January 2023

Ai Fairness: Assessment For Prediction Of Agitation In Mental Health Patients At Emergency Rooms

William Pang
williampangbest1@gmail.com

AI Fairness: Assessment for Prediction of agitation in mental health patients at Emergency Rooms

William Long-Hin Pang
Department of Epidemiology of Microbial Diseases
Yale School of Public Health

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF PUBLIC HEALTH

in Epidemiology of Microbial Diseases

Primary Advisor: Richard Taylor, MD, MHS
Secondary Advisor: Ambrose Wong, MD, MSEd, MHS

Abstract

Machine-Learning (ML) algorithms are increasingly used to assist clinicians in decision-making. Recent studies have shown, however, that machine learning algorithms might be prone to unfairness, albeit unintentional. As such, it is important to ensure that algorithms do not discriminate against certain groups. This thesis examines an in-house algorithm used to predict agitation among mental-health patients in the emergency department and addresses biases that might arise from generated predictions.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Machine learning (ML) algorithms are increasingly used in our daily lives, whether it is in autonomous vehicles, recommendation systems, and many other applications. In the clinical space, ML has received significant attention thanks to advances in computing power, combined with new data storage tools, which have allowed for efficient processing of large patient information (colloquially known as "big data") in the form of electronic health records [2]. The general idea is that by harnessing troves of patient data, computer algorithms could identify patterns that are otherwise hidden to clinicians, serving as an invaluable tool to aid physicians in making clinical decisions.

While ML tools offer great promise, recent research has begun to shed light on the potential pitfalls when ML algorithms are used indiscriminately. One famous example is the use of ML algorithms in facial recognition technologies, where researchers found that algorithms performed substantially worse in darker skinned individuals as compared to light-skinned individuals [3]. Another example is how natural language processing algorithms incorporate human-like biases [4], where researchers showed that algorithms made gender-based associations (such as male names and career choices).

In this project, we explore a dataset that was generated by a parent study investigating the use of ML algorithms to predict restraint among psychiatric patients in the emergency medicine department. By understanding patterns in patient data, we hope to perform a risk assessment upon patient intake and recommend pre-emptive strategies to clinicians with the goal of minimizing use of restraint.

We also recognize that healthcare datasets are imbued with inherent biases, perpetuated by structural inequalities, stigma, as well as implicit biases [5]. This is particularly salient in the mental health community, where psychiatric conditions are often stigmatized, resulting in disparate care among historically marginalized and socioeconomically disadvantaged populations [6].

As such, there is a great need to ensure that clinical-support tools are not only accurate, but also perform *fairly and equitably* regardless of race, gender, and economic status. The objective of this thesis is to illustrate strategies that can examine, report, and mitigate bias in a predictive algorithm used for early risk assessment among psychiatric patients.

1.2 Literature Review

Machine learning models are increasingly incorporated in clinical settings, thanks to its ability to process and identify patterns from large datasets spanning information contained in millions of patients. However, many of these datasets are embedded with structural biases, and as such poses a legitimate concern that recommendations made by ML algorithms will lead to a continued perpetuation of structural biases and racism in the healthcare system.

In the fairness in machine learning space, an equally vexing challenge is the lack of a singular definition for *fairness*. Perhaps the most famous example is in recidivism prediction, where algorithms are used to assist judges

in predicting whether an arrested defendant will recidivate. In Pro Publica's investigation of one such algorithm (COMPAS), they found that Blacks were more often flagged as high risk and therefore more often denied bail as compared to Whites. However, the company who created the algorithm (Northpointe) countered Pro Publica's claim and argued that COMPAS satisfied their fairness metric [7], resulting in heated discussions among the online and academic community.

