1–20. https://www.courtinnovation.org/sites/default/files/media/document/2019/Beyond_The_Algorithm.pdf

[17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. *31st Conference on Neural Information Processing Systems* (2017), 1–10. https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

[18] Andrea Romei and Salvatore Ruggieri. 2013. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* 00:0 (2013), 1–54. https://doi.org/10.1017/S000000000000000

[19] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2008. Data Mining for Discrimination Discovery. *Process of KDD* (2008), 1–49. http://pages.di.unipi.it/ruggieri/Papers/tkdd.pdf

[20] Matthew Stewart. 2019. *Handling Discriminatory Biases in Data for Machine Learning.* Towards Data Science. Retrieved May 1, 2021 from https://towardsdatascience.com/machine-learning-and-discrimination-2ed1a8b01038

[21] RStudio Team. 2021. *R Shiny: an R package that makes it easy to build interactive web apps straight from R.* RStudio, Boston, Massachusetts. https://shiny.rstudio.com/

[22] RStudio Team. 2021. *RStudio: An Integrated Development Environment (IDE) for R.* RStudio, Boston, Massachusetts. https://www.rstudio.com/

[23] R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[24] Richard Zemel, Yu (Ledell) Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. *Proceedings of the 30th International Conference on Machine Learning* (2013), 1–9. https://www.cs.toronto.edu/~toni/Papers/icml-final.pdf

8 FIGURES

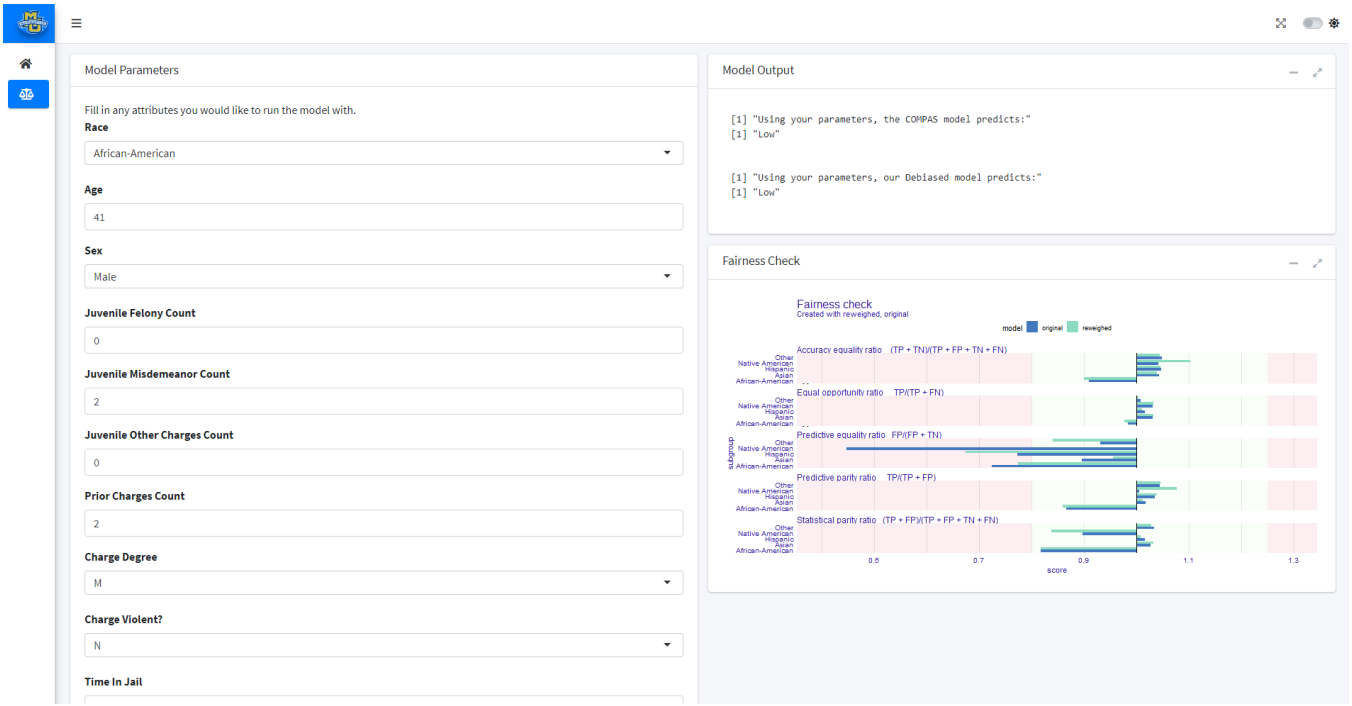

Figure 1: Model tab on R Shiny interactive Dashboard