Literature and Proposed Plan of Research for the Use of Pre-processing Techniques to Mitigate Bias in Machine Learning Models

An interim report
submitted in partial fulfilment
of the requirements for the degree of
Master of Science in Data Analytics

By
Elton Grivith D'Souza
D00264329

Under the supervision of
Dr. Natalia Budarina

Department of Computing Science & Mathematics
Dundalk Institute of Technology
Louth, Ireland

June 2024

Abstract

Machine learning models and its applications have been used across various aspects of society, often in situations where automation can help an organisation be more efficient and manage their resources effectively. While these solutions provide a streamlined approach fixing large-scaled, redundant, and time consuming processes, they may contain problems that cause bias in regard to certain features and groups. An example of this is the COMPAS model that the United States government (USA) used to assess the probability of recidivism of inmates. The model, proven to be biased, showed behaviour that was favoured towards White inmates as compared to Black and Hispanic racial groups.
In this study, we aim to use and discover methods that can help to efficiently track and reduce the impact of bias using pre-processing techniques. The proposed work will be conducted on a subset of the COMPAS recidivism dataset that consists of all cases recorded by the state of Florida, USA. This academic report, explores the literature surrounding the domain, ethical concerns inherent in this research and defines the research methodology while presenting our preliminary findings.

Contents

Chapter 1

Introduction

Research on artificial intelligence (AI) and machine learning (ML) has bought about significant advancements over the past few years. ML models have been increasingly used as decision-making entities that have a significant impact on the lives of many people. These models are used independently or, in some cases, in lieu with domain experts, used in sensitive domains such as healthcare, medicine, politics, and even judicial systems.

While these applications of ML models are impressive and noteworthy, if used in such sensitive domains without through evaluation, problems can arise Dressel & Farid (2021). This is more apparent in predictive models where the models can reinforce the bias present in the data. And when these models are deployed without a comprehensive analysis on its behaviour, it can have major impact on peoples lives Janssen & Kuk (2016). Here, the analysis of the models performance plays a vital role, and detection of these errors can help us build better performing and ethically aligned ML models.

Pre-processing is a series of procedures executed to better align the data to suit the needs of the model which can help attain better results Bellamy et al. (2019). Through comprehensive review of the literature, we present a review of this domain that consists of the theoretical framework aligned with this study, definitions and metrics that help detect and mitigate bias, ethical considerations of the study and some of the analysis techniques that we found were commonly used across the domain.

In addition to this, we have also provided our planned approach for the completion of research and preliminary findings that validate the need for *debiasing* methods to be implemented.

Chapter 2

Literature Review

The increasing reliance on machine learning algorithms in judicial and law enforcement systems has raised significant ethical concerns, particularly regarding bias and fairness Barenstein (2019). Recent studies have shown that these algorithms can perpetuate and even exacerbate existing biases, leading to unfair treatment of certain demographic groups Mehrabi et al. (2021), Chouldechova (2017).

The literature on bias detection and mitigation is primarily focused on quantifying a measure of bias and developing algorithms to mitigate it Zandee (2021). In this chapter, we aim to critically evaluate existing studies on preprocessing techniques that address bias detection and mitigation while exploring evaluation metrics and the applicability of these algorithms in real-world scenarios.

2.1 Literature Selection Criteria

This review examines publications by decomposing our research objective and compiling information from the exploration of smaller related problems. The reviewed papers majorly consist of publications in the domain of Artificial Intelligence (AI), Machine Learning (ML) and Ethics. We explored several databases, journals, and proceedings from Elsevier, Springer, IEEE Transactions, arXiv, and key conferences such as NeurIPS and ACL among others. AI tools such as Paper Digest (PD) and Elicit were used to target solutions and explore work done in different domains. Recognising the amount of research done in this field, we have included preprints, independent and seminal works to ensure comprehensive evaluation.

2.2 Theoretical Framework

The theoretical framework for detecting and mitigating bias in data is primarily focused on FAIR principles. Raza et al. (2024) systematically explores these principles and an overview of their research is summarized in Section 2.2.2.

2.2.1 Definitions

This section has summarised some of the key terminologies used in this field of research. This work is largely based on the work of AIF360 Bellamy et al. (2019), Zandee (2021) and Raza et al. (2024).

A **protected attribute** refers to a label or value that is considered sensitive or prone to discrimination. Using these attributes without proper data augmentations could lead to unfair outcomes. Common examples of these attributes are race, gender, and age, among others.

A **privileged group** refers to a subset of the protected attribute that historically or systematically receives preferential treatment over the others. This is often contrasted by groups that are discriminated against called unprivileged groups. Examples of these factors include race, socioeconomic differences, and disability, among others.

Fairness is used to denote impartial outcomes. In the context of this research, fairness is treated as a quantifiable measure that is summarised as **bias metrics**. These metrics are described in further detail in section 3.3.

Baseline refers to a simple model that is used as a point of reference for more complex models. These are straightforward to implement and can provide an accurate view of the efficiency of preprocessing techniques while using the same model. **Ground truth** is a similar concept which describes a real-world situation. Ground truth is beneficial over baseline in certain scenarios as they benchmark the results not just by how well the algorithm mitigates bias, but it also considers how the results change compared to the situation.

2.2.2 FAIR Principles

The FAIR principles is an abbreviation for the attributes of Findability, Accessibility, Interoperability and Reusability. Given the constraints, the definitions for each of these principles is provided below.

1. **Findability**: This principle ensures that data and resources are easily locatable and accessible. To enhance the readability and discoverability of data, Raza et al. (2024) suggests the use of metadata, persistent identifiers, and promoting efficient data indexing and searchability. It is also important to ensure trustworthiness of the data source to promote findability among different environments.

2. **Accessibility**: This is defined by the ease of obtaining and using the data once located. The application of FAIR principles in the context of accessibility is explored by Lamprecht et al. (2020). The works emphasise the significance of incorporating accessibility considerations into research software, platform support, and implementation challenges Raza et al. (2024).

3. **Interoperability**: This refers to the ability of different systems to efficiently work togetherSantos et al. (2021). It requires standardised data

formats and protocols, enabling easy data exchange and integration across diverse systems Raza et al. (2024).

4. **Reusability**: This is a key component in FAIR data principles that ensures data is stored and documented for future retrieval and reuse Raza et al. (2024). A structured approach with planning FAIRification of data in regards to reusability involves providing rich metadata, legal and ethical considerations, and potential societal impactGonzález-Cebrián et al. (2024).

2.2.3 Measuring Fairness in ML

This section outlines some of the metrics and methodologies used to measure fairness with decision making systems. Understanding these methodologies helps us identify and mitigate bias that enables us to build an equitable model which follows the FAIR principles.

Discrimination Control

Discrimination is the prejudicial treatment of an individual based on membership in a legally protected group such as a race or gender. Direct discrimination occurs when protected attributes are explicitly used in making decisions, which is known as *disparate treatment* in law du Pin Calmon et al. (2017). Quantification and treatment of discrimination on an algorithmic level involves the modification of the training data, the learning algorithm, or the decisions itself Loukas & Chung (2023). These are known as pre-processing, in-processing, and post-processing, respectively. This approach utilises balanced error rates and predictive bias to focus on pre-processing that achieves equitive results.

Based on the work by du Pin Calmon et al. (2017), we arrive at a general formulation as such:

Given a dataset consisting of n i.i.d. samples $\{(D_i, X_i, Y_i)\}_{i=1}^{n}$ from a joint distribution $p_{D,X,Y}$ with domain $D \times X \times Y$. We define the following:

- D: Discriminatory variables such as gender and race.
- X: Non-protected variables used for decision-making.
- Y: Outcome random variable.

Assumptions:

1. D and X are discrete and finite domains.
2. Y is binary, i.e., $Y \in \{0, 1\}$.
3. There are no restrictions on the dimensions of D and X.

For instance, in the context of recidivism prediction using the COMPAS dataset:

- D_i could represent the demographic information (e.g., race, gender) of individual i.
- X_i could represent the prior criminal history and other non-protected attributes of individual i.
- Y_i could represent whether individual i recidivates ($Y_i = 1$) or not ($Y_i = 0$).

The objective is to analyze and model the relationship between (D, X) and Y while considering the potential impacts of discriminatory variables on the outcome.

Equal Opportunity

Equal opprtunity is a metric that is used to capture the benefits of model predictions. This is done by using True Positive Rates (TPR) where true positive implies the scenario in which the model's positive decision aligns with the real observation. This can be denoted as:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where:

TP : Correctly predicted positive instances,

FN : Positive instances incorrectly predicted as negative.

Using this, Equal opportunity can be defined as a scenario in which

$$TPR_0 = TPR_1$$

In practice, we provide some flexibility for statistical uncertainty.

$$TPR_1 - TPR_0 < \theta$$

where:

$$\theta = \text{Threshold}$$

As a ratio, this can be represented as:

$$\frac{TPR_0}{TPR_1} > \theta$$

A statistical representation of this principle is written below:

$$P(Y = 1|A = a, \hat{Y} = t) = P(Y = 1|A = b, \hat{Y} = t)$$

$$\forall a, b \in \text{Dom}(A), \forall t \in \text{Range}(\hat{Y}),$$

where:

Y : Binary outcome variable indicating the occurrence of an event of interest,

A : Protected attribute, such as race or gender,

$\hat{Y}$: Predicted outcome from a predictive model.

Equalised Odds

Unlike equal opportunity, Equalised Odds uses False Positive Rates (FPR). In our study this can be used to represent the number of individuals that do no re-offend but are categorized as potential re-offenders. FPR is denoted as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

where:

FP : Negative instances incorrectly predicted as positive,

TN : Negative instances correctly predicted as negative.

Similar to Equal Opportunity, Equalised Odds can be denoted as:

$$TPR_0 = TPR_1$$

$$FPR_0 = FPR_1$$

Through this definition, we can see that Equalised Odds is a stricter definition is compared to Equal Opportunity. A statistical version of this definition is provided as a reference Hardt et al. (2016), Bellamy et al. (2019), Feldman et al. (2015):

$$P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = b)$$

$$P(\hat{Y} = 0|Y = 0, A = a) = P(\hat{Y} = 0|Y = 0, A = b)$$

$$\forall a, b \in \text{Dom}(A), \forall t \in \text{Range}(\hat{Y}),$$

where:

Y : Binary outcome variable indicating the occurrence of an event of interest,

A : Protected attribute, such as race or gender,

$\hat{Y}$: Predicted outcome from a predictive model.

Disparate Impact

Disparate impact is addressed by the principles of statistical parity and group fairness du Pin Calmon et al. (2017). Also known as adversial impact, it occurs when a neutral policy disproportionately affects a group of observations Feldman et al. (2015). While disparate treatment addresses intentional discrimination, disparate impact occurs unintentionally. This metric is a result derived from the PPP (predicted as positive) rates which can be denoted as:

$$\%\text{Predicted as Positive (PPP)} = \frac{TP + FP}{N}$$

where:

$$\begin{aligned} TP &: \text{True Positive,} \\ FP &: \text{False Positive,} \\ N &: \text{Number of Observations.} \end{aligned}$$

The definition of a fair model given this PPP value is:

$$PPP_0 = PPP_1$$

Again, adjusting to statistical uncertainty,

$$\frac{PPP_0}{PPP_1} > \theta$$

In the United States, there is a legal precedent to set θ at 80% Odyssey (2023). Hence we will be using the same.

2.3 Ethical Considerations

To understand and portray the ethical implications of the research, we conducted a SWOT(Strengths, Weaknesses, Opportunities, and Threats) Analysis. This enables us to evaluate our research and determine stakeholders. Through this analysis, we aim to explore and understand the ethical landscape of the legal system that can help build towards responsible and principled decision making.

1. **Strengths**:

    - *Focus on pre-processing*: It is important to address model performance at the data level. Doing this allows researchers to address issues apparent in the data before constructing and evaluating models Chouldechova (2017). The COMPAS dataset has many issues with the raw data that ProPublica had not addressed. This makes their evaluation of the model unfit for a comparative analysis Hoffmann (2019). Issues such as these must be addressed before using model evaluation and performance strategies.

- *Fair Decision Making*: Implementing bias mitigation strategies can lead to fairer decision-making systems, transcending individual biases and social hierarchies within the COMPAS dataset Corbett-Davies et al. (2017). A successful implementation of a fair model would lead to truly equitable outcomes.

2. **Weaknesses**:

   - *Partial Mitigation*: While pre-processing techniques are beneficial, it can only partially address the issues within the dataset and cannot be considered as standalone solutions Chouldechova (2017). This is an inherent limitation of all research regarding preprocessing algorithms.
   - *Potential for introducing new bias*: Hardt et al. (2016) elucidates upon the risk of introducing new bias within the data while trying to mitigate existing ones. Concerns that algorithms may discriminate against certain groups have led to numerous efforts to 'blind' the algorithm to race Kleinberg et al. (2018). It is crucial to have a through understanding about the research domain and methodologies used to make sure that the algorithms do not introduce non-existent bias.

3. **Opportunities**:

   - *Advancement of ML models*: The successful implementation of bias mitigation strategies can advance the performance and applicability of ML models Kleinberg et al. (2018). Our research can provide insight and new opportunities to develop and improve upon existing ML applications in the criminal justice domain.
   - *Ethical Decision Making*: The use of ethically aligned algorithms in the criminal justice system, particularly when applied to the COMPAS dataset, can potentially improve fairness, accountability, and transparency Chiao (2019). This provides opportunities to address real-world challenges with better efficiency.

4. **Threats**:

   - *Regulatory Challenges*: The lack of regulatory frameworks that defines fairness and the acceptable measures of standard errors poses significant challenges in defining and maintaining ethically fair models Chiao (2019), Hardt et al. (2016). This lack of a standard definition for fairness makes it hard for us to conduct and justify a comparative analysis with other research works.
   - *Ethical Ambiguity*: The lack of standard regulations and guidance in volatile fields of real-world impact makes it hard for organisations to adopt a new effective approach regardless of the results produced-Chouldechova (2017).

- *Change of data over time*: Valavi et al. (2022) reviews the effects of time over the validity of data. The COMPAS dataset comprises of information from 2013 and 2014. There might be significant changes in behavioural mechanics over time.

Chapter 3

Research Methodology

3.1 Research Design

The research design is intended to be a guide for planning and conducting a research study. In this section, we discuss our planned approach to completing the ongoing research.

The initial analysis involves identifying and categorising the different types of bias present in the dataset while exploring our variables and its relations independently. Statistical metrics and algorithms will be employed to quantify and illustrate these biases. This will be considered an initial baseline. Once a practical baseline is determined, we will use some of the data analysis techniques mentioned in Section 3.3. This will then be documented across multiple preprocessing techniques to conduct a comparative study as shown in Figure 3.1.

Through this approach, we also plan to derive a novel pipeline or algorithm that can efficiently mitigate bias. With successful implementation, we plan to compare its performance with the other techniques to check its effectiveness.

3.2 Data Collection

The selection of data needed several considerations. This comprised of using or obtaining data that was proven to be biased while consisting of qualities that could be addressed by this research.

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset is a comprehensive repository of information that is used by judges and parole officers in the United States of America to predict a criminal defendant's likelihood of re-offending (recidivism). This data consists of extensive details such as demographics, prison and jail time, type and details of the offense among others.

Through a public records request, ProPublica obtained two years worth of COMPAS scores from the Broward County Sheriff's Office in Florida Larson

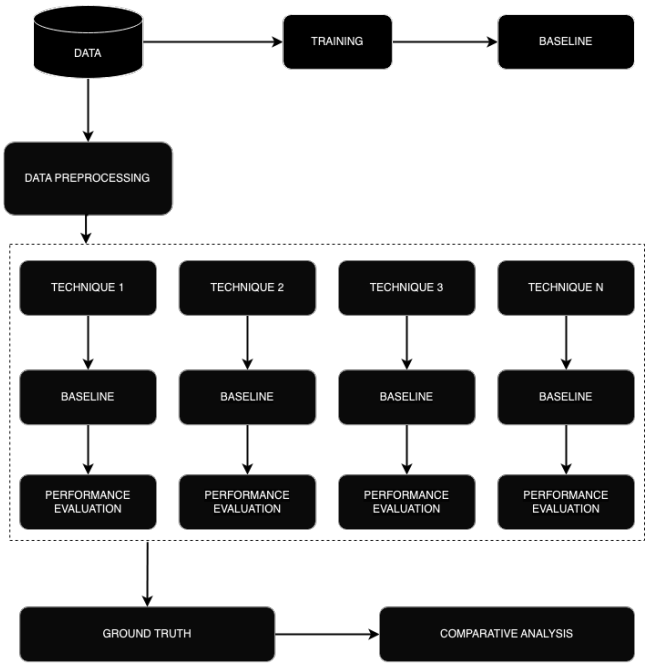

Figure 3.1: Flowchart of the Research Design for Comparative Analysis of Preprocessing Techniques. This diagram illustrates the systematic steps involved, from data collection to the final comparative analysis, ensuring a robust methodology for evaluating different preprocessing methods against a baseline model and ground truth.

et al. (2016). This data consists of all the criminal defendants processed through the COMPAS framework during 2013 and 2014. This data was then made open source, and we were able to obtain a version of it through Kaggle.

The choice of data is justified for our research for several reasons:

1. **Widespread Usage**: Automations withing the criminal justice system can have significant real-world impact. Addressing these issues can lead to fair decision making that can improve and obtain equitable outcomes for varied demographic groups Dressel & Farid (2021).

2. **Existing concerns of bias**: Larson et al. (2016) have investigated the data to provide substantial proof regarding the bias that exists within the data. By focusing on reducing these racial disparities, researchers can help contribute to fairer justice administration.

3. **Comprehensive Data**: The richness of data about the criminal defendants allows us to explore multiple pre-processing techniques on varied dimensions Mehrabi et al. (2021) such as racial and demographic bias among others.

Feature Name	Description	Variable Type
sex	Sex of the individual	Binary
age	Age of the individual	Continuous
race	Ethnicity of the individual	Categorical
priors_count	Number of previous offenses recorded	Continuous
c_charge_degree	Type of offense	Categorical (F: Felony; M: Misdemeanor)
two_year_recid	Binary outcome of whether the individual re-offended within two years of release	Binary(1: re-offended; 0: did not re-offend)

Table 3.1: Description of features in the COMPAS dataset

By decomposing our problem statement, we chose a core data subset from the original version provided by ProPublica. This provisional decision allows us to focus on the core features that lead to the decision outcome. The data card of our subset is shown in Table 3.1

3.3 Data Analysis Techniques

du Pin Calmon et al. (2017), Hardt et al. (2016), Bellamy et al. (2019), Loukas & Chung (2023) in their works, portray various means of pre-processing techniques used to detect and mitigate bias. Through evaluation of these techniques, we conclude that the majority of the approaches fall into these categories:

1. **Re-sampling Techniques**: These techniques manipulating the number of instances in different groups in order to balance their effect on the decision outcome. Some of the well known methods to execute this involve:

   (a) *Under-sampling*: This technique involves reducing the number of observed instances in the privileged group. This helps in preventing the decisions to be discriminating against the unprivileged groups.

   (b) *Over-sampling*: This technique involves increasing the amount of observations in the unprivileged group to help balance the overall group value counts. It is commonly executed by duplicating or generating synthetic observations (SMOTE) of the unprivileged group.

2. **Re-weighting Techniques**: This method focuses on manipulating the contributions of individual instances based on its likelihood to introduce bias. A commonly used method to execute this concept is the *Inverse Propensity Scoring* (IPS) algorithm. The IPS algorithm works by assigning weights to the instances in such a way that the weights are inversely proportional to the instances' likelihood of being sampled or treated in

a biased manner. This results with instances that have lower IPS scores having higher weights which results in debiasing of the classification.

3. **Data Transformation Techniques**: This category of techniques focus on mitigating bias by augmenting the data. Some fundamental concepts like data scaling and stratification among others can be considered effective approaches to mitigating bias in certain scenarios. Also, based on the type of bias being detected, we can use algorithms such as *disparate impact remover* (used for discrimination such as race and gender) and optimised pre-processing pipelines to balance the models biasdu Pin Calmon et al. (2017).

Chapter 4

Preliminary Findings

In this section, we briefly outline some of the significant discoveries made during the Exploratory Data Analysis (EDA). We have comprehensively explored the data to find several key insights regarding the distribution and imbalance among others.

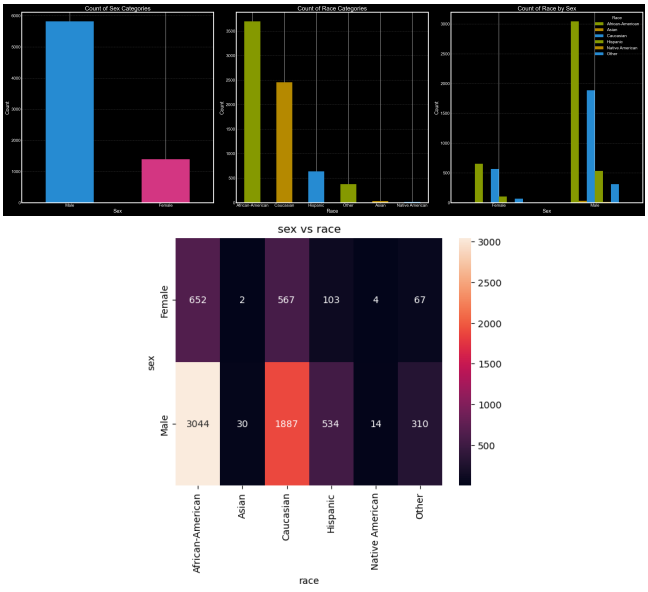

Figure 4.1: Distributions of race and sex along with a contingency table.

The sex and race category were found to have drastic amounts of imbalance as shown in Figure 4.1. Another example to portray the differences are the violent decile (Risk score with a range of 0 - 10) scores . This is provided in Figure 4.2. While there is a significant difference in the distributions, this alone is not enough to capture the the bias present in the variables. Moving

forward, We have provided a few more contingency table that depict significant imbalance between groups shown in Figure 4.3.

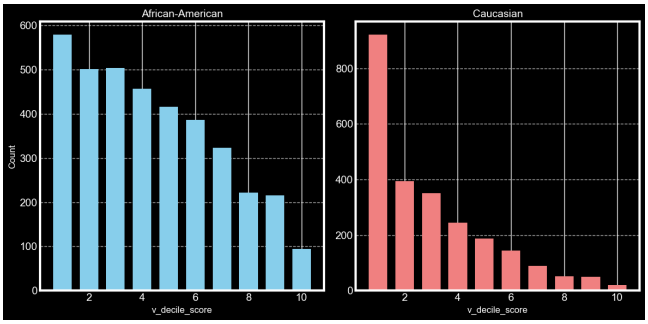

Figure 4.2: Comparison of violent decile scores of Black defendants and Caucasian defendants

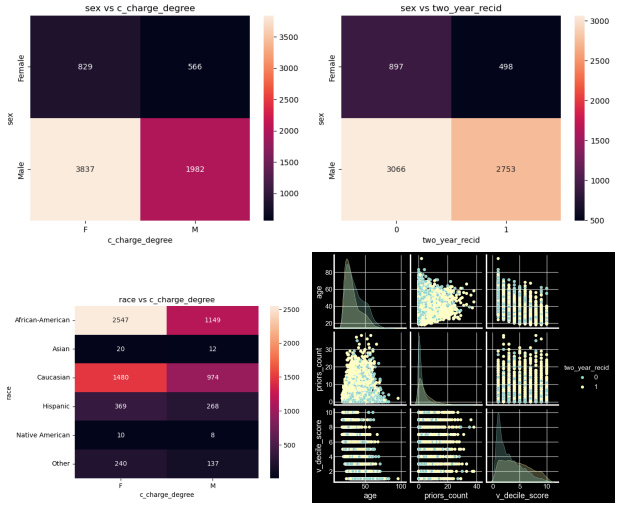

Figure 4.3: Contingency tables that show significant difference between groups and a pairplot summarising all the core variables.

We can observe from the pairplot that there are several instances of group imbalance, skew and outliers. These factors provide opportunities to improve upon bias mitigation using pre-processing techniques.

As an initial examination of bias, we have calculated the odds ratios and found that White (Caucasian) defendants are 72% less likely than Black (African-American) defendants to be categorised as re-offenders. the summary of the Logistic regression is provided in Figure 4.4.

The calculations associated and the code for these visualisations have been provided in the appendix.

```
Optimization terminated successfully.
         Current function value: 0.660308
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 7214
Model:                          Logit   Df Residuals:                     7206
Method:                           MLE   Df Model:                            7
Date:                Wed, 12 Jun 2024   Pseudo R-squ.:                0.04063
Time:                        22:29:42   Log-Likelihood:                -4763.5
converged:                       True   LL-Null:                       -4965.2
Covariance Type:            nonrobust   LLR p-value:                4.399e-83
==============================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                     1.1744      0.080     14.725      0.000       1.018       1.331
age                      -0.0307      0.002    -14.022      0.000      -0.035      -0.026
race_Asian               -0.8563      0.401     -2.136      0.033      -1.642      -0.070
race_Caucasian           -0.3269      0.055     -5.983      0.000      -0.434      -0.220
race_Hispanic            -0.5153      0.090     -5.697      0.000      -0.693      -0.338
race_Native American      0.2261      0.482      0.469      0.639      -0.718       1.170
race_Other               -0.5985      0.115     -5.223      0.000      -0.823      -0.374
c_charge_degree_M        -0.3644      0.052     -7.072      0.000      -0.465      -0.263
==============================================================================
```

Figure 4.4: Model summary of the logistic regression

Chapter 5

Conclusion and Future Plans

The report explores the literature and proposed methodologies to mitigate bias during the pre-processing stages of model development. To explore opportunities in this domain of research, the COMPAS model was chosen. The selection for the dataset was justified through multiple studies showing the inherent bias present in the dataset. A SWOT analysis presented key takeaways of this research while identifying the stakeholders.

Some of the important conclusions from the report include:

1. Studies from Bellamy et al. (2019), du Pin Calmon et al. (2017), González-Cebrián et al. (2024) among others, show the importance of analysing the data and pre-processing it to suit the deployment environment. This ensures equitable outcomes while reducing the likelihood of biased predictions.

2. Techniques such as re-sampling, re-weighting, and transformation methods were found to be the basis of most studies relating to pre processing techniques.

3. In a similar manner, equalised odds, disparate impact, and equal opportunity were found to be the evaluation metrics associated with the field of research.

4. Ethical implications of the research was critically analysed through a SWOT analysis. This provides a structured approach while providing insights into the areas of potential research.

Through a flowchart provided in Figure 3.1, we show the proposed plan of future work. The next steps of this study involves implementing and documenting a comparative analysis of different methodologies. Should a viable concept be developed within the specified time frame, we plan to develop a novel methodology or pipeline to effectively mitigate bias during the pre-processing stages.

Bibliography

Barenstein, M. (2019), 'Propublica's compas data revisited', *arXiv preprint arXiv:1906.04711* .

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A. et al. (2019), 'Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias', *IBM Journal of Research and Development* **63**(4/5), 4–1.

Chiao, V. (2019), 'Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice', *International Journal of Law in Context* **15**(2), 126–139.

Chouldechova, A. (2017), 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', *Big data* **5**(2), 153–163.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. (2017), Algorithmic decision making and the cost of fairness, *in* 'Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining', pp. 797–806.

Dressel, J. & Farid, H. (2021), 'The dangers of risk prediction in the criminal justice system'.

du Pin Calmon, F., Wei, D., Ramamurthy, K. N. & Varshney, K. R. (2017), 'Optimized data pre-processing for discrimination prevention', *CoRR, abs/1704.03354* .

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. & Venkatasubramanian, S. (2015), Certifying and removing disparate impact, *in* 'proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining', pp. 259–268.

González-Cebrián, A., Bradford, M., Chis, A. E. & González-Vélez, H. (2024), 'Standardised versioning of datasets: a fair–compliant proposal', *Scientific Data* **11**(1), 358.

Hardt, M., Price, E. & Srebro, N. (2016), 'Equality of opportunity in supervised learning', *Advances in neural information processing systems* **29**.

Hoffmann, A. L. (2019), 'Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse', *Information, Communication & Society* **22**(7), 900–915.

Janssen, M. & Kuk, G. (2016), 'The challenges and limits of big data algorithms in technocratic governance'.

Kleinberg, J., Ludwig, J., Mullainathan, S. & Rambachan, A. (2018), Algorithmic fairness, *in* 'Aea papers and proceedings', Vol. 108, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 22–27.

Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., Van De Sandt, S., Ison, J., Martinez, P. A. et al. (2020), 'Towards fair principles for research software', *Data Science* **3**(1), 37–59.

Larson, J., Mattu, S., Kirchner, L. & Angwin, J. (2016), 'How we analyzed the compas recidivism algorithm', *ProPublica (5 2016)* **9**(1), 3–3.

Loukas, O. & Chung, H.-R. (2023), 'Demographic parity: Mitigating biases in real-world data', *arXiv preprint arXiv:2309.17347* .

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021), 'A survey on bias and fairness in machine learning', *ACM computing surveys (CSUR)* **54**(6), 1–35.

Odyssey, A. D. (2023), 'Definitions of fairness in machine learning — equal opportunity, equalized odds disparate impact', *YouTube* .

Raza, S., Ghuge, S., Ding, C. & Pandya, D. (2024), 'Fair enough: How can we develop and assess a fair-compliant dataset for large language models' training?', *arXiv preprint arXiv:2401.11033* .

Santos, K. S. S., Pinheiro, L. B. L. & Maciel, R. S. P. (2021), Interoperability types classifications: A tertiary study, *in* 'Proceedings of the XVII Brazilian Symposium on Information Systems', pp. 1–8.

Valavi, E., Hestness, J., Ardalani, N. & Iansiti, M. (2022), 'Time and the value of data', *arXiv preprint arXiv:2203.09118* .

Zandee, R. (2021), 'A comparative study of bias mitigation methods applied on a multitude of classification algorithms', *no. June* .

Chapter 6

Appendix

Use this link to access the code: https://github.com/elton-dsza/Bias-mitigation

ChatGPT commands:

- General structure of an academic interim report focused on literature review
- Most of the formulas were translated to latex code using ChatGPT
- privileged groups in ml
- Is the COMPAS model used anywhere other than America
- Tools to automate and search for literature review
- Simplify the systematic review process and provide task list
- \begin{table} \centering \begin{tabular}—c—c—} \hline Feature& Description\ \hline sex& Sex of the Individual\ \hline age& Age of the Individual\ \hline race& Ethnicity of the Individual\ \hline priors_count& Number of previous offences recorded\ \hline c_charge_degree& Type of offence (F: Felony; M: misdemeanour)\ \hline two_year_recid& Binary Outcome of wether the individual re-offended within two years of release\ \hline \end{tabular} \caption{COMPAS data subset card} \label{tab:data card} \end{table} fix the errors
- difference between baseline and ground truth
- What is theoretical framework and how to structure literature inside it?
- Metrics commonly used for bias detection
- odds ratio interpretation
- set theory fundamentals

- survival analysis and its applications in debiasing
- Kaplan Meier explanation
- logit statsmodels assumptions
- automating EDA for univariate, bivariate and multivariate analysis in python