While there's no one-size-fit-all definition for fairness, a few common frameworks on approaching fairness can be defined. The most basic definition of fairness is demographic parity [8], which is the idea that a model's classification is independent of the protected attribute p. In the COMPAS example above, this would mean that risk prediction $\hat{y}$ (i.e., classifying whether a person is high risk or low risk) is independent of the sensitive attribute. Mathematically, this can be expressed as:

$$Pr(\hat{y}|p) = Pr(\hat{y}) \tag{1.1}$$

An extension to the demographic parity framework would be to ensure that our model does not systematically make errors for protected groups; we would want the model to have the same accuracy regardless of the protected attribute p [9]. More concretely, we would want the true positive rate and false positive rate to be roughly equal so that the systematic errors are uniform across each group. Note that this is distinct from the demographic parity framework as we're trying to assess the rate of *errors*, which means that we would need a ground truth label. Digging back to the COMPAS example, this would mean that the algorithm will classify low risk scores and high risk scores at equal rates regardless of the racial group. Mathematically, this can be expressed as:

$$Pr(\hat{y}|y, p) = Pr(\hat{y}|y) \tag{1.2}$$

Note once again that $\hat{y}$ refers to the predicted outcome and y refers to the ground truth outcome.

Finally, we also define the equality of opportunity, which is the idea that the classifier would predict the preferred outcome at equal rates regardless of the attribute [10]. In the COMPAS example above, this would mean that the rate of low risk classification should not depend on a sensitive attribute like race. Mathematically, this can be expressed as:

$$Pr(\hat{y}|y = 1, p) = Pr(\hat{y}|y = 1) \tag{1.3}$$

where y=*1* is the "preferred" outcome.

In practice, choosing which framework to implement poses multiple challenges [11]. First, there will always be some trade-off between fairness and accuracy; we could hypothetically constrain our model in such a way that would make 'undesirable' outcomes extremely rare for either groups, but would result in very low accuracy. Another challenge is that different frameworks might not be simultaneously compatible, as was the case with the COMPAS recidivism example.

Chapter 2

Methodology

2.1 Data Collection

Data for our prediction model was obtained from an existing cohort of 2.1 million patient visits from five clinical sites across the Yale New Haven healthcare system. This is data provided by the parent study led by Dr. Ambrose Wong (secondary thesis advisor).

2.2 Data Cleaning

Data cleaning was mostly processed by YCAS. Briefly, all data that was coded in strings and characters were converted into a numeric type, including missing values (encoded as "NA"s) which were set to an empty character string and then also converted to numeric type. Simple imputations were also performed on missing continuous variables using mean imputation; for categorical variables, a new missing category (coded as "999") was created.

To select for predictors that are meaningful to our model, a hybrid of computational strategies as well as domain expertise was deployed. On the computational side, variable selection was selected after univariate analysis by examining the odds ratio relationship between the predictor and out-

come variable, followed by a stepwise selection of variables. On the domain-expertise side, Dr. Ambrose Wong provided a list of predictors that should be retained. The combination of selection from the computational method and domain-expertise resulted in a final set of variables that will be used to train our model.

2.3 Data Preprocessing

Before feading our data into a machine learning model, we need to divide our dataframe into three parts: a training set, a testing set, as well as a validation set.

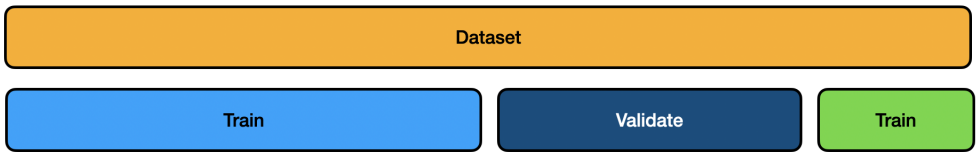

Figure 2.1: An illustration of the dataset splits

The idea of splitting the dataset is to ensure that the model does not overfit to our data. The bulk of the split will be reserved for the training dataset, as we would like to have as much data as possible for the machine learning model to learn from. Generally, this should be at least 50% of the split.

Once the model is trained, we can then check the performance of our results using the validation set. The validation dataset is set aside to provide an unbiased evaluation of the training set. If the model performs poorly on the validation dataset, fine tuning is then made to ensure that the model does not overfit the training dataset. Generally, this should be at least 20-30% of the split.

Finally, the validation dataset is used to provide an unbiased evaluation of the final model. The reason we want to reserve an additional chunk of the

dataset is because the validation dataset might become more biased overtime as the model iteratively "learns" from the validