**Investigating Bias in Decision-Making Software**

Creating a supervised learning model to analyse the

presence of algorithmic bias in recidivism software used to

make judicial decisions in the United States

Logo, company name

Description automatically generated

Megan E. Tennies

Candidate Number 181443

Supervisor Dr Viktoriia Sharmanska

Submitted in partial satisfaction of the requirements

for the degree of

BSc (Hons) Computer Science and Artificial Intelligence

School of Engineering and Informatics

University of Sussex

May 2022

## Declaration

This report is submitted as part requirement for the degree of Computer Science and Artificial Intelligence BSc at the University of Sussex. It is the product of my own labour except where indicated in the text. The report may be freely copied and distributed provided the source is acknowledged. I hereby give permission for a copy of this report to be loaned out to students in future years.

Signature

Megan E. Tennies

## Acknowledgements

## Summary

**Author** Megan E. Tennies

**Year** 2022

**Subject** Investigating Bias in Decision-Making Software

**Supervisor** Dr Viktoriia Sharmanska

**Abstract**

Automated decision-making software is becoming the norm in every sector form education and healthcare to banking and policing. As the use of algorithms and AI for decision-making increases, so does the presence of algorithmic bias. The basis of this project is to show the presence and impact of algorithmic bias in decision-making software, demonstrated using the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm. Given information about a currently incarcerated individual, COMPAS predicts a decile score from 1-10 which is their risk of recidivism.

The three objectives for this research and analysis are as follows:

* To show how the data used for training the algorithm is biased
* To show that using Race as a predictor for recidivism is not based in fact, and
* To show that COMPAS produces unfair predictions, which could be improved by using a range of bias mitigation techniques.

Objective 1 will be

Objectives 2 and 3 will be

Background research – types of bias, explanations

Programming – race as a predictor, unfair predictions

The results show

1. **Unfair / biased data used for training (research) COMPAS algorithm**
   1. **Can’t rely on race-based criminal statistics to infer criminal behaviour without considering the context in which those statistics arise**
   2. **Illustrates the difficulty / impossible task of making an unbiased algorithm – but it can be fairer (links to objective 3)**
2. **Incorrect to use race as a predictor for recidivism**
   1. **Criminal history is the strongest indicator – demonstrated using Random Forest of Trees for Feature Importance**
3. **COMPAS produces unfair predictions**
   1. **The application of bias mitigation algorithm increases each fairness metric**
   2. **Doesn’t pass the 80% the rule**

**Keywords** Algorithm, Algorithmic Bias, Recidivism

**Word** **Count**

**Pages**

## Table of Contents

Declaration ii

Acknowledgements iii

Summary iv

Table of Contents vi

List of Figures x

List of Tables xii

Nomenclature xiii

Chapter 1 14

Introduction 14

1.1 Project Aims & Objectives 14

1.2 Relevance 15

1.3 Target Group and Benefit of the Project 16

1.4 Resource Overview 16

Professional and Ethical Considerations 18

2.1 BCS Code of Conduct 18

2.2 Ethical Issues 19

Requirements Analysis 21

3.1 Mandatory Requirements 21

3.2 Desirable Requirements 22

Background Research & Existing Analyses 23

4.1 COMPAS, Software for Automating Risk 23

4.2 History of Risk Automation for Predictive Policing 24

4.3 Existing Analyses of Risk Automation Software 25

4.3.1 Desmarais, S. & Singh, J. (2013) 25

4.3.2 Skeem, J. & Lowenkamp, C. (2016) 25

4.3.3 Whiteacre, K. (2006) 25

4.3.4 ProPublica (2016) 25

4.3.5 Equivant (2016) 26

Racial Disparity in Policing and Sentencing 27

5.1 *‘Blacks have a higher rate of recidivism than whites’*: An Unsubstantiated Claim 27

5.1.1 Charge degree determined by race, not crime 28

5.1.2 Unlawful targeting by the police 29

5.1.3 Likelihood of arrest influenced by an individual’s race 30

5.1.4 Harsher sentencing 32

5.1.5 Poverty 32

5.1.6 Bail decisions and the inability to post bail 32

5.2 Identifying Sources of Bias in COMPAS 32

5.2.1 Biased training data (see 4.4) 32

5.2.2 Implicit biases 32

5.2.3 Explicit biases 33

5.2.4 Lack of diversity 33

5.2.5 No bias detection and/or mitigation 35

Methodology 36

6.1 Algorithmic Bias Mitigation 36

6.1.1 Demographic Parity 37

6.1.2 Statistical Parity Difference 38

6.1.3 Equality of Opportunity 38

6.1.4 Average Odds Difference 39

6.2 Supervised Learning Models 39

6.2.1 Model Input 39

6.2.2 Baseline Logistic Regression Model 39

6.2.3 Random Forest Classifier 40

6.2.4 Reweighing 40

6.2.5 Adversarial Debiasing 41

6.2.6 Prejudice Remover 42

6.2.7 Calibrated Equalised-odds Difference 43

6.2.8 Reject-option Classification 44

7.3 Feature Importance with Random Forest Classifier 49

7.5 Model Results using Bias Mitigation Techniques 52

7.5.1 Baseline Model 52

7.5.2 Model with Reweighing 53

7.5.3 Adversarial Debiasing Model 55

7.5.4 Prejudice Remover Model 57

7.5.6 Model with Calibrated Equalised-odds Difference 58

7.5.7 Model with Reject-option Classification 59

7.6 Comparison 61

7.7 Fairness Model 61

Testing & Evaluation 63

8.1 Model Evaluation 63

8.2 Quantitative Analysis 63

8.3 Model Cross-Validation (K-Fold) 63

Proposing Solutions 64

9.1 Ethical Datasets for Training 64

9.1.1 Using Equally Weighted Sample Sizes for Training 64

9.1.2 Representative Datasets 64

9.2 Using More Accurate Attributes to Predict Recidivism 64

9.2.1 Removing Race as an Attribute 64

9.2.2 Weighting Attributes 64

9.2.3 Including Dynamic Attributes 64

9.3 Testing for Bias 64

9.4 AI Fairness 360 for Bias Testing and Mitigation 64

9.5 Bias Intervention Checks 64

9.6 Transparency in the Decision-Making Process 64

9.7 Implicit and Explicit Bias Training 65

9.8 Increasing Diversity at Equivant 65

9.9 Legality 65

Conclusion 66

References 67

Appendices 70

Appendix A – Imprisonment rates of U.S. residents based on sentenced prisoners under the jurisdiction of state or federal correctional authorities, by jurisdiction, sex, and race or ethnicity, 2010-2020 70

Appendix B – Sentenced prisoners under the jurisdiction of state or federal correctional authorities, by jurisdiction, sex, and race or ethnicity, 2010–2020 71

Appendix C – Cumulative percent of state prisoners released in 24 states in 2008 who were arrested following release, by sex, race or ethnicity, age at release, and year following release 72

Appendix D – Results from all bias mitigation models 73

## List of Figures

[Figure 1. Racial distribution of the US Prison population in 2020 26](https://universityofsussex-my.sharepoint.com/personal/met28_sussex_ac_uk/Documents/FYP%20Report.docx#_Toc101884329)

[Figure 2. Imprisonment rates per 100,000 US residents for White and Black populations between 2010 and 2020 26](https://universityofsussex-my.sharepoint.com/personal/met28_sussex_ac_uk/Documents/FYP%20Report.docx#_Toc101884330)

[Figure 3. Cumulative re-arrest (recidivism) rates for US prisoners released in 2008 in the following decade, by race 27](#_Toc101884331)

[Figure 4. A graph to show the number of reported Stop-and-Frisks performed by the NYPD per year during 2002-2019 28](#_Toc101884332)

[Figure 5. Graphs comparing the proportion of Black individuals in NYC against the proportion of Stop-and-Frisks carried out on Black individuals by the NYPD in 2011 28](https://universityofsussex-my.sharepoint.com/personal/met28_sussex_ac_uk/Documents/FYP%20Report.docx#_Toc101884333)

[Figure 6. Graphs comparing the percentage of Black and White marijuana users in the American population with the arrest rates for Black and White individuals for marijuana possession between 2002-2010 30](#_Toc101884334)

[Figure 7. Graphs showing the percentage of women employed in Computing and Engineering between each decade between 1990-2019 34](https://universityofsussex-my.sharepoint.com/personal/met28_sussex_ac_uk/Documents/FYP%20Report.docx#_Toc101884335)

[Figure 8. Graphs showing the race distribution of employees in Computing and Engineering in 2019 34](https://universityofsussex-my.sharepoint.com/personal/met28_sussex_ac_uk/Documents/FYP%20Report.docx#_Toc101884336)

[Figure 9. Diagram illustrating where AIF360 Bias Mitigation algorithms and metrics will be introduced in the design pipeline 35](https://universityofsussex-my.sharepoint.com/personal/met28_sussex_ac_uk/Documents/FYP%20Report.docx#_Toc101884337)

[Figure 10. Adversarial Debiasing pipeline 41](https://universityofsussex-my.sharepoint.com/personal/met28_sussex_ac_uk/Documents/FYP%20Report.docx#_Toc101884338)

[Figure 11. Calibration of classifiers H\*1 and H\*2 plotted in the false-positive, false-negative plane for the two classifiers h1 and h2, for satisfying different cost constraints (FPR/FNR) 43](#_Toc101884339)

[Figure 3. Graphs to show the distribution of race, sex, charge degree, and age for individuals within the COMPAS dataset 46](https://universityofsussex-my.sharepoint.com/personal/met28_sussex_ac_uk/Documents/FYP%20Report.docx#_Toc101884340)

[Figure 4. Distribution of the number of prior crimes committed by individuals in the COMPAS dataset 47](#_Toc101884341)

[Figure 5. Confusion Matrix for the privileged group, White 47](#_Toc101884342)

[Figure 6. Confusion matrix for the unprivileged group, Non-white 47](#_Toc101884343)

[Figure 7. Feature importance for individual and grouped attributes using Mean Decrease in Impunity (MDI) and Feature Permutation (FP) 49](https://universityofsussex-my.sharepoint.com/personal/met28_sussex_ac_uk/Documents/FYP%20Report.docx#_Toc101884344)

[Figure 8. Results from the Reweighing model 52](#_Toc101884345)

## List of Tables

[Table 1. COMPAS attributes used for data analysis 20](#_Toc101884346)

[Table 2. Interpretation of COMPAS risk scores 23](#_Toc101884347)

[Table 3. Table showing the percentage and number of Black and White marijuana users in the United States in 2010 31](#_Toc101884348)

[Table 4. Table showing implicit biases which may be present in the decision-making process 33](#_Toc101884349)

[Table 1. Table showing the interpretation of Disparate Impact (DI) scores 38](#_Toc101884350)

[Table 2. Types of prejudice that can be present in ML algorithms 42](#_Toc101884351)

[Table 3. TPR and TNR for the privileged (White) and unprivileged (Non-white) groups 49](#_Toc101884352)

[Table 4. Difference in mean outcomes for the privileged and unprivileged groups for the training, validation, and testing data 49](#_Toc101884353)

[Table 5. Results from the baseline Logistic Regression model 52](#_Toc101884354)

[Table 6. Results from hyperparameter tuning for Reject-option Classification: Fairness metric constraints 60](#_Toc101884355)

[Table 7. Results from hyperparameter tuning for Reject-option Classification: Upper and lower bounds for the constrained fairness metric 60](#_Toc101884356)

## Nomenclature

*Ŷ = 1* Positive predictions

*Ŷ = 0* Negative predictions

*FP / FPR* False Positive / False Positive Rate

*FN / FNR* False Negative / False Negative Rate

*TP / TPR* True Positive / True Positive Rate

*TN / TNR* True Negative / True Negative Rate

*DI* Disparate Impact

*1 – DI*  1 – Disparate Impact, distance to Demographic Parity

*AOD* Average Odds Difference

*EOD* Equal Opportunity Difference

*TI* Theil Index

*ACC* (Balanced) Accuracy

*LR* Logistic Regression

*RW* Reweighing, Reweighed

*AD* Adversarial Debiasing

*PR* Prejudice Remover

*CEOD* Calibrated Equalised-odds Difference

*ROC* Reject-option Classification

## Chapter 1

## Introduction

Algorithms are adept at making calculated automated decisions as a result of data and statistical analysis and identifying patterns by finding new connections within a dataset. The process of automating decision-making, whether that’s in the context of social welfare, healthcare, credit decisions, employment screening, or criminal justice, was designed to take away the responsibility from human decision-makers. Automated decisions are thought to be rational, objective, and unbiased – a fairer alternative to decisions made by people. Despite the fact that algorithms are not inherently biased, the way in which they have been developed can inadvertently produce discriminatory results, which contradicts the principal intention for using an algorithm over a human decision-maker. These unfair outcomes caused by systemic errors within an algorithm are characterised as Algorithmic Bias.

Algorithmic bias can typically be divided into two categories – accuracy and impact. A decision-making algorithm can have different rates of accuracy across different demographical groups. Similarly, the algorithm can make considerably different decisions when applied to different populations. The US criminal justice system often relies on the use of automated decision-making tools to aid in the prediction of criminal risk. In terms of algorithmic bias within this context, these two concepts of accuracy and impact can translate to overpredicting risk in one racial group whist simultaneously underpredicting risk in another.

This report is an investigation into the presence of algorithmic bias within decision-making software, and the approaches which can be taken to mitigate bias in the design and application processes. It will give the reader a preliminary understanding of automating decision-making and sources and causes of algorithmic bias within the process. This is written in chapter 4, which also explores the history of risk automation in the US criminal justice system. Chapter 5 explores the statistics and data behind the COMPAS algorithm. Chapter 6 will formalise the implementation of the ML model and the application of approaches towards fairness. Chapter 7 will discuss the results from the models in terms of predictions and fairness metrics and provide an evaluation of the models. Proposed solutions to implement fairness in software will be outlined in chapter 8.

### 1.1 Project Aims & Objectives

The overall aim of this project is to investigate the presence of algorithmic bias within automated decision-making systems. The first of the primary objectives is to evaluate the use of automated decision-making over human decision-making, in the context of predicting recidivism[[1]](#footnote-2). Algorithms are used to increase efficiency and perform high-level data interpretation, but human decision-makers have the capability to take into consideration that people are complex and distinctive. COMPAS is a decision support tool used to assess the likelihood of recidivism of an offender once released from prison. The algorithm with COMPAS needs to be able to recognise differences between individual offenders and make individualistic decisions. Risk scores determined by the algorithm should be unique to each individual rather than following a pattern based on an individual’s identity. To evaluate the presence of bias within COMPAS and subsequently the use of decision-making software for predicting recidivism, the following questions will be answered –

* To what extent is COMPAS free of predictive bias?
* To what extent does COMPAS output different scores based on different factors such as race, gender, or age?
* Which risk factors are associated with the highest score disparities between racial groups, and are these risk factors associated with race?
* Can the disparities in the scores be explained by offenders’ criminal record or the type of crimes they were committed for?

To answer these questions, in-depth research will be done to find the possible sources and causes of bias within risk automation software, specifically looking at bias within COMPAS. Using this information, a supervised learning model will be created to make predictions of identity, i.e. race, gender, and age, from the risk scores produced by the COMPAS algorithm. This is to look for patterns among racial, gender and age groups to evaluate whether COMPAS is able to make individualistic decisions or whether decisions are made based on an individual’s identity.

The second of the primary objectives is to apply fairness metrics such as Bias Amplification (Chang, K. *et al*, 2017) to measure the rate at which different groups of individuals receive the same risk scores.

### 1.2 Relevance

Studying Computer Science, we are taught about how to create our own systems to perform tasks however we rarely speak about the possible malicious applications or implementations of making such software. Computer science professor Hao Li talks about how software such as Photoshop was designed for creative purposes but is used to manipulate the public through deepfakes. He goes on to explain that these deepfakes are being used to spread harmful misinformation, regardless of the fact that Photoshop was not created with this intention. Li warns about the ethical concerns of creating software that has inadvertently malicious applications and outcomes. Putting this into the context of our degree, we were tasked to build a model to perform facial recognition. Facial recognition is a controversial technology for which many ethical and privacy concerns have been identified. One example is that most facial recognition is more accurate for white, male faces than for people of colour or women. This results in more likely false positives for women, or more misidentification of men of colour, putting women of colour at the intersection of being most likely to be misidentified (Grother, Hanoka & Ngan, 2019). When we created our own facial recognition software, we did not test our systems for fairness across different groups of people. Lack of testing for fairness could overlook inaccurate and underperforming pieces of software, leading to the amplification of unconscious biases.

The motivation behind this project is to analyse COMPAS and apply different fairness metrics to show that well-intentioned systems can produce discriminatory and unjustified outcomes. Performing the analysis will illustrate the importance of putting AI ethics in practice, whether that is in regard to a university-level project, or a real-world software.

### 1.3 Target Group and Benefit of the Project

There are specific risk assessment tools that perform better at predicting recidivism in specific populations of offenders such as male versus female offenders, or younger versus older offenders. Each tool uses a different number of factors, derived from a variety of sources. Some use an offenders’ arrest report, a self-report, a report filed by a correctional officer, an interview, or a combination of these. This means that each risk assessment tool is going to perform differently and have different outcomes. Many different researchers have tested the validity and fairness of existing risk assessment tools, the majority of which have concluded that there is a disparity in accuracy between different sub-groups of offenders (Desmarais & Singh, 2013).

It is clear that the needs of the users (the offenders being scored) are not being met by the current risk assessment tools. This project will show the different predictive accuracies for different groups of users. Highlighting the disparity between risk scores and identifying why these could be occurring could contribute to an evolution in risk assessment tools to produce fairer outcomes. I believe that an ideal solution for the users would be a tool which is able to make individualistic decisions rather than basing decisions off of general data from a population of offenders who share a similar characteristic, such as race, gender, religion, or age.

### 1.4 Resource Overview

I will be using Anaconda to launch Jupyter Notebook. I’ve chosen to use a computational notebook such as Jupyter as it combines software code and computational and visual output with explanatory text. Jupyter will function as an effective presentation tool to show the data analysis I have performed on the COMPAS software. I have chosen to program in Python because it is an ideal language for data analysis. It allows for data collection and cleaning, data exploration, data modelling and data visualisation. Python has an extensive number of libraries for data analysis that I will use, such as –

* Scikit-Learn and TensorFlow for feature and model selection
* AIF360 (AI Fairness 360) for metrics and algorithms to mitigate bias in datasets and models
* Matplotlib and Seaborn for data visualisation
* NumPy for mathematical functions
* Pandas for loading and manipulating the .csv data files
* IPython for displaying comments, titles, and images.

# Chapter 2

## Professional and Ethical Considerations

### 2.1 BCS Code of Conduct

This research will be conducted in line with BCS Code of Conduct (The Chartered Institute for IT, 2021). Outlined below are the points from the Code of Conduct which relate to this project, and how they will be satisfied.

The data being used for analysis is a subset of the data COMPAS collects. Any identifiable information, or information that can be combined with another piece of information to identify an individual within the dataset is not being loaded for the analysis. This means that no confidential information will be disclosed (point 3.d[[2]](#footnote-3)) in order to keep the individuals’ anonymity to ensure their privacy and security (point 1.a[[3]](#footnote-4)). The data being analysed is from outside of the European Economic Area (EEA) as it has been collected from Broward County in Florida. Under the General Data Protection Act (GDPR), the University of Sussex must have a lawful basis for processing personal data. To be compliant with points 2.d[[4]](#footnote-5) and 3.a[[5]](#footnote-6), the lawful bases for handing the dataset being used for analysis have been satisfied identified as Public Task[[6]](#footnote-7) and Legitimate Interest[[7]](#footnote-8).

The content of this report may be upsetting to some third parties. In accordance with point 1.b[[8]](#footnote-9), the topic of algorithmic bias and discriminatory policing has been briefly introduced in the summary. This gives readers the chance to understand the potentially difficult themes which will be explored throughout the report.

In order to ensure that differing viewpoints will be respected (points 2.e[[9]](#footnote-10)) in regard to the use of automated decision-making software, the research carried out will also evaluate the use of alternative decision-making (i.e. by human decision-makers). Different organisations have claimed different opinions of COMPAS. ProPublica, a non-profit organisation, carried out analysis which demonstrated cases where COMPAS produced discriminatory results for non-White individuals. Contrastingly, the organisation that produced COMPAS (Equivant) carried out their analysis to reject the results of the ProPublica study. The method(s) and validity of both of these analyses will be considered before making any statements on bias within the software. The results and findings from analysis will be presented without any edits or omissions, to ensure that professional opinions of the research can be formed.

As per point 2.a[[10]](#footnote-11), b[[11]](#footnote-12) and c[[12]](#footnote-13), the author holds competency in Python, Data Analysis, and Predictive Modelling having studied modules at the University of Sussex such as –

* Natural Language Engineering and Advanced Natural Language Engineering
* Fundamentals of Machine Learning
* Acquired Intelligence & Adaptive Behaviour.

These existing competencies, along with the awareness of technical knowledge regarding the research area of algorithmic bias, will be developed by undertaking this project.

### 2.2 Ethical Issues

The ethical issue surrounding my project is the use of criminal information data – as using the data lacks informed consent, and it includes potentially identifiable data. The justification for not acquiring informed consent for this research is that the data being analysed is that the dataset is in the public domain, and available to anyone who wishes to read or use it. Florida – the state which the data is from – has open record laws which allowed ProPublica to obtain two years’ worth of COMPAS scores through a public records request to the Broward County Sheriff’s Office. They linked this data with public incarceration records from the Florida Department of Corrections to produce a dataset with the records of individuals and their recidivism risk score. To ensure that the analysis does not comprise of any potentially identifying information, the data being used for analysis is from the AIF360 COMPAS dataset[[13]](#footnote-14) which is a subset of the dataset used for COMPAS. As explained in 2.1, the individuals within the analysis will not be able to be identified, as the identifiable attributes such as date and place of birth, or place and time of arrest which COMPAS takes, are not contained in the dataset. Given that only 6 out of 402[[14]](#footnote-15) attributes for each individual will be loaded, which do not directly identify an individual in themselves and are extremely unlikely to be able to identify individuals when combined with other information. Additionally, the database relates to over 6,000 individuals, meaning that the data being used does not amount to personal data. The following table lists the 6 unidentifiable attributes included in the AIF360 COMPAS dataset which will be used for analysis.

|  |  |
| --- | --- |
| Gender | Categorical variable indicating the individuals’ gender – Male or Female |
| Age | Categorical variable categorising the individual into three groups based on age – x > 25, 25-45, or x < 45 |
| Race | Categorical variable representing the individuals’ race – Black[[15]](#footnote-16), White, Asian, Hispanic[[16]](#footnote-17), Native American/Native Alaskan, or Other[[17]](#footnote-18) |
| Priors Count | Integer representing the number of prior crimes committed |
| Charge Degree | Character representing the degree of the original criminal charge – M (misdemeanour) or F (felony) |
| Two-Year Recidivism Score | Binary variable of whether the individual re-offended within two years of release – 0 or 1 |

Table 1. COMPAS attributes used for data analysis

# Chapter 3

## Requirements Analysis

This chapter will lay out the essential (mandatory) and non-essential (desirable) requirements of the project. \*NEEDS TO BE UPDATED

### 3.1 Mandatory Requirements

|  |  |
| --- | --- |
| **M.1** | In-depth research into the history of automated decision-making software:   * Application of risk automation software * Automating risk leads to automating bias * Sources and causes of bias * Lack of transparency in the decision-making process |
| **M.2** | Appropriate data pre-processing techniques that extract only the required fields from each criminal record for analysis, to:   * Keep the identities of the offenders confidential in alignment with Data Protection practice. This will be achieved by only loading 6 out of 402 attributes from each criminal record, ensuring anonymity. * Data-smoothing to remove noise from the dataset in order to clearly identify patterns or correlations in the data * Case-deletion to remove any records with missing values from the dataset, as using any method of value substitution would be making false and unjustified assumptions |
| **M.3** | Creating a supervised learning model to analyse the outputs of COMPAS |
| **M.4** | Implementing AI Fairness approaches:   * Bias Amplification * AIF360 Algorithms * Performance Measures |
| **M.5** | Testing and Validation of the model:   * Testing and training data * Hyperparameter Tuning |
|  | * K-Fold Cross Validation * Accuracy |
| **M.6** | Evaluation of the model:   * Model evaluation * Identifying correlations in the outputs * Qualitative analysis against actual recidivism rates in Florida |
| **M.7** | Proposed solutions for avoiding bias in predictive recidivism software |

### 3.2 Desirable Requirements

|  |  |
| --- | --- |
| **D.1** | Research into bias mitigation techniques |
| **D.2** | Research into alternatives to automated risk assessment tools such as COMPAS |
| **D.3** | Using more attributes from the original dataset to find other possible biases from COMPAS  \*Reliant on resubmitting an ethical approval application receiving approval which may not be possible as adding more attributes may make individuals within the dataset identifiable |

# Chapter 4

## Background Research & Existing Analyses

### 4.1 COMPAS, Software for Automating Risk

The US judicial system often relies on the use of algorithms in the form of actuarial risk assessment instruments (ARAIs) to inform decisions regarding crime prediction, crime reduction strategies, and sentence processing. Prominent examples of these are PredPol, CompStat, and COMPAS. These tools were designed with the goal of reducing crime, lowering imprisonment rates, or predicting recidivism and promoting desistence[[18]](#footnote-19). However, their implementation and applications are seen as controversial. One common critique being that the algorithms are built using biased datasets and will therefore output biased results.

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a risk assessment and decision support tool developed by the consulting and research firm Equivant. It is used within the judicial system in the United States to determine the likelihood of an offender committing a future crime upon release from prison. The algorithm takes information from a 137-item questionnaire about a defendant’s background, such as their level of education, socio-economic background, or whether the defendant has a job. From this information, a decile score between 1 and 10 is derived which rates their risk of recidivism. Decile scores are obtained by ranking the scores obtained for a group of individuals in ascending order and splitting them into ten equal sized groups. The decile scores range from 1 (lowest) to 10 (highest). A score of 1 indicates an individual’s risk score is in the lowest 10% of all scores in the group and a score of 10 an individual’s risk score is in the highest 10% of all scores in the group (Equivant, 2015). The scores can be interpreted as follows:

|  |  |
| --- | --- |
| 1 – 4 | Low risk in relation to other offenders |
| 5 – 7 | Medium risk in relation to other offenders |
| 8 – 10 | High risk in relation to other offenders |

Table 2. Interpretation of COMPAS risk scores

The basis of using this tool is that if an algorithm were to be able to accurately predict whether an individual is likely to reoffend upon release, the criminal judicial system could be fairer (Grove et al., 2000). In theory, this could also reduce overincarceration rates and prison populations by selectively deciding who is incarcerated and the length of their sentence. Using an algorithm to determine a recidivism score raises one significant question – what happens if the predictions are wrong? If the algorithm predicts a high score incorrectly, the result could be an offender receiving a harsher sentence or a longer parole date, leaving the offender in prison for an unequitable time. If the algorithm predicts a low score incorrectly, a potentially dangerous offender would be released from prison, free to commit another crime which will ultimately increase recidivism and imprisonment rates – undermining the justification for using COMPAS. If the algorithm arrives at these decisions through biased methods, the bias within crime statistics regarding recidivism and desistance will be amplified.

### 4.2 History of Risk Automation for Predictive Policing

Predictive policing as a way to analyse whether offenders with a shared certain characteristic are more likely to commit a crime has been researched by criminologists at length. As a result, risk assessment tools based on statistics were developed as effective methods to predict an individual’s chances of offending or reoffending (Quinsey et al, 1998; Quinsey et al. 2006). Historically, risk prediction began with educated guesses regarding identifying factors that were believed to be corelated with higher rates of crime and recidivism and evolved into evidence-based predictions and models. The first generation[[19]](#footnote-20) of risk assessment tools were based off of professional judgement alone, using clinical or correctional staff’s personal experiences and training to make judgements of an individual’s characteristics and behaviour to assess their risk of recidivism. These assessments that were subject to human error, subjectivity to biases, and unempirical outcomes, and were therefore widely critiqued for their poor predictive accuracy.

Second generation[[20]](#footnote-21) risk assessment tools relied on evidence-based numerical predictions using static risk factors – such as history of drug use, age of their first offence and prior criminal history – that had been statistically linked to a higher risk of recidivism, as well as static historical factors such as the individual’s gender and age, all which are not amenable to change. These predictions were based off of the analysis from paroled individuals and were thought to be more objective and evidence-based than the previous generation of assessments. However they were critiqued for their ‘fixed’ prediction of risk which failed to recognise the dynamic changes in an individual’s behaviour, attitude, characteristics and needs over time.

The third generation[[21]](#footnote-22), often referred to as ‘risk-need-responsivity’ instruments, operated under the assumption that risk prediction needed to incorporate both the static and dynamic factors linked to criminal behaviour in order to make assessments. This is because dynamic risk factors such as problem-soling ability or temper control are able to change over a long period of time, which incarceration provides. This generation of models worked under the presumption that correctional facilities were effective at rehabilitation and positively changing factors in an offender’s life. The argument in support of the effectiveness of these models was that if rehabilitation and correctional treatment were responsive to an individual’s dynamic needs, the risk of recidivism within the population would decrease. This theory, also seen in fourth generation[[22]](#footnote-23) tools, de-individuals the risk assessment process by putting offenders into distinct categories regarding their unalterable characteristics, such as age, race, gender, or religion. From this, it can be inferred that the resulting recidivism risk scores from third- and fourth-generation tools are made on the basis of who an offender is rather what an offender has done. Assessments based on the history of a population of incarcerated individuals are inevitably going to encompass a varying amount of bias, largely because they are not an accurate representation of the population outside of prison. Both past and present arrest and sentencing patterns disproportionally target Black, Asian, Hispanic, Native-American, and other non-White individuals, which skews the demographics of the prison population and alters statistics regarding characteristics such as race, and the corresponding rate of crime and recidivism.

The wrongful assumption of causation from correlation in datasets in terms of assessing risk between different groups of people encourages discriminatory biases within risk assessment tools. One of the biggest criticisms of risk assessment tools is that they are programmed using biased data which incorrectly assumes links between factors such as the race of an individual with the risk of recidivism. This is a type of attribution bias. One way that the attributions may be biased is from making a mistake known as the group attribution error[[23]](#footnote-24). This refers to people’s assumption that characteristics of an individual within a group are reflective of the group as a whole. Another explanation is the fundamental attribution error[[24]](#footnote-25) which is a tendency to categorise an individual’s actions and behaviours on a person’s characteristics rather than on social or environmental situations that influence the individual’s actions.

4.3 Existing Analyses of Risk Automation Software \*NEEDS TO BE WRITTEN OR CUT FROM REPORT DEP. ON WORD COUNT

#### 4.3.1 Desmarais, S. & Singh, J. (2013)

#### 4.3.2 Skeem, J. & Lowenkamp, C. (2016)

#### 4.3.3 Whiteacre, K. (2006)

#### 4.3.4 ProPublica (2016)

ProPublica, an independent non-profit organisation that conducts investigative journalism in the interest of the general public, launched an examination into the study of recidivism risk scores (Anwin *et al.* 2016). In their analysis of COMPAS, they illustrated that COMPAS was using datasets that pushes the narrative that Black defendants are more likely to reoffend. This was because statistically there are more Black people in prison than White people in the US. This is an example of where attribution error can be seen in the software – attributing the higher proportion of Black individuals in prison to a shared characteristic rather than looking at the societal forces which result in a higher proportion of Black individuals being imprisoned than their White counterparts (NAACP, 2021). It fails to take into consideration any context(s) which results in this statistic, two key examples being discrimination by law enforcement and socio-economic factors.

#### 4.3.5 Equivant (2016)

After ProPublica’s report was released, publicly criticising the validity of COMPAS in terms of producing discriminatory results on the basis of race, Equivant produced a counter-report rejecting ProPublica’s conclusions. The rationale for this being –

ProPublica focused on classification statistics that did not take into account the different base rates of recidivism for blacks and whites. When the correct classification statistics are used, the data do not substantiate the claim of racial bias towards blacks. The proper interpretation of the results in the samples used by ProPublica demonstrates that the General Recidivism Risk Scale (GRRS) and Violent Recidivism Risk Scale (VRRS) are equally accurate for blacks and whites. (Equivant, 2016)

The issue with this claim is that recidivism statistics for Black individuals are skewed heavily by racism and don’t accurately or fairly represent a Black person’s likelihood to commit a crime.

# Chapter 5

## Racial Disparity in Policing and Sentencing

### 5.1 *‘Blacks have a higher rate of recidivism than whites’*: An Unsubstantiated Claim

In 2016, then-Northpointe created a report refuting ProPublica’s claim of racial bias in COMPAS. Their statement – *‘We also point out that in comparison with blacks, whites have much lower base rates of general recidivism, and violent recidivism’* – is misleading. They are equating higher rates of recidivism and imprisonment for Black individuals with a higher tendency to commit subsequent crimes.

Figure 1. Racial distribution of the US Prison population in 2020

Statistically, they can back up this claim as there are more Black than White individuals in prison (Figure 1), or because Black individuals have a much larger imprisonment rate per 100,000 than White individuals (Figure 2).

Figure 2. Imprisonment rates per 100,000 US residents for White and Black populations between 2010 and 2020

Or, because Black individuals have marginally higher rates of recidivism (Figure 3).

Figure 3. Cumulative re-arrest (recidivism) rates for US prisoners released in 2008 in the following decade, by race

But what is imperative to consider is *why* these statistics exist. It’s easy – but irresponsible – to assume that the reason behind these statistics is that Black people are more likely to commit crimes than White people. The following sections will illustrate how criminal and sentencing data can be misinterpreted without consideration of context to support the above statement from Equivant, by exploring case studies where Black individuals received unfair outcomes due to their race.

#### 5.1.1 Charge degree determined by race, not crime

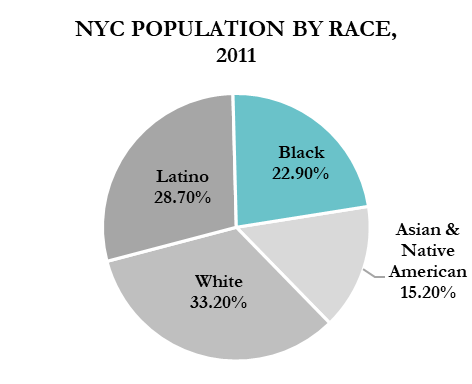
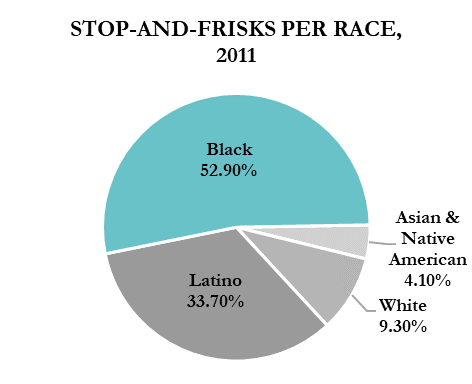
\*NEEDS TO BE WRITTEN OR CUT FROM REPORT DEP. ON WORD COUNT

#### 5.1.2 Unlawful targeting by the police

The NYCLU produced a report analysing crime statistics and the rate of Stop-and-Frisks from the NYPD during the de Blasio era (NYCLU, 2019). A Stop-and-Frisk is the practice of temporarily detaining for questioning a civilian on the street, often accompanied by searching their person for weapons or other contraband. The NYCLU illustrated that members of the NYPD were targeting individuals with the Stop-and-Frisk program on the basis of the individuals’ race. Since 2002, the NYPD reported stopping 5,174,072 individuals, the distribution as follows –

Figure 4. A graph to show the number of reported Stop-and-Frisks performed by the NYPD per year during 2002-2019

Looking at the above graph, the number of Stop-and-Frisks peaked in 2011 at 685,724 individuals being stopped. The sharp decline of stops in the following years is a result of a trio of class-action lawsuits filed against the NYPD, the most relevant of which being Floyd v. City of New York[[25]](#footnote-26). This lawsuit asserted that the NYPD’s Stop-and-Frisk practices were a culmination of racial profiling and the unconstitutional stopping of pedestrians. Below are two charts which visualise the racial profiling and targeting by the police, as outlined in the class-action lawsuits. Using the data[[26]](#footnote-27) from 2011 preceding Floyd v. City of New York, the below graphs illustrate the disparity in the racial distribution of the NYC population against the races of the individuals who get stopped by the police.



Black individuals make up 22.9% of the NYC population, however they make up 53.9% of Stop-and-Frisk cases. In comparison, White individuals make up 33.2% of the NYC population but only 9.3% of cases. If individuals were stopped due to factors other than race, then the proportion of the stops would be consistent with the population. As the percentage of Black individuals being stopped by police is over double the percentage of Black individuals within the population, it can easily be concluded that Black individuals (and other non-White individuals) are overwhelming and unfairly targeted by police activity.

Figure 5. Graphs comparing the proportion of Black individuals in NYC against the proportion of Stop-and-Frisks carried out on Black individuals by the NYPD in 2011

Not every individual who is stopped by police in the Stop-and-Frisk practice is frisked – the officer is required to have reasonable suspicious that the individual has a weapon that poses a threat to the officer’s safety. However, of the 66% of stops that led to frisks, over 93% results in no weapon being found. Out of the 66% of stops that led to frisks, Black and Latino people were more likely to be frisked than White people, however when frisked were less likely to be found with a weapon. This is a clear example of racial profiling by law enforcement – more Black and Latino individuals are targeted by the police, despite the fact that they pose no more risk to the safety of the officers or general publics than White individuals.

#### 5.1.3 Likelihood of arrest influenced by an individual’s race

In 2013, the ACLU[[27]](#footnote-28) produced a report comparing marijuana possession rates[[28]](#footnote-29) by race for all 50 states in America. The graph below shows the disparity in arrest rates for possession of marijuana for Black and White individuals over a decade, even though marijuana usage for both racial groups is comparatively the same.

Figure 6. Graphs comparing the percentage of Black and White marijuana users in the American population with the arrest rates for Black and White individuals for marijuana possession between 2002-2010

Looking at the latest data for 2010, Black people are 3.73 times more likely to be arrested for possession, with an arrest rate of 716 per 100,000, compared to the arrest rate for white people which is 192 per 100,000. Combining this with population data for 2010 (U.S. Census Bureau, 2021), the rough number of White marijuana users is almost 4.5 times more than the number of Black marijuana users.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Population | Percentage of Users | Number of Users |
| White | 196,817,552 | 9.1 | 17,910,367 |
| Black | 37,685,848 | 10.7 | 4,032,286 |

Table 3. Table showing the percentage and number of Black and White marijuana users in the United States in 2010

With this knowledge, it can be inferred that the higher rate of arrests for possession of marijuana for Black people is not due to more Black people smoking or possessing the drug. Just like how the NYPD disproportionally targets Black people with Stop-and-Frisks, law enforcement throughout the U.S. target Black marijuana users.

#### 5.1.4 Harsher sentencing

\*NEEDS TO BE WRITTEN OR CUT FROM REPORT DEP. ON WORD COUNT

#### 5.1.5 Poverty

\*NEEDS TO BE WRITTEN OR CUT FROM REPORT DEP. ON WORD COUNT

#### 5.1.6 Bail decisions and the inability to post bail

\*NEEDS TO BE WRITTEN OR CUT FROM REPORT DEP. ON WORD COUNT

These are examples of situations which causes more arrests of Black people than their White counterparts for the same crime, which leads to disproportionally increased number of Black individuals in prison (NAACP, 2021). The COMPAS algorithm assumes that this higher proportion is due to the fact that Black people are pre-disposed to offend. This is racially biased and untrue. Perpetuating racial stereotypes such as this will result in the algorithm determining recidivism scores based on bias rather than objective facts and statistics regarding recidivism. If the US judicial system continues to use ARAIs such as COMPAS to aid sentencing, the systems and the predictions outputted must be tested for fairness and bias. Otherwise, algorithmically biased systems which produce discriminatory results will be implemented and relied upon to make significant decisions about an individual’s life.

### 5.2 Identifying Sources of Bias in COMPAS

#### 5.2.1 Biased training data (see 4.4)

\*NEEDS TO BE FINISHED

#### 5.2.2 Implicit biases

Two sources of bias have been identified above – the fundamental attribution error and group attribution error. These are both examples of Implicit Bias[[29]](#footnote-30), which is where biased attitudes about a particular group of people are operating without any conscious unawareness. Although they are conveyed indirectly and often without malicious intent, they still cause harm when present in the design and application of decision-making software. Below is a table of three implicit biases that are likely to appear in the automated decision-making process.

|  |  |
| --- | --- |
| Attribution Bias | The tendency to incorrectly reason about the cause of others’ behaviours – exhibited most commonly when explaining the behaviours of different groups of people than ones’ own |
| Confirmation Bias | The tendency to search for or interpret information in a way that supports ones’ own preconceptions or beliefs |
| Bias Blind Spot | The tendency to recognise the impact of bias within others’ judgement while failing to recognise the impact on ones’ own judgement |

Table 4. Table showing implicit biases which may be present in the decision-making process

#### 5.2.3 Explicit biases

Explicit bias[[30]](#footnote-31) refers to biased attitudes that a person is consciously aware of and voluntarily activates. These attitudes are prejudiced and discriminatory against a group (or groups) of people. Even though explicit biases are conveyed voluntary and easy to categorise, they are hard to identify within a system, often because information about the design of an algorithm, what data it uses, or how it’s used are not publicly disclosed. Explicit biases are often discovered in the outputs of a decision-making system, when an algorithm is making vastly different decisions for one group of people compared to another, or when an algorithm cannot perform accurately on a certain group.

#### 5.2.4 Lack of diversity

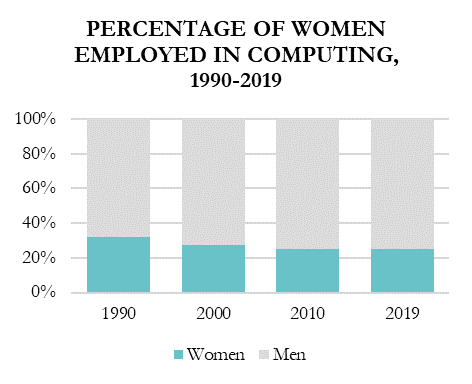
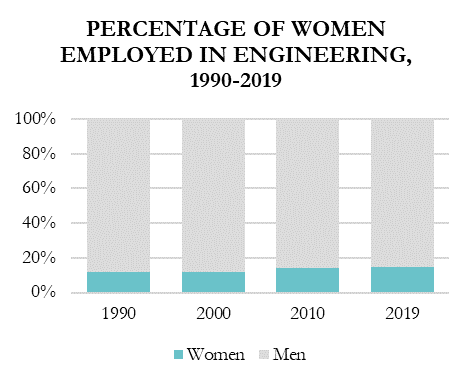
In 2014, Amazon developed an AI Résumé Screening Tool designed to score job candidates on a scale from one to five stars. It would rank a number of resumes against each other and output the top five candidates as a hiring recommendation. Amazon’s models were trained to vet potential candidates using data from resumes submitted to company over the previous 10 years, and whether the candidate from the resume in question received a job offer. The models were taught to recognise over 50,000 terms which appeared on successful resumes. The algorithms learnt to assign a low significance to skills found on the majority of resumes, such as fast typing speed or fluency in Microsoft Excel. Skills like programming in 5+ languages would be ranked higher. Most of the resumes the model used for training came from men. Consequently, the model learnt to favour male candidates over female candidates, as more men were successful in receiving a job offer. This resulted in three things –

* It ranked graduates from all-women’s colleges lower than graduates from gender non-specific colleges
* It penalised resumes which included the word ‘women’ – i.e. ‘women’s football coach’ or ‘women’s representative’
* It favoured resumes where candidates had described their work experience with more masculine language (verbs more commonly found on male applicants resumes), i.e. ‘executed’ or ‘captured’

As a result, the AI would repeatedly recommend women as less preferable candidates to men. The way that the algorithm learnt to favour men was through the training data – more resumes came from men than from women, a reflection of the male dominance within the tech industry. In the same way that Amazon’s AI Résumé Screening Tool produced better outcomes for men and worse outcomes for women, Equivant’s COMPAS produces better outcomes for White people and worse outcomes for Black people. Comparatively, racial groups such as Black, Asian, Hispanic, and other non-White groups are vastly underrepresented in the tech industry to the same extent as women.

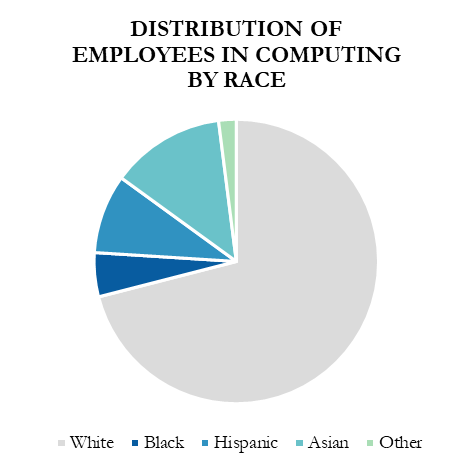
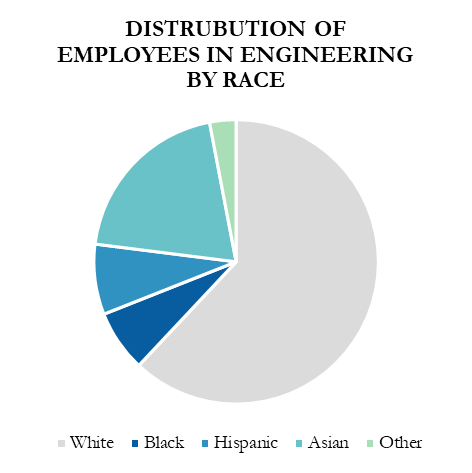
The Pew Research Centre produced a report after analysing federal government data[[31]](#footnote-32), to show how minorities are underrepresented in the tech workforce, a statistic which has remained unchanged for decades.

Figure . Graphs showing the percentage of women employed in Computing and Engineering between each decade between 1990-2019



The lack of diversity – both gender and racial – is an important factor to look at when analysing the sources and causes of algorithmic bias. How can we know that an algorithm designed by one group of people is going to work in favour of another?

Figure . Graphs showing the race distribution of employees in Computing and Engineering in 2019



\*NEEDS TO BE FINISHED

#### 5.2.5 No bias detection and/or mitigation

\*NEEDS TO BE FINISHED

# Chapter 6

## Methodology

This chapter will explore the methodology behind objectives 2 and 3, illustrating that COMPAS produces unfair and biased predictions, and uses race as a criminality attribute incorrectly. Supervised learning models will be created to show unfairness in the predictions of COMPAS using a variety of bias mitigation techniques and fairness metrics.

### 6.1 Algorithmic Bias Mitigation

Algorithmic bias mitigation is a supervised deep learning task where the goal is to predict an output variable, *Y* given an input variable or set of variables, *X* while remaining unbiased with respect to variable, *Z*. The predictor, *Ŷ = f(X)* is constructed of *(X, Y, Z)* tuples, where:

* *X* is the input features – age category, priors count, sex, and charge degree
* *Y* is the prediction of recidivism
* *Z* is the protected attribute – race.

Prediction *Y* is binary, the favourable label being *Y = 0* and unfavourable label being *Y = 1*, as the favourable outcome that we want from the COMPAS algorithm is ‘No Recidivism’. The output from a predictor can be judged using fairness metrics to determine the level and type of bias present in the classifier. There are multiple forms of bias which can be present in classification algorithms – prejudice, underestimation, and negative legacy. Prejudice implies a statistical dependence between the sensitive (protected) variable, *Z* and the target variable, *Y* or non-sensitive input variable, *X*. This can be in the form of direct prejudice, where the sensitive variable is used in a prediction model, or indirect prejudice, where there is a statistical dependence between the sensitive attribute and target variable. As race is used as an attribute in the COMPAS model, it can be deduced that the classification results depend on the race of the individual, thereby generating predictions that have direct discrimination. Using Equivant’s own description of the COMPAS algorithm[[32]](#footnote-33), the model also contains indirect discrimination, as predictions contain a statistical correlation with an individual’s race. This shows that the predictions depend on sensitive information.

The AI Fairness 360 (AIF360) Toolkit (Bellamy *et al.* 2018) is an open-source Python toolkit for detecting, understanding, and mitigating algorithmic bias. AIF360 will be used to create a set of models that use a range of pre-, in-, and post-processing techniques to mitigate bias. Figure 9 shows where the AIF360 algorithms and explainers will be introduced in the supervised learning pipeline.

Figure . Diagram illustrating where AIF360 Bias Mitigation algorithms and metrics will be introduced in the design pipeline

Raw

Data

Data

Pre-processing

Training

Data

Testing

Data

Classifier

Deployment

*Training*

*Testing*

**Dataset**

**Metric**

**Dataset Metric Explainer**

**Classifier Metric Explainer**

**Classifier**

**Metric**

**In-processing**

**Algorithm**

**Pre-processing**

**Algorithm**

**Post-processing**

**Algorithm**

To be able to determine bias in the models, a range of fairness metrics can be applied to the dataset, model, and predictions. The following definitions of fairness which have been defined by Hardt *et al.* (2016) and refined by Beutel *et al.* (2017) will be used for each model.

#### 6.1.1 Demographic Parity

Demographic parity indicates that a predictor is free of bias if the predictions, *Y* are independent of protected attribute, *Z* so that:

Disparate Impact (DI) is a measure of discrimination, which can be defined as the deviation from Demographic Parity. Given the privileged class, *Z = 1* and unprivileged class, *Z = 0*, DI can be measured as:

Which is the proportion of favourable outcomes between the privileged and unprivileged classes. The values of DI can be interpreted as follows:

|  |  |
| --- | --- |
| **DI < 1** | The privileged class, Z *= 1* has a higher proportion of predicted positive outcomes than the unprivileged class, Z *= 0* – this is **negative bias** |
| **DI = 1** | The proportion of individuals within each group of the protected class receive positive outcomes at the same rate – this is **demographic parity** |
| **DI > 1** | The unprivileged class, Z *= 0* has a higher proportion of predicted positive outcomes than the privileged class, Z *= 1* – this is **positive bias** |

Table . Table showing the interpretation of Disparate Impact (DI) scores

Where DI is equal to 1, demographic parity has been achieved and the model can be considered free of predictive bias. To be able to class predictions as fair, the models will be compared to the 80% as a fairness threshold. The 80% rule[[33]](#footnote-34) originated as a way to help companies determine if their company was using discriminatory hiring practices. The rule stated that companies needed to be hiring from a protected group at a rate that is 80% of the unprotected group. For example, if a company hired 50 men *(Z = 1)* and 20 women *(Z = 0)*,DI could be used as a measure to judge whether the company had a discrimination problem. This hiring ratio would be calculated by , or 40%. As this is lower than the 80% threshold, the company would need to provide a legitimate reason as to why women (individuals from the protected group) were being hired at a lower rate.

#### 6.1.2 Statistical Parity Difference

Statistical Parity Difference (SPD) is the same calculation as DI, however instead of being represented as a ratio it is the difference between the number of favourable outcomes for each group, as such:

#### 6.1.3 Equality of Opportunity

A predictor, *Ŷ* satisfies Equality of Opportunity for predictions, *Y* with respect to class *y*, if *Ŷ* and *Z* are conditionally independent on *Y = y* as such:

Compared to demographic parity, equality of opportunity allows prediction, *Ŷ* to depend on *Z*, but only through the target variable *Y*. The metric Equal Opportunity Difference (EOD) can be measured by deviation from Equality of Opportunity, as:

Which is the also difference in *TPR* between the privileged and unprivileged classes. A value of zero for EOD implies that each class has equal benefit, and the model can be considered fair.

#### 6.1.4 Average Odds Difference

Average Odds Difference (AOD) is the average of the difference in false positive rates *(FPR)* and True Positive Rates *(TPR)* between the privileged and unprivileged classes, as:

A value of 0 for AOD implies that each class has equal benefit, and the model can be considered fair.

### 6.2 Supervised Learning Models

#### 6.2.1 Model Input

The COMPAS dataset used for analysis contains records for 5,278 individuals, along with their age category, race, sex, number of prior offences, charge degree, and two-year recidivism score. The categorical data has been encoded using a One-Hot Encoder, which splits the attributes into 10 input features, *X*:

* age\_cat = 25 to 45, age\_cat = Greater than 45, and age\_cat = Less than 25
* priors\_count = 0, priors\_count = 1 to 3, and priors\_count = More than 3
* c\_charge\_degree = F and c\_charge\_degree = M
* sex = Male and sex = Female.

The protected attribute for the data is race, where:

* The privileged class, *Z = 1* is White
* The unprivileged class, *Z = 0* is Non-White.

To feed the dataset into each classifier, it has been split into training (70%), testing (15%) and validation (15%) sets.

#### 6.2.2 Baseline Logistic Regression Model

The first model is a Logistic Regression (LR) classifier, which takes the input variables, *X* and predicts the two-year recidivism score, *Y*. LR is the most appropriate technique for prediction as it exhibits the relationship between the response variable – *two\_year\_recid,* and the multiple predictor variable – *race*, *age\_cat*, *sex*, *priors\_count*, and *c\_charge\_degree*, which can be used to determine discrimination. This model will be used as the baseline for comparison to see how the different bias mitigation techniques have increased fairness.

#### 6.2.3 Random Forest Classifier

In order to determine how important each attribute in the dataset is for calculating an individuals’ risk of recidivism, feature importance using a random forest classifier will be used. The LR model described above makes predictions which are a weighted sum of the input variables (coefficients). These coefficients can then be used for finding individual feature importances for classification.

To do this, the input data is scaled using a Standard Scaler. The Random Forest Classifier (RFC) is then fit on the dataset to get the coefficient property which contains the coefficients found for each variable. Our task is binary classification, so we have class *1* (no recidivism) and class *0* (recidivism). The coefficient values which are positive indicate that the feature predicts the positive class, *1*, and the negative coefficient values indicate that a feature predicts the negative class, *0*. The two importance metrics that will be used to investigate the bias in variable importance are Mean Decrease in Impurity (MDI[[34]](#footnote-35)) and Feature Permutation Importance (FPI).

MDI is categorised as the improvement in the ‘Gini gain’ splitting criterion, which incorporates a weighted mean of the individual trees’ improvement in the splitting criterion produced by each variable. The Gini impurity index, G, is defined as:

where *nc* is the number of classes in the target variable and *pi* is the ratio of the given class. The decrease in impurity, *I*, is then defined as:

Where *G* is the Gini index for each node and *P* is the proportion of the data each split takes relative to the parent node. MDI is then calculated by taking the mean of these values.

FPI works by considering a variable important only if it has a positive effect on prediction accuracy. First, prediction accuracy is calculated on the predictions from the RFC. Any association between the variable of interest, *Xi* and the prediction *Ŷ* is broken by permuting the values of all the observations for *Xi*. Prediction accuracy is then recalculated. The difference between the two accuracy scores is known as the permutation importance of *Xi*. Taking the average of all of the importances gives the permutation importance of the variable. The positive and negative results of FPI can be interpreted the same as with MDI.

#### 6.2.4 Reweighing

Reweighing (Kamiran & Calders, 2012) is a pre-processing technique whereby the *(X, Y, Z)* tuples in the training dataset are reweighed. This gives cases where the protected attribute, *Z* that predicts the positive outcome, a heavier weighting. The reweighed dataset is then used as the input for the Logistic Regression classifier, which utilises the updated weights in the cost function. The goal of Reweighing is to learn a classifier which optimises accuracy but does not exhibit any discrimination in the predictions for the testing set.

Using the reweighing technique, the performance of classifier *Ŷ* is measured by the trade-off between high accuracy and low discrimination. The training dataset, *D* contains sensitive attribute, *Z*, which is ‘Race’, with domain *Z* *∈* {0, 1} which are ‘Non-White’ and ‘White’ respectively. The non-sensitive attributes are defined as *X* – but it is important to note that X may correlate with Z. We have a binary classification problem, with target attribute, *Y*, domain *Y* *∈* {0, 1}, where *Y* = 0 is the favorable and *Y* = 1 is the unfavorable label. The discrimination of the classifier is defined as:

*P (Ŷ = 0|Z = 1) – P (Ŷ = 0|Z = 0)*

A value greater than 0 means that the privileged class has a higher chance of being assigned the positive label by the classifier than the unprivileged class does. To try to mitigate this bias within the training dataset and classifier, tuples *(X, Y, Z)* are assigned weights. This makes the dataset discrimination free with respect to *Z* without having to change the labels.

If dataset *D* is unbiased, sensitive attribute *Z* and prediction *Y* would be statistically independent, with expected probability as:

However, the observed probability may be different:

If *Pexp* is greater than *Pobs*, it shows that the there is bias towards the unprivileged class, *Z* = 0 as they are less likely to be receiving the favourable label. To compensate for this bias, the objects in the unprivileged class which received the favourable outcome, i.e. Non-White individuals who did not reoffend, are given a higher weighting. Weights, *W* are calculated as follows:

The calculated weights are then used to fit the LR model.

#### 6.2.5 Adversarial Debiasing

Diagram

Description automatically generatedAdversarial Debiasing (Zhang *et al*. 2018) is an in-processing algorithm which reduces evidence of the protected attribute *Z* in predictions *Ŷ.* The goal of this technique is to maximise predictive accuracy whilst simultaneously reducing the adversary’s ability to determine the protected attribute from the predictions. This leads to a fair classifier, as the predictions will not be carrying any group discrimination that the adversary would be able to exploit.

Figure . Adversarial Debiasing pipeline

Figure 10 shows how predictor model *Ŷ = f(X)* is trained by adjusting weights *W* to minimise a given loss function *Lp(Ŷ, y)* using stochastic gradient descent. The output layer of the predictor is then used as the input layer to an adversary network, *g*. The adversary is trained by adjusting weights *U* to minimise the loss function *LA(Ẑ, z),* in order to predict the value of *Z*.

The hyperparameters which can be tuned for the classifier in Adversarial Debiasing are batch size, epochs, learning rate, the number of hidden units in the classifier and adversary loss weight, which determines the strength of the adversarial loss function. As a default, a batch size of 32 with 100 epochs will be used. Zhang *et al*. set the number of hidden units at 200 with an adversary loss weight of 0.1, but for this analysis a range of hidden unit numbers and weights will be explored to find the best combination. The parameters which return the highest classification accuracy for the adversary will be used.

#### 6.2.6 Prejudice Remover

Prejudice Remover (Kamishima et al., 2012) is an in-processing algorithm which adds two extra regularization parameters to the output of the logistic regression classifier to remove prejudice. Prejudice within the model can be categorised as follows:

|  |  |
| --- | --- |
| **Direct Prejudice** | Where the model predictions, *Y*, are directly dependant on sensitive attribute, *Z* |
| **Indirect Prejudice** | Where the model predictions, *Y*, are statistically dependant on sensitive attribute, *Z* |
| **Latent Prejudice** | Where the model predictions, *Y*, are statistically dependant on non-sensitive attribute(s), *X* |

Table . Types of prejudice that can be present in ML algorithms

The regularisers are designed to minimise the mutual information between the protected attribute *Z*, and the prediction, *Y*. These give the ability to adjust the distribution of *Y* depending on the value of *Z*. The first regulariser, is *L2*, which is designed to avoid overfitting. The second regulariser, where *D = {(Y, Z = 0, Z = 1)},* is the prejudice remover regulariser designed to enforce fair classification. We want the value of the this regulariser to be as small as possible, to constrain the dependence between *Y* and *Z* as possible. The regularisers become a constraint of a no-indirect-prejudice condition as follows:

This condition ensures the independence between *Y* and *Z*. Where *θ* is the model parameters, and *η* is the fairness parameter. The larger the value of *η*, the more fairness that is enforced.

The prejudice remover model can be evaluated using accuracy alongside the other models, and we can also use the measure of Normalised Mutual Information (NMI) which defines how fair the predicted classes are. NMI is the mutual information between a predicted class and sensitive feature, normalised into the range [0, 1] as follows:

Where *Y* represents the predictions labels, *Z* represents the class labels, *H*(*label*) gives the entropy of the label, and *I*(*Y*;*Z*) gives the mutual information shared between *Y* and *Z*. Where NMI = 0, variables Y and Z are independent, demonstrating that the prejudice has been removed.

#### 6.2.7 Calibrated Equalised-odds Difference

Calibrated Equalised-odds Difference[[35]](#footnote-36) (Hardt et al., 2016) is a post-processing technique which constrains a classification algorithm so that false-positive and false-negative errors are received at proportionate rates with any population group. The goal of the technique is to satisfy a single equal-cost constraint whilst maintaining calibration for each group *Z*. The cost function, *gt*, is defined as a linear function for the rate of false positives, *CFP(ht)* and false negatives, *CFN(ht)* as such:

where *at* and *bt* are non-negative constants that are specific to each group. For the COMPAS algorithm, equal false-positive conditions, and equal false-negative conditions essential. Meaning, non-reoffending individuals of any race cannot be disproportionately predicted as high-risk or given a recidivism score of 1, and reoffending individuals of any race cannot be disproportionally predicted as low risk or given a recidivism score of 0. Given the cost function g*t* in the form of two binary classifiers h1 and h2 that classify samples from *G1* and *G2* respectively, we can say that calibrated equalised-odds have been achieved if the classifiers are calibrated and satisfy the cost constraint

*g1*(*h1*) = *g2*(*h2*).

As shown in the figure below, if two calibrated classifiers, and are plotted along the false positive and false negative plane, being able to ensure the false positive and false negative conditions are satisfied requires the classifiers to lie on a convex (Figure 11d).

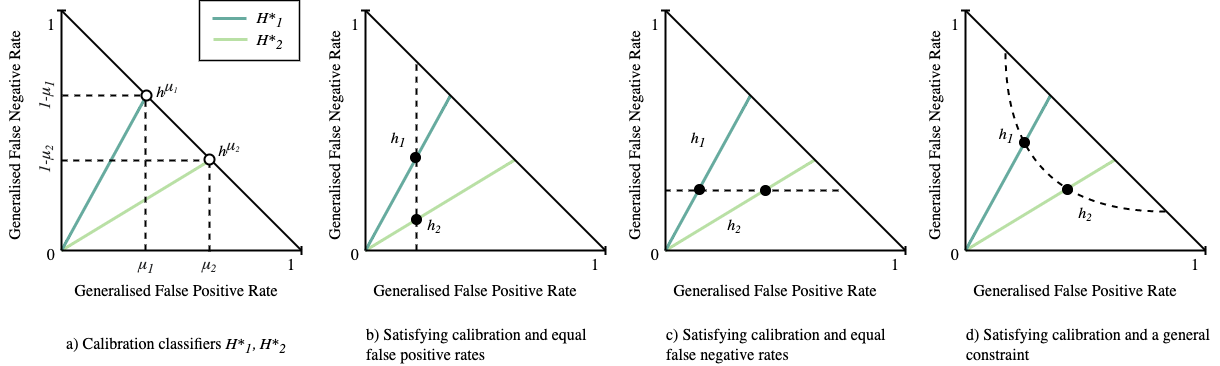


Figure 11. Calibration of classifiers H\*1 and H\*2 plotted in the false-positive, false-negative plane for the two classifiers h1 and h2, for satisfying different cost constraints (FPR/FNR)

This requires a weighted combination of error rates for each classifier. Therefore, the cost constraint hyperparameter used for the algorithm will be ‘weighted’. To achieve calibrated equalised-odds, we want to arrive at a calibrated classifier for group *G2* such that *g1*(*h1*) = *g2*(). This is done by withholding predictive information for a randomly chosen subset of group *G2*. The way this works is by holding out the validation dataset and instead of returning *h2*(*x*) for the samples, the group’s mean probability is returned. This arrives at the optimal false-negative/false-positive solution for , where is calibrated and has cost equal to *h1*.

#### 6.2.8 Reject-option Classification

Reject Option Classification (Kamiran et al., 2021) is a post-processing technique which exploits the low confidence region of a classifier for discrimination reduction. This is to remove the discriminatory decisions which are often found close to decision boundaries due to a decision-makers bias. Reject-option Classification relabels predictions within this region, giving the unprivileged group favourable labels and the privileged group unfavourable labels to reduce discrimination.

The ROC classifier gets trained on the original training data, implementing the algorithms debiasing function. This estimates the optimal classification threshold and the critical region boundary (ROC margin) which optimise a given fairness metric. We can define the ROC margin, 𝜃, as:

*Ŷ ← Ŷ +* 𝜃 if *Z = 0* and *0.5 –* 𝜃 ≤ *Ŷ ≤ 0.5 +* 𝜃

*Ŷ ← Ŷ –* 𝜃 if *Z = 1* and *0.5 –* 𝜃 ≤ *Ŷ ≤ 0.5 +* 𝜃

The predictions which lie within the ROC margin are adjusted to give class Z = 0 (unprivileged class) the prediction of *Y = 0*, and class *Z = 1* (privileged class) the predictions of *Y = 1.* This gives Non-white individuals the label of ‘No Recidivism’, and White individuals the label of ‘Recidivism’.

ROC has the following hyperparameters that can be tuned to get the ROC margin and produce fairer predictions –

* The fairness metric to be optimised – Average Odds Difference, Equal Opportunity Difference or Statistical Parity Difference
* Upper and lower bounds for the fairness metric being constrained, *-1 < x < 1*
* Upper and lower classification thresholds, *0 < x < 1*
* Number of ROC Margins to be used in the optimisation search, *x > 0*

The hyperparameters that will be chosen will be those that maximise the classification threshold whilst satisfying the fairness constraint with the lowest ROC margin.

**Chapter 7**

**Results & Discussion**

**7.1 Measuring Model Fairness**

As outline in 6.4, the fairness metrics used in the following sections are:

* Disparate Impact (DI) / Demographic Parity (1 – DI)
* Average Odds Difference (AOD)
* Statistical Parity Difference (SPD)
* Equal Opportunity Difference (EOD)

The Balanced Accuracy (ACC) for each model will be calculated by:

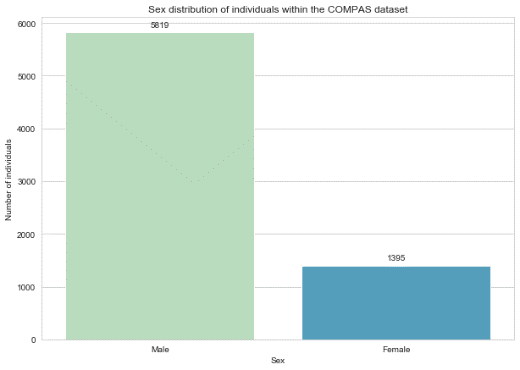
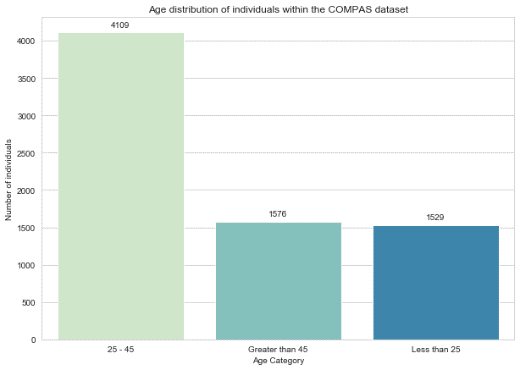
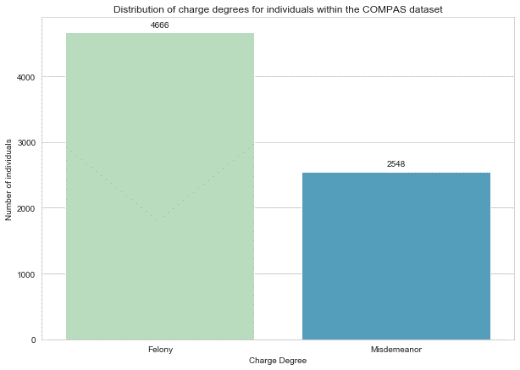
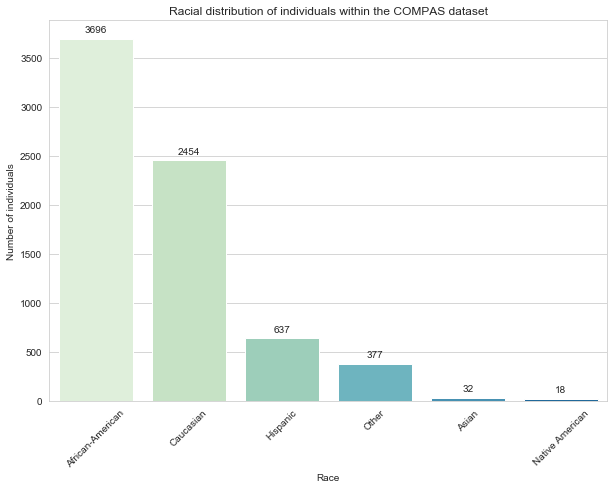
Where TPR is the percentage of positive outcomes the model correctly predicts (also known as *sensitivity*) and TNR is the percentage of negative outcomes the model correctly predicts (also known as *specificity*). However, this metric is not being used as to determine fairness. This is because the model may be predicting recidivism accurately, but it does not directly imply that the algorithm is fair as it is predicting using biased data and statistics (see Chapter 5).

**7.2 Analysis of the COMPAS Dataset**

As discussed in 6.2.2, the COMPAS dataset contains the records for 5,278 individuals. The distribution of different groups such as race, age, and prior offences committed is shown in the following figures. The issue that this analysis produces is that the dataset is neither representative of the US resident or prison population and doesn’t contain an equal distribution of groups within the population. This is not an issue specific to this analysis, rather the COMPAS algorithm (input and output) itself.

The training data for COMPAS has a much higher proportion of records for African-American individuals, followed by Caucasian, then Hispanic individuals. They used relatively little data for Asian, Native American, and individuals classed as Other. As you can see in Figure 12a, The biggest disparity in groups is between African-Americans and Native-Americans, with a ratio of just over 205:3. An unrepresentative dataset is not providing sufficient information to learn the problem, which is happening because there are far few too many examples for races other than African-American. This is going to cause an imbalanced classification problem. Figures 12b, 12c and 12d show the same issue – there is a considerably higher number examples for Charge degree = Felony, Age category = 25 to 45 and Sex = Female respectively.

Figure 12. Graphs to show the distribution of race, sex, charge degree, and age for individuals within the COMPAS dataset



(Code expansion: PCA to visualise recid/no-recid labels/clusters)

Figure 13 shows that the range of the number of prior crimes committed by individuals in the dataset is also skewed, however this is less of a problem than the distributions above. This is because it is representative of US prison populations and criminality statistics.

Chart

Description automatically generated

Figure 13. Distribution of the number of prior crimes committed by individuals in the COMPAS dataset

The Logistic Regression algorithm, as well as most other machine learning algorithms, assume that the distribution of examples for each class in the training data is equal. Where this is not true, the predictive accuracy of the model is likely to be poor. Using the True Positive Rates (TPR) and True Negative Rates (TNR) shown in the confusion matrices below, we can calculate an overall predictive accuracy of 56.5%, meaning that 43.5% of predictions are incorrect. This is a very low accuracy, and one of the reasons behind this can be attributed to the unequal distribution of attributes in the COMPAS dataset. We can also calculate that the accuracy of predictions is 57% and 56% for the privileged and unprivileged groups respectively, meaning that Non-white individuals have a marginally higher error rate than White individuals. This demonstrates unfairness, as the unprivileged class is not receiving correct predictions at the same rate as the privileged group.

|  |  |  |  |
| --- | --- | --- | --- |
| Actual Label | Positive | 32.5% | 25.5% |
| Negative | 17.5% | 24.5% |
|  |  | Positive | Negative |
|  |  | Predicted label | |

Figure 14. Confusion Matrix for the privileged group, White

|  |  |  |  |
| --- | --- | --- | --- |
| Actual Label | Positive | 27.0% | 21.0% |
| Negative | 21.0% | 29.0% |
|  |  | Positive | Negative |
|  |  | Predicted label | |

Figure 15. Confusion matrix for the unprivileged group, Non-white

Remembering that a negative prediction, *Y* = 0 corresponds to the favourable label, and a positive prediction, *Y* = 1 corresponds to the unfavourable label, a negative prediction is that of ‘No Recidivism’, and a positive prediction is that of ‘Recidivism’. This means that the False Negative (FN) Rate corresponds to the group of individuals who were predicted to reoffend, but didn’t, and the False Positive (FP) Rate corresponds to the group of individuals who reoffended, after receiving a prediction of not reoffending.

|  |  |  |
| --- | --- | --- |
|  | False Negative Rate | False Positive Rate |
| White | 17.5% | 22.5% |
| Non-white | 21.0% | 21.0% |

Table 7. TPR and TNR for the privileged (White) and unprivileged (Non-white) groups

Table 3 shows the TP and TN rates for White and Non-white individuals, which guide two conclusions:

1. Non-white individuals are more likely to be predicted to reoffend, and then not reoffend
2. White individuals are more likely to be predicted to not reoffend, but then go on to reoffend.

Recidivism is being underpredicted in White individuals, and overpredicted in Non-white individuals. This is indirect prejudice – the predictions of recidivism are statistically dependant on race.

Below, **Table 8** shows the mean difference in the two-year recidivism labels assigned to the privileged and unprivileged groups in the dataset. Each value is negative, which again shows that individuals in the unprivileged group are being assigned a prediction of recidivism at a higher rate than the privileged group.

|  |  |
| --- | --- |
| **Training Data** | -0.123760 |
| **Validation Data** | -0.145289 |
| **Testing Data** | -0.158760 |

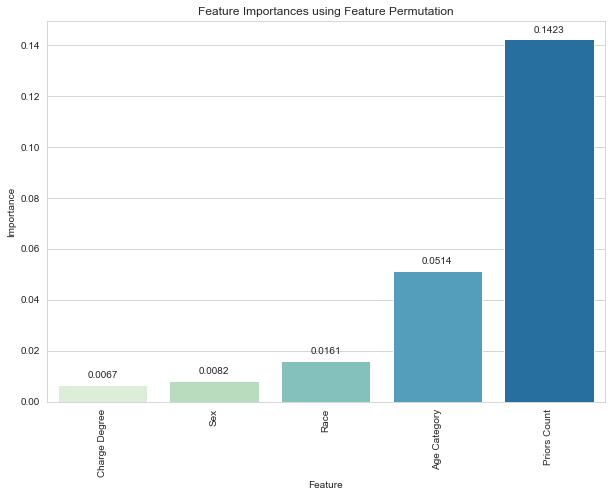
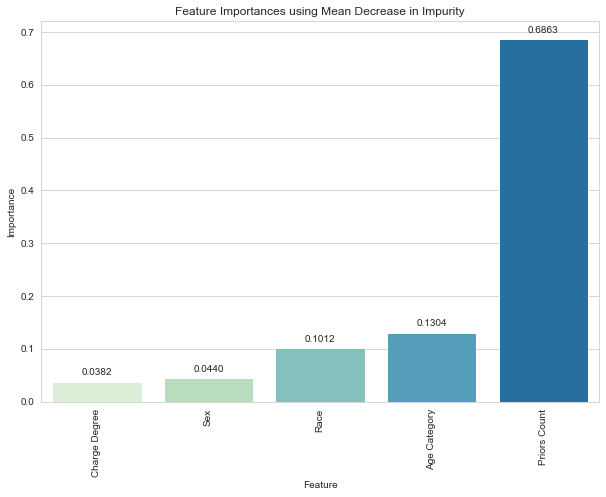
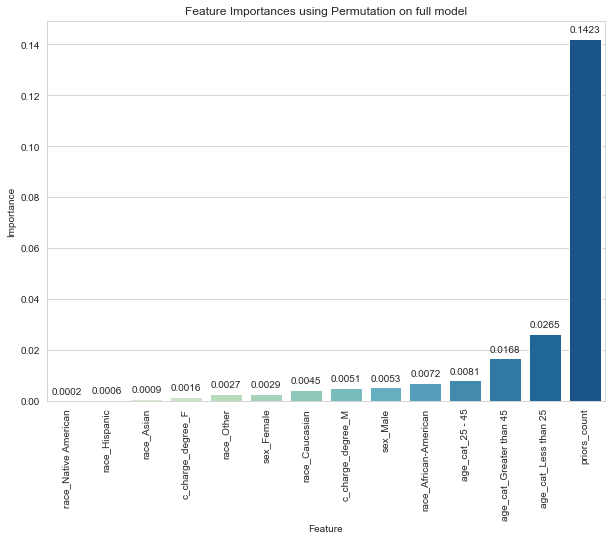
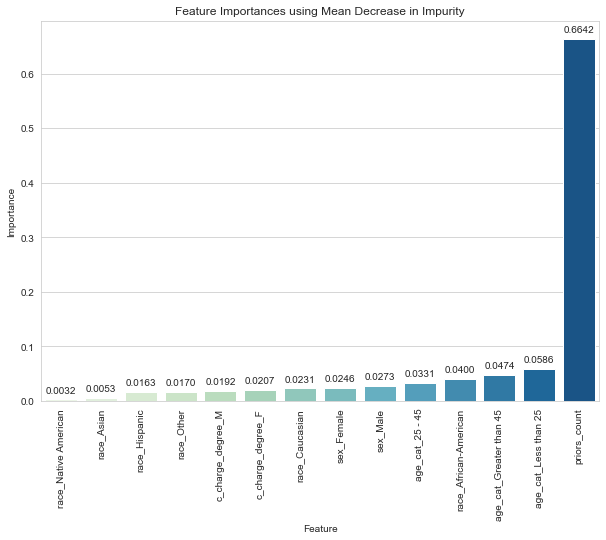
Table 8. Difference in mean outcomes for the privileged and unprivileged groups for the training, validation, and testing data

(Idea for expansion: remove race as an attribute and recalculate TP/TN/FP/FN rates to see if predictions change. If they do, there is also direct prejudice)

### 7.3 Feature Importance with Random Forest Classifier

The research done in chapter 4.4 demonstrated why statistics based on racial identity cannot provide an accurate indicator of criminal intention. Below are the results from using a forest of trees to evaluate the importance of each indicator (attribute) in the dataset for predicting recidivism. The graphs in figure 16show that the input feature with the highest level of importance using both Mean Decrease in Impunity (MDI) and Feature Permutation (FP) is priors\_count, with a value over ten times larger than the next most important feature.

Figure 16. Feature importance for individual and grouped attributes using Mean Decrease in Impunity (MDI) and Feature Permutation (FP)



These results are consistent with recidivism research, with the Bureau of Justice Statistics showing in a 9-year study that the best predictor of recidivism is an individual’s criminal history. The conclusion drawn from this part of the analysis is that race isn’t the most important feature to predict recidivism, so it does not need to be as highly weighted in the COMPAS algorithm.

### 7.5 Model Results using Bias Mitigation Techniques

#### 7.5.1 Baseline Model

Below are the results from the baseline model, which uses Logistic Regression (LR) to predict the two-year recidivism score, where 0: Did not reoffend[[36]](#footnote-37) and 1: Did reoffend**.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Results from the optimal classification threshold | | | | |
|  | **ACC** | **DI** | **AOD** | **SPD** | **EOD** |
| Goal | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| Model | 0.6347 | 0.4693 | -0.2755 | -0.3114 | -0.2637 |

Table 9. Results from the baseline Logistic Regression model

The table above shows that at the optimal classification threshold (0.5247), the model performs poorly in terms in fairness. Looking at the value for Disparate Impact, we can come to two conclusions:

1. The model does not pass the 80% rule, which demonstrates that the privileged class, White receives positive outcomes at a much higher rate than the unprivileged class, Non-White
2. The model does not produce fair results.

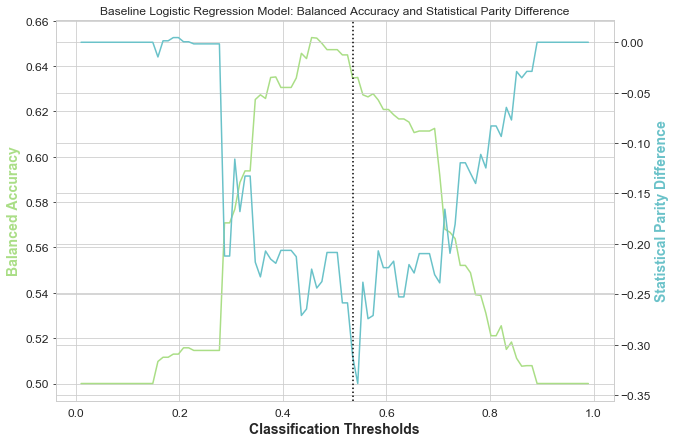
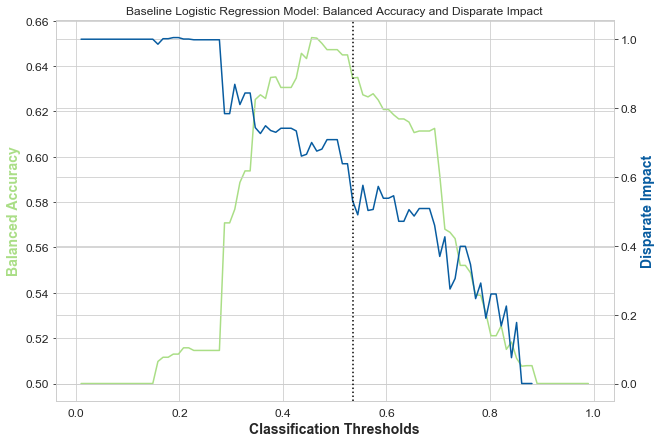
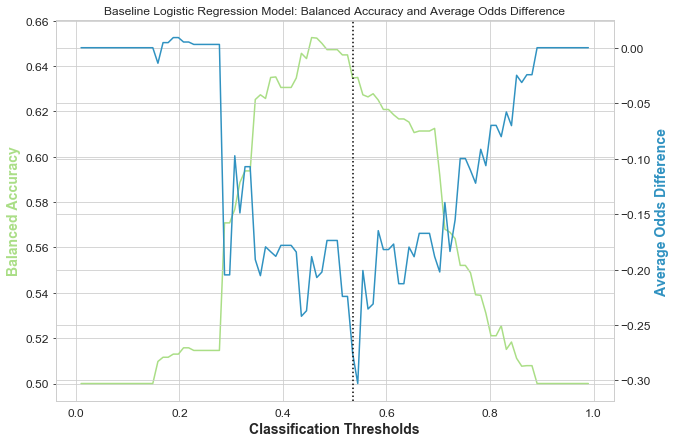


Figure . Graphs showing DI, AOD, SPD and EOD over the classification thresholds for the baseline LR model

#### 7.5.2 Model with Reweighing

Below are the results from the LR model using pre-processing technique Reweighing. What you can see is that reweighing the tuples in the dataset has increased the fairness of the baseline model with regards to each fairness metrics, while the balanced accuracy of the prediction stays constant.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Results from the optimal classification threshold | | | | |
|  | **ACC** | **DI** | **AOD** | **SPD** | **EOD** |
| Goal | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| Baseline | 0.6347 | 0.4693 | -0.2755 | -0.3114 | -0.2637 |
| Model | 0.6380 | 0.9399 | 0.0715 | 0.0287 | -0.1052 |

Table . Results from the RW Model

This model has a successful balance between accuracy and fairness. This is meaningful as a perfectly balanced dataset will have no trade-off, or rather does not have the need for a trade-off. It also demonstrates an improvement over the LR model, where the trade-off favours accuracy for fairness. Looking at the value for Disparate Impact, we can come to two conclusions:

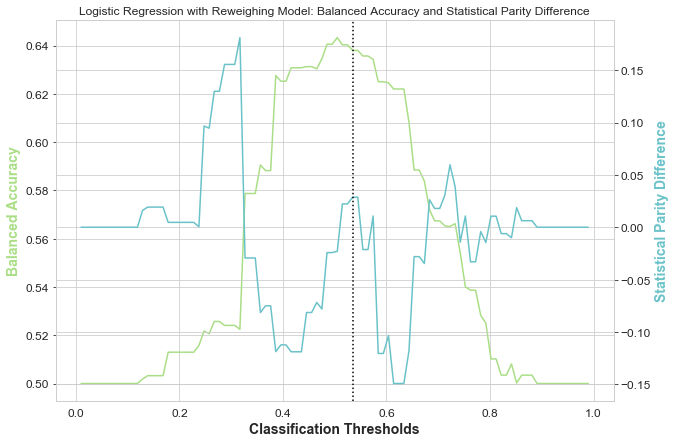
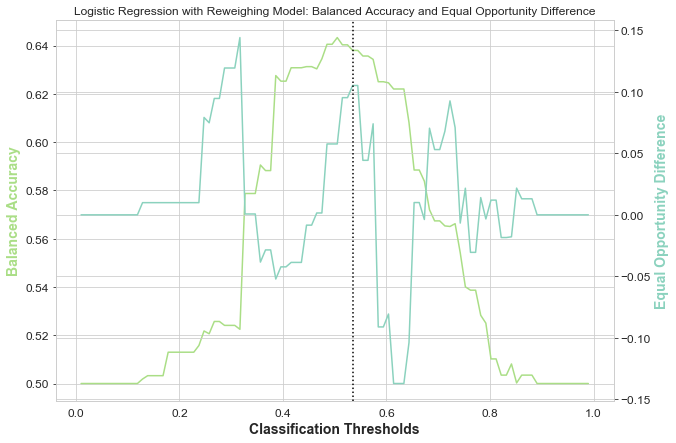
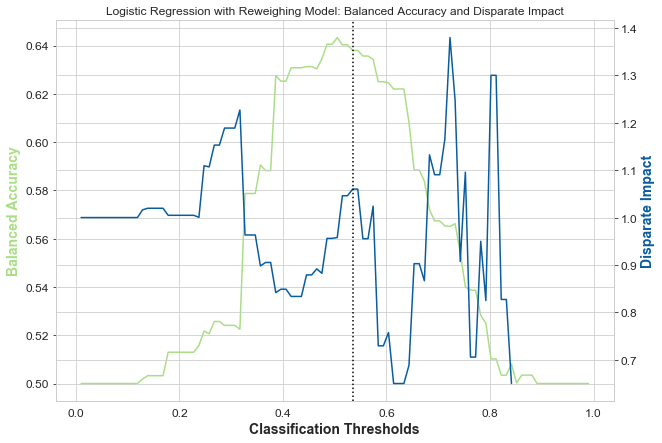
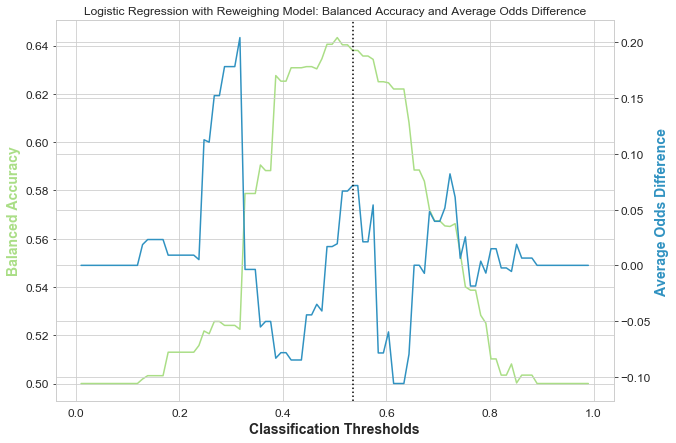
1. The model passes the 80% rule, which demonstrates that the privileged class, White receives positive outcomes at a proportionate rate to the unprivileged class, Non-White
2. The model produces fairer results than the baseline model.

The weights to be used for the classifier were as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **Weights** | |
| **Race** | **Label** | **Original** | **Reweighed** |
| White | 0.0 | 0.8758 | x |
| White | 1.0 | 1.1907 | x |
| Non-white | 0.0 | 1.1051 | x |
| Non-white | 1.0 | 0.9030 | x |

Table . Weights created by Reweighing

Figure . Graphs showing DI, AOD, SPD and EOD over the classification thresholds for the RW model



#### 7.5.3 Adversarial Debiasing Model

Below are the results from the Adversarial Debiasing (AD) model. What you can see is that moving from the LR algorithm to the AD algorithm increased fairness with regards to each fairness metrics, but at the cost of predictive accuracy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Results from the optimal classification threshold | | | | |
|  | **ACC** | **DI** | **AOD** | **SPD** | **EOD** |
| Goal | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| Baseline | 0.6347 | 0.4693 | -0.2755 | -0.3114 | -0.2637 |
| Model | 0.3464 | 0.9228 | -0.0801 | -0.0423 | -0.0806 |

Table . Results from the AD model

In order to maximise demographic parity for the AD model, the decision boundaries for predictions have been moved. The low predictive accuracy combined with high fairness metrics infer that the accuracy/fairness trade-off that occurs by moving decision boundaries is skewed in favour of fairness. So whilst the model is fair, it is not accurate. Looking at the value for Disparate Impact, we can come to two conclusions:

1. The model passes the 80% rule, which demonstrates that the privileged class, White receives positive outcomes at a proportionate rate to the unprivileged class, Non-White
2. The model produces fairer, but less accurate results than the baseline model.

These results were obtained running the AD model after finding the best values for the two hyperparameters, adversarial loss weight and the number of hidden units in the classifier. As discussed in 6.2.5, we wanted to select the hyperparameters that return the highest classification accuracy for the adversary.

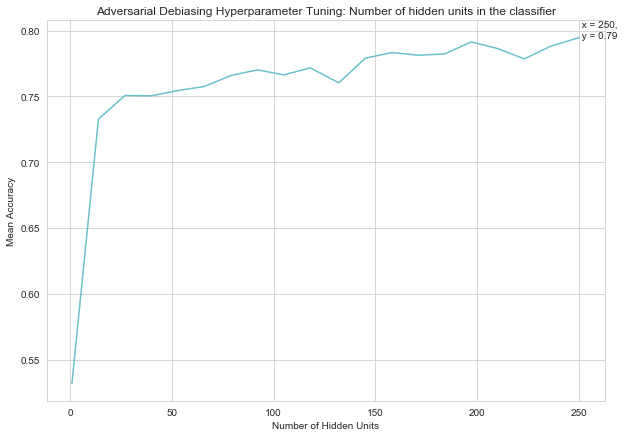
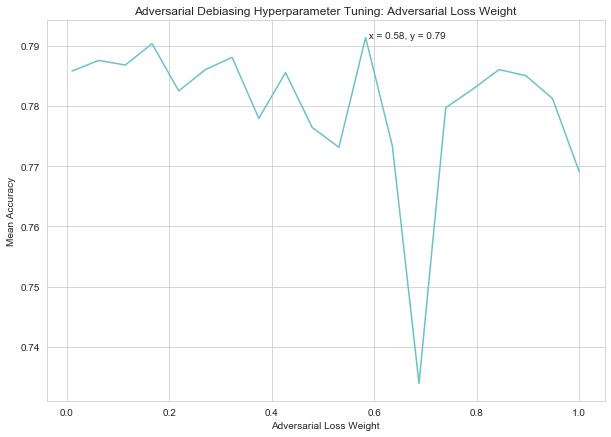
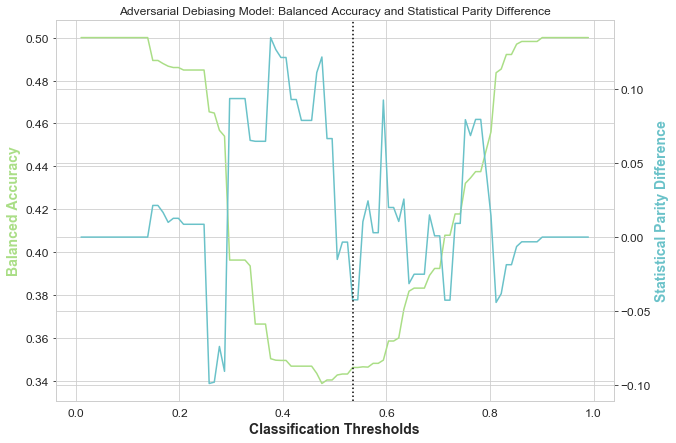
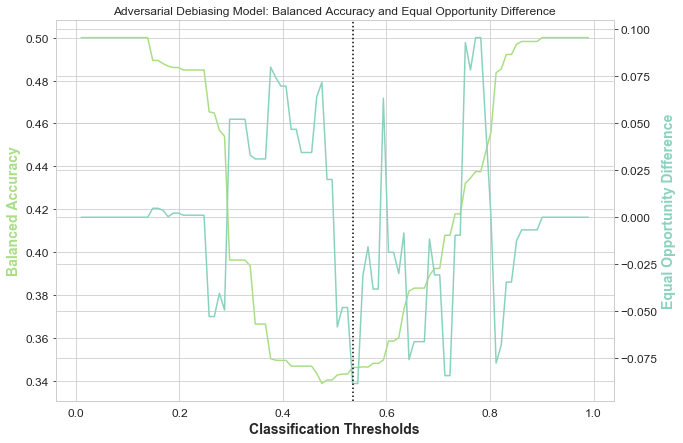
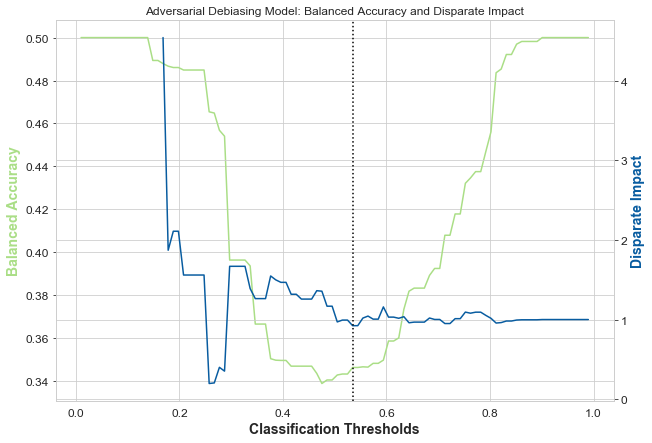
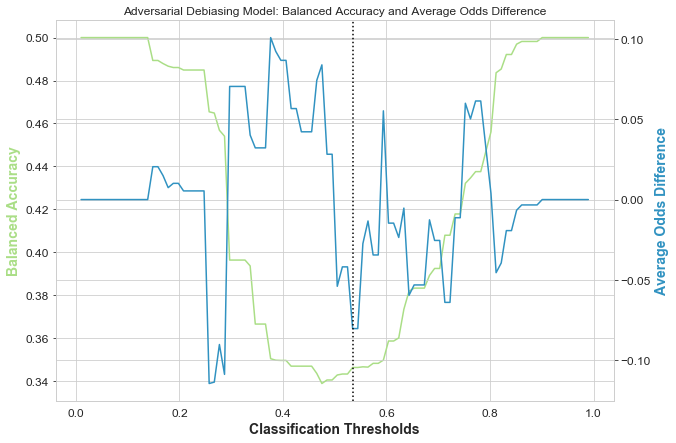


Figure . Hyperparameter tuning for the AD model

Figures 19a and 19b show that the optimum hyperparameters for the AD model are an adversary loss weight of 0.58, with 250 hidden units in the classifier.

Figure . Graphs showing DI, AOD, SPD and EOD over the classification thresholds for the AD model



#### 7.5.4 Prejudice Remover Model

Below are the results from the Prejudice Remover (PR) model. What you can see is that moving from the LR algorithm to the PR algorithm increased fairness with regards to each fairness metrics, but at the cost of predictive accuracy. This result is consistent with the AD model, where the poor predictive accuracy is a result of the accuracy/fairness trade-off.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Results from the optimal classification threshold | | | | |
|  | **ACC** | **DI** | **AOD** | **SPD** | **EOD** |
| Goal | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| Baseline | 0.6347 | 0.4693 | -0.2755 | -0.3114 | -0.2637 |
| Model | 0.4980 | 0.8985 | -0.0412 | -0.0476 | -0.1191 |

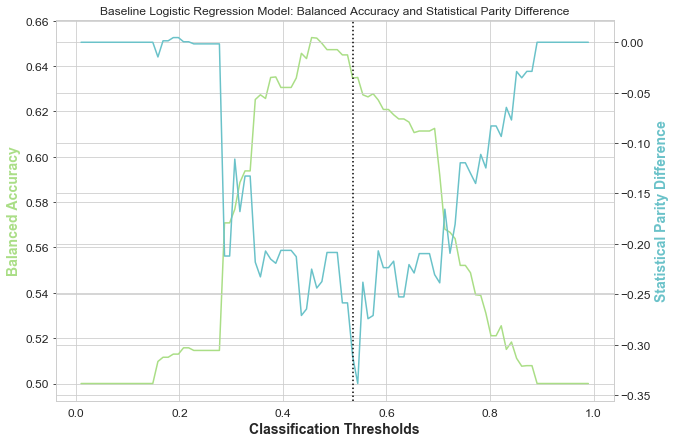
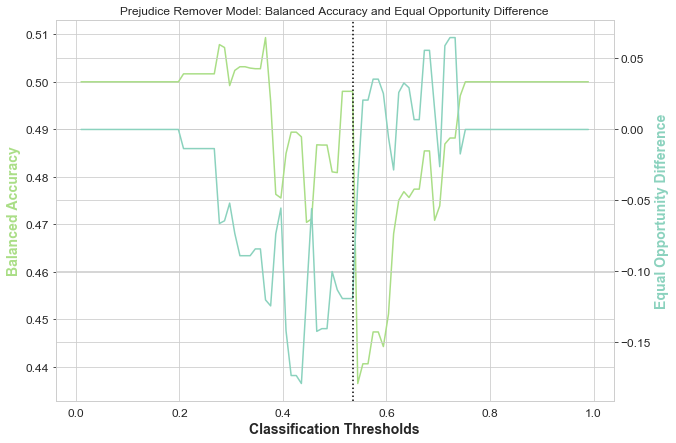
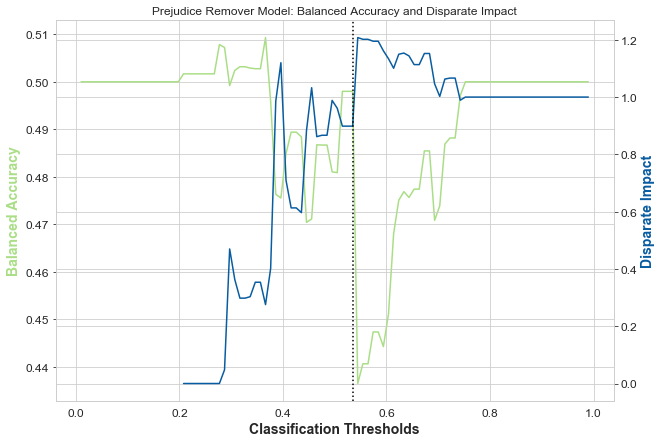
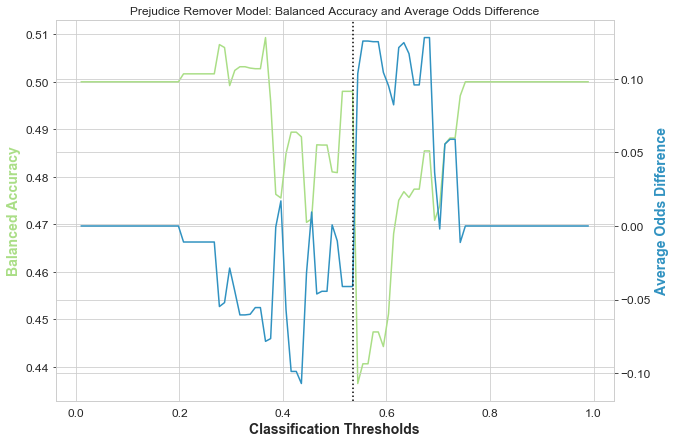
Table . Results fron the PR model

\*ADD IN NMI RESULTS

Looking at the value for Disparate Impact, we can come to two conclusions:

1. The model passes the 80% rule, which demonstrates that the privileged class, White receives positive outcomes at a proportionate rate to the unprivileged class, Non-White
2. The model produces fairer, but less accurate results than the baseline model.

Figure . Graphs showing DI, AOD, SPD and EOD over the classification thresholds for the PR model



#### 7.5.6 Model with Calibrated Equalised-odds Difference

Below are the results from the LR model using the post-processing technique Calibrated Equalised-odds Difference (CEOD). What you can see is that applying CEOD to the baseline LR model increased fairness only in terms of Disparate Impact, with the other fairness metrics having worsened and the balanced accuracy of the prediction stays constant.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Results from the optimal classification threshold | | | | |
|  | **ACC** | **DI** | **AOD** | **SPD** | **EOD** |
| Goal | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| Baseline | 0.6347 | 0.4693 | -0.2755 | -0.3114 | -0.2637 |
| Model | x | x | x | x | x |

Table . Results from the CEOD model

Looking at the value for Disparate Impact, we can come to two conclusions:

1. The model passes the 80% rule, which demonstrates that the privileged class, White receives positive outcomes at a proportionate rate to the unprivileged class, Non-White
2. The model produces fairer results than the baseline model, in terms of DI.

The table below shows the difference in generalised FP and FN rates between unprivileged and privileged groups, with the original data and the transformed data, where the cost constraint has been tuned.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Validation Data** | | **Testing Data** | |
|  |  | **FPR** | **FNR** | **FPR** | **FNR** |
| **Original** | | -0.1370 | 0.0789 | -0.0997 | 0.0948 |
| **Transformed** | **Cost Constraint = FPR** | -0.0532 | 0.1533 | -0.0413 | 0.1430 |
| **Cost Constraint = FNR** | x | x | x | x |
| **Cost Constraint = Weighted** | x | x | x | x |

Table . Hyperparameter tuning for the CEOD model

These results showed that the cost constraint to be chosen CEOD should be ‘weighted’, which optimises weighted combination of false-positive and false-negative rates.

#### 7.5.7 Model with Reject-option Classification

Below are the results from the LR model using the post-processing technique Reject-option Classification (ROC). What you can see is that applying ROC to the baseline LR model increased fairness with regards to each fairness metrics, while the balanced accuracy of the prediction stays constant.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Results from the optimal classification threshold | | | | |
|  | **ACC** | **DI** | **AOD** | **SPD** | **EOD** |
| Goal | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| Baseline | 0.6347 | 0.4693 | -0.2755 | -0.3114 | -0.2637 |
| Model | x | x | x | x | x |

Table . Results from the ROC model

Looking at the value for Disparate Impact, we can come to two conclusions:

1. The model passes the 80% rule, which demonstrates that the privileged class, White receives positive outcomes at a proportionate rate to the unprivileged class, Non-White
2. The model produces fairer results than the baseline model.

These results were obtained after finding the best values for the hyperparameters, which are the fairness metric to be constrained and corresponding the upper and lower bounds. As discussed in 6.2.8, we wanted to select the hyperparameters that maximise the classification threshold whilst satisfying the fairness constraint with the lowest ROC margin.

|  |  |  |
| --- | --- | --- |
|  | **Optimal Classification Threshold** | **ROC Margin** |
| **AOD** | 0.5247 | 0.0485 |
| **EOD** | 0.5346 | 0.0190 |
| **SPD** | 0.5148 | 0.0594 |

Table 17. Results from hyperparameter tuning for Reject-option Classification: Fairness metric constraints

Table 17 shows the optimal classification threshold and ROC margins model, where a different fairness metric is being constrained each time. The metric chosen to be constrained was Equal Opportunity Difference, as it has the highest classification threshold and the lowest ROC margin.

|  |  |  |  |
| --- | --- | --- | --- |
| **Bounds** | | **Results** | |
| **Lower** | **Upper** | **Optimal Classification Threshold** | **ROC Margin** |
| -1.0 | 1.0 | 0.4555 | 0.0093 |
| -0.05 | 1.0 | 0.4753 | 0.0291 |
| -0.01 | 1.0 | 0.4456 | 0.0273 |

Table 18. Results from hyperparameter tuning for Reject-option Classification: Upper and lower bounds for the constrained fairness metric

Table 18[[37]](#footnote-38) shows the optimal classification threshold and ROC margins for the model, where different upper and lower bounds for the constrained metric are being used each time. The upper and lower bounds chosen were 1.0 and -1.0 respectively as they have the best trade-off in terms of maximising the optimal classification threshold while minimising the ROC margin.

### 7.6 Comparison

ADD GRAPHS

### 7.7 Optimised Fairness Model

The results from section 7.6 guided the design of the Optimised Fairness (OF) model, which is an LR Classifier with the RW pre-processing algorithm and ROC post-processing algorithm applied. These were the techniques that produced the highest fairness and accuracy. The results below show that the FM model has increased the fairness of the baseline model with regards to each fairness metrics, while the balanced accuracy of the prediction stays constant.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Results from the optimal classification threshold** | | | | | |
|  | **ACC** | **DI** | **1 – DI** | **SPD** | **AOD** | **EOD** |
| **Goal** | 1.0[[38]](#footnote-39) | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **Baseline** | 0.6723 | 0.5613 | 0.4387 | -0.3463 | -0.3065 | -0.2410 |
| **Model** | 0.6481 | 1.006 | -0.0064 | 0.0028 | -0.0009 | -0.0069 |

Table . Results for the OF model

The hyperparameters for the LR classifier were found using Grid Search, iterating through each solver, regularisation penalty and C parameter below:

* Solvers: newton-cg, lbfgs, liblinear
* Regularisation penalties: none, L1, L2, elasticnet
* C parameters: C in [100, 10, 1.0, 0.1, 0.01]

The hyperparameters that returned the highest accuracy were a liblinear solver, L2 regularisation penalty and C = 0.01, which were used to train the LR classifier.

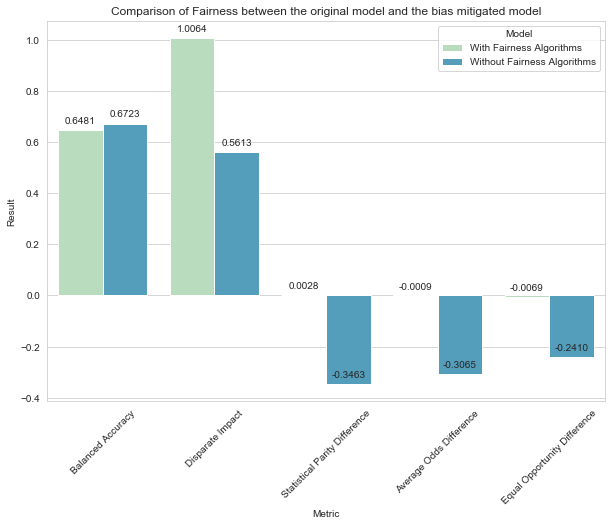


Figure . Graphs showing DI, AOD, SPD and EOD over the classification thresholds for the OF model

# Chapter 8

## Proposing Solutions

### 8.1 Ethical Datasets for Training

#### **8.1.1 Using Equally Weighted Sample Sizes for Training**

#### **8.1.2 Representative Datasets**

### 8.2 Using More Accurate Attributes to Predict Recidivism

#### 8.2.1 Removing Race as an Attribute

#### 8.2.2 Weighting Attributes

#### 8.2.3 Including Dynamic Attributes

### 8.3 Testing for Bias

### 8.4 AI Fairness 360 for Bias Testing and Mitigation

### 8.5 Bias Intervention Checks

### 8.6 Transparency in the Decision-Making Process

### 8.7 Changing Culture

#### 8.7.1 Implicit and Explicit Bias Training

#### 8.7.2 Increasing Diversity

### 8.8 Legality

# Chapter 9

## Conclusion

## References

1. ACLU (2013) *The war on marijuana in black and white*, New York: ACLU Foundation. Available at: https://www.aclu.org/report/report-war-marijuana-black-and-white [Accessed 20 December 2021]
2. Anwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016) *Machine Bias*. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Accessed 18 September 2021]
3. Alper, A., Durose, M. R. & Markman, J. (2018) *2018 update on prisoner recidivism: a 9-year follow-up period* (2005-2014), Bureau of Justice Statistics Special Report. Available at: <https://bjs.ojp.gov/content/pub/pdf/18upr9yfup0514.pdf> [Accessed 07 April 2022]
4. Barton, G., Lee, N. & Resnick, P. (2019) *Algorithmic bias detection and mitigation: best practices and policies to reduce consumer harms*. Available at: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ [Accessed 25 September 2021]
5. BCS, The Chartered Institute for IT (2021) *Code of conduct for BCS members*. Available at: https://www.bcs.org/media/2211/bcs-code-of-conduct.pdf [Accessed 01 November 2021]
6. Bellamy, R. et al. (2019) ‘AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias’, *IBM Journal of Research and Development*, 63(4), pp. 4:1-4:15, DOI:10.1147/JRD.2019.2942287
7. Bureau of Justice Assistance (no date) *History of Risk Assessment*. Available at: https://bja.ojp.gov/program/psrac/basics/history-risk-assessment [Accessed 07 November 2021]
8. Chang, K. et al. (2017) ‘Men also like shopping: reducing gender bias amplification using corpus-level constraints’, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 7-11 September, DOI:10.18653/v1/D17-1323
9. Dastin, J. (2018) ‘Amazon scraps secret AI recruiting tool that showed bias against women’, *Reuters*. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [Accessed 19 December 2021]
10. Data Collaborative for Justice (2020) *Tracking enforcement trends in New York City: 2003-2018.* Available at: https://datacollaborativeforjustice.org/wp-content/uploads/2020/09/2020\_08\_31\_Enforcement.pdf [Accessed 17 December 2021]
11. Desilver, D., Fahmy, D. & Lipka, M. (2020) *10 things we know about race and policing in the U.S.* Available at: https://www.pewresearch.org/fact-tank/2020/06/03/10-things-we-know-about-race-and-policing-in-the-u-s/ [Accessed 17 December 2021]
12. Equivant (2016) *Response to ProPublica: demonstrating accuracy equity and predictive parity.* Available at: https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/ [Accessed 18 December 2021]
13. Géron, A. (2019) *Hands-on machine learning with Scikit-learn*, Keras & TensorFlow. California: O’Reilly Media, Inc.
14. Hannah-Moffat, K. (2012) ‘Actuarial sentencing: an “unsettled” proposition’, *Justice Quarterly*, 30(2), pp. 270-296, DOI:10.1080/07418825.2012.682603
15. Harcourt, B. (2010) *Risk as a proxy for race*. John M. Olin Program in Law and Economics Working Paper No. 535. Available at: https://chicagounbound.uchicago.edu/law\_and\_economics/433/ [Accessed 06 November 2021]
16. Heilweil, R. (2020) *Why algorithms can be racist and sexist*. Available at: https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency [Accessed 19 December 2021]
17. Jones, E. & Harris, V. (1967) ‘The attribution of attitudes’, *Journal of Experimental Social Psychology*, 3(1), pp. 1-24, DOI: 10.1016/0022-1031(67)90034-0
18. Kirkpatrick, K. (2017) ‘It’s not the algorithm, it’s the data’, *Communications of the ACM*, 60(2), pp. 21-23. DOI:10.1145/3022181 [Accessed 12 December 2021]
19. NAACP (2021) *Criminal justice fact sheet*. Available at: https://naacp.org/resources/criminal-justice-fact-sheet [Accessed 04 November 2021]
20. National Institute of Justice (2008) *Recidivism is a core criminal justice concern*. Available at: https://nij.ojp.gov/topics/articles/recidivism-core-criminal-justice-concern#citation--0 [Accessed 06 November 2021]
21. Noble, S. U. (2018) *Algorithms of oppression*. New York: NYU Press.
22. NYCLU (2012) *Stop-and-frisk 2011*. Available at: https://www.nyclu.org/sites/default/files/publications/NYCLU\_2011\_Stop-and-Frisk\_Report.pdf [Accessed 17 December 2021]
23. NYCLU (2019) *Stop-and-frisk in the de Blasio era*, New York: New York Civil Liberties Union. Available at: https://www.nyclu.org/en/publications/stop-and-frisk-de-blasio-era-2019 [Accessed 17 December 2021]
24. Pew Research Centre (2021) *STEM jobs see uneven progress in increasing gender, racial and ethnic diversity.* Available at: https://www.pewresearch.org/science/2021/04/01/stem-jobs-see-uneven-progress-in-increasing-gender-racial-and-ethnic-diversity/ [Accessed 19 December 2021]
25. Ross, L (1977) ‘The intuitive psychologist and his shortcomings: distortions in the attribution process’, *Advances in Experimental Social Psychology*, 10, pp. 173-220, DOI: 10.1016/S0065-2601(08)60357-3
26. United States Census Bureau (2010-2020) *Census return for racial and ethnic diversity in the United States: 2010 census and 2020 census, United States*. Public Record Office: (2021) Available at: https://www.census.gov/library/visualizations/interactive/racial-and-ethnic-diversity-in-the-united-states-2010-and-2020-census.html [Accessed 20 December 2021]
27. University of Sussex (2021) *Division of general counsel, governance and compliance: processing personal data*. Available at: https://www.sussex.ac.uk/ogs/policies/information/dpa/processingdata [Accessed 21 October 2021]
28. U.S. Department of Health and Human Services (2011) *Results from the 2010 national survey on drug use and health: summary of national findings*, Rockville: Substance Abuse and Mental Health Services Administration. Available at: https://www.samhsa.gov/data/sites/default/files/NSDUHNationalFindingsResults2010-web/2k10ResultsRev/NSDUHresultsRev2010.pdf [Accessed 20 December 2021]

## Appendices

#### Appendix A – Imprisonment rates of U.S. residents based on sentenced prisoners under the jurisdiction of state or federal correctional authorities, by jurisdiction, sex, and race or ethnicity, 2010-2020

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Per 100,000 U.S. residents within each demographic group** | **Asian** | 108 | 107 | 103 | 99 | 98 | 98 | 96 | 93 | 92 | 88 | 74 |  |  | -31.8 | -16.1 |
| **American Indian/Alaska Native** | 1,044 | 983 | 927 | 846 | 903 | 863 | 853 | 881 | 873 | 885 | 779 |  |  | -25.4 | -12.1 |
| **Hispanic** | 658 | 650 | 629 | 630 | 611 | 592 | 591 | 575 | 555 | 530 | 446 |  |  | -32.2 | -15.8 |
| **Black** | 1,489 | 1,438 | 1,377 | 1,348 | 1,302 | 1,239 | 1,199 | 1,161 | 1,124 | 1,088 | 938 |  |  | -37.0 | -13.8 |
| **White** | 248 | 243 | 238 | 237 | 234 | 228 | 223 | 221 | 218 | 214 | 183 |  |  | -26.5 | -14.9 |
| **Female** | 66 | 65 | 63 | 65 | 65 | 64 | 64 | 63 | 63 | 61 | 47 |  |  | -28.6 | -21.7 |
| **Male** | 948 | 932 | 910 | 907 | 891 | 865 | 848 | 831 | 811 | 789 | 678 |  |  | -28.4 | -14.0 |
| **Per 100,000 U.S. residents** | **State** | 439 | 429 | 418 | 418 | 412 | 403 | 397 | 391 | 381 | 371 | 315 |  |  | -28.1 | -15.1 |
| **Federal** | 61 | 63 | 62 | 61 | 60 | 55 | 53 | 51 | 50 | 48 | 43 |  | **Percentage change** | -29.9 | -10.6 |
| **Total** | 500 | 492 | 480 | 479 | 472 | 459 | 450 | 442 | 431 | 419 | 358 |  | -28.4 | -14.5 |
|  | **Year** | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |  | 2010-2020 | 2019-2020 |

Source: Bureau of Justice Statistics, Federal Justice Statistics Program, 2020 (preliminary), National Corrections Reporting Program, 2019, National Prisoner Statistics, 2010–2020, Survey of Inmates in State and Federal Correctional Facilities, 2004, and Survey of Prison Inmates, 2016; and U.S. Census Bureau, postcensal resident population estimates for January 1 of the following calendar year.

#### Appendix B – Sentenced prisoners under the jurisdiction of state or federal correctional authorities, by jurisdiction, sex, and race or ethnicity, 2010–2020

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Per 100,000 U.S. residents within each demographic group** | **Asian** | 108 | 107 | 103 | 99 | 98 | 98 | 96 | 93 | 92 | 88 | 74 |  |  | -31.8 | -16.1 |
| **American Indian/Alaska Native** | 1,044 | 983 | 927 | 846 | 903 | 863 | 853 | 881 | 873 | 885 | 779 |  |  | -25.4 | -12.1 |
| **Hispanic** | 658 | 650 | 629 | 630 | 611 | 592 | 591 | 575 | 555 | 530 | 446 |  |  | -32.2 | -15.8 |
| **Black** | 1,489 | 1,438 | 1,377 | 1,348 | 1,302 | 1,239 | 1,199 | 1,161 | 1,124 | 1,088 | 938 |  |  | -37.0 | -13.8 |
| **White** | 248 | 243 | 238 | 237 | 234 | 228 | 223 | 221 | 218 | 214 | 183 |  |  | -26.5 | -14.9 |
| **Female** | 66 | 65 | 63 | 65 | 65 | 64 | 64 | 63 | 63 | 61 | 47 |  |  | -28.6 | -21.7 |
| **Male** | 948 | 932 | 910 | 907 | 891 | 865 | 848 | 831 | 811 | 789 | 678 |  |  | -28.4 | -14.0 |
| **Per 100,000 U.S. residents** | **State** | 439 | 429 | 418 | 418 | 412 | 403 | 397 | 391 | 381 | 371 | 315 |  |  | -28.1 | -15.1 |
| **Federal** | 61 | 63 | 62 | 61 | 60 | 55 | 53 | 51 | 50 | 48 | 43 |  |  | -29.9 | -10.6 |
| **Total** | 500 | 492 | 480 | 479 | 472 | 459 | 450 | 442 | 431 | 419 | 358 |  | **Percentage change** | -28.4 | -14.5 |
|  | **Year** | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |  | 2010-2020 | 2019-2020 |

Source: Bureau of Justice Statistics, Federal Justice Statistics Program, 2020 (preliminary); National Corrections Reporting Program, 2019; National Prisoner Statistics, 2010–2020; Survey of Inmates in State and Federal Correctional Facilities, 2004; and Survey of Prison Inmates, 2016.

#### Appendix C – Cumulative percent of state prisoners released in 24 states in 2008 who were arrested following release, by sex, race or ethnicity, age at release, and year following release

Source: Bureau of Justice Statistics, Recidivism of State Prisoners Released in 2008 data collection, 2008–2018.

<https://bjs.ojp.gov/BJS_PUB/rpr24s0810yfup0818/Web%20content/508%20compliant%20PDFs>, page 4

#### Appendix D – Results from all bias mitigation models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Theil Index** | 0.2684 | 0.0280 | 0.4531 | 0.4060 |  |  | 0.2100 |
| **Statistical Parity Difference** | -0.3818 | 0.0287 | -0.0423 | -0.0476 |  |  | -0.0440 |
| **Equal Opportunity Difference** | -0.2637 | 0.1052 |  | -0.1191 |  |  | -0.0051 |
| **Average Odds Difference** | -0.3725 | 0.0715 | -0.0806 | -0.0412 |  |  | -0.0481 |
| **Distance from Demographic Parity** | 0.5395 | 0.0600 | 0.0772 | 0.1015 |  | 0.0000 | 0.0683 |
| **Disparate Impact** | 0.4605 | 0.9400 | 0.9228 | 0.8985 |  | 1.000 | 0.9317 |
| **Balanced Accuracy** | 0.6672 | 0.6380 | 0.3464 | 0.4980 |  | 0.5919 | 0.6298 |
|  | **Logistic Regression** | **Logistic Regression with Reweighing** | **Adversarial Debiasing** | **Prejudice Remover** | **Logistic Regression with Reject-Option Classification** | **Logistic Regression with Calibrated Equalised-Odds Difference** | **Logistic Regression with Reweighing and Reject-Option Classification** |
|  |  | **Pre-processing** | **In-processing** | | **Post-processing** | | **Pre- and Post-processing** |

1. *Recidivism* refers to a person’s relapse into criminal behaviour, often after the person receives sanctions or undergoes intervention for a previous crime. It is measured by criminal acts that result in rearrest, reconviction, or return to prison with or without a new sentence during the three-year period after the person’s release (National Institute of Justice, 2008) [↑](#footnote-ref-2)
2. BCS Code of Conduct 3.d. “You shall NOT disclose or authorise to be disclosed or use for personal gain or to benefit a third party, confidential information except with the permission of your Relevant Authority, or as required by Legislation.” [↑](#footnote-ref-3)
3. BCS Code of Conduct 1.a. “You shall have due regard for public health, privacy, security and wellbeing of others and the environment.” [↑](#footnote-ref-4)
4. BCS Code of Conduct 2.d. “You shall ensure that you have the knowledge and understanding of Legislation and that you comply with such Legislation, in carrying out your professional responsibilities.” [↑](#footnote-ref-5)
5. BCS Code of Conduct 3.a. “You shall carry out your professional responsibilities with due care and diligence in accordance with the Relevant Authority’s requirements whilst exercising your professional judgement at all times.” [↑](#footnote-ref-6)
6. Meaning that the processing is necessary for the University to perform a task in the public interest or as part of its official functions (University of Sussex, 2021) [↑](#footnote-ref-7)
7. Meaning that the processing is necessary for the legitimate interests of the University or a third party, unless there is a good reason to protect the individual’s personal data which overrides those legitimate interests (University of Sussex, 2021) [↑](#footnote-ref-8)
8. BCS Code of Conduct 1.b. “You shall have due respect for the legitimate rights of Third Parties.” [↑](#footnote-ref-9)
9. BCS Code of Conduct 2.e. “You shall respect and value alternative viewpoints and, seek, accept and offer honest criticisms of work.” [↑](#footnote-ref-10)
10. BCS Code of Conduct 2.a. “You shall only undertake to do work or provide a service that is within your professional competence.” [↑](#footnote-ref-11)
11. BCS Code of Conduct 2.b. “You shall NOT claim any level of competence that you do not possess” [↑](#footnote-ref-12)
12. BCS Code of Conduct 2.c. “You shall develop your professional knowledge, skills, and competence on a continuing basis, maintaining awareness of technological developments, procedures, and standards that are relevant to your field” [↑](#footnote-ref-13)
13. AIF360 (AI Fairness 360) is an open-source library of fairness metrics for datasets and machine learning models, and algorithms to mitigate bias in datasets and models (Bellamy*, et al*. 2019) [↑](#footnote-ref-14)
14. COMPAS takes 402 attributes about an individual from their criminal record, a subject interview, and correctional record to produce a recidivism risk score [↑](#footnote-ref-15)
15. Both the US Census and COMPAS use ‘African-American’ and ‘Caucasian’ to classify race, but I have used ‘Black’ and ‘White’ respectively throughout this report, which are more racially-sensitive terms [↑](#footnote-ref-16)
16. Hispanic (or Latino) is officially classified as an ethnicity rather than a racial group, however in America, Hispanic individuals do not receive the same treatment as White individuals and are often targets of discrimination (United States Census Bureau, 2020) [↑](#footnote-ref-17)
17. ‘Other’ in this context represents those who identify as a racial group not specified by the Broward County Sheriff’s Office, such as Native Hawaiian/Pacific Islander or Multiracial Americans [↑](#footnote-ref-18)
18. *Desistance* refers to the process by which a person arrives at a permanent state of non-offending (National Institute of Justice, 2008) [↑](#footnote-ref-19)
19. First generation (1900-1970s) [↑](#footnote-ref-20)
20. Second generation (1970s-1990s) [↑](#footnote-ref-21)
21. Third generation (1990s-2000s) [↑](#footnote-ref-22)
22. Fourth generation (current) [↑](#footnote-ref-23)
23. A term coined by Scott T. Allison and David M. Messick after research carried out between 1970-1985 [↑](#footnote-ref-24)
24. A term coined by Lee Ross in 1977 after a series of experiments carried out by Edward E. Jones and Victor A. Harris (1967) to determine the attribution of attitudes [↑](#footnote-ref-25)
25. Filed on January 31st, 2008, resulting in a historic ruling on August 12th, 2013, which found the NYPD liable for a pattern and practise of racial profiling and unconstitutional stops [↑](#footnote-ref-26)
26. Data from Stop-and-Frisk 2011 (NYCLU, 2012) [↑](#footnote-ref-27)
27. American Civil Liberties Union [↑](#footnote-ref-28)
28. Using data from the Federal Bureau of Investigation’s Uniform Crime Reporting Program, U.S. annual census data from 2001-2010, and the 2010 National Survey on Drug Use and Health [↑](#footnote-ref-29)
29. Also known as unconscious bias [↑](#footnote-ref-30)
30. Also known as conscious bias [↑](#footnote-ref-31)
31. PRC pulled data from the U.S. Census Bureau’s 1990-2000 U.S. decennial censuses, 1990-2016 U.S. Census Bureau, and aggregated 2014-2016 and 2017-2019 American Community survey data [↑](#footnote-ref-32)
32. See chapter 5.1. [↑](#footnote-ref-33)
33. Originally framed by the Technical Advisory Committee on Testing (TACT) assembled by the State of California Fair Employment Practice Commission (FEPC) in 1971. [↑](#footnote-ref-34)
34. Also referred to as *Gini* [↑](#footnote-ref-35)
35. Also referred to as Disparate Mistreatment [↑](#footnote-ref-36)
36. A score of 0 technically means the individual did not get charged or imprisoned for a crime but does not necessarily mean a crime was not committed. For the sake of this analysis where it is not possible to see if the individual committed a crime they were not charged for, we will assume 0 means that the individual has desisted from crime. [↑](#footnote-ref-37)
37. Lower bounds of -1, -0.5, -0.1, -0.05 and -0.01 were used with upper bounds of 1, 0.5, 0.1, 0.05 and 0.01 which produced 25 combinations. Only three combinations have been displayed in the table as the results were identical for the other pairs. [↑](#footnote-ref-38)
38. A high accuracy in this case is not the main goal – the predictions may be accurate but this does not mitigate existing bias or imply fairness [↑](#footnote-ref-39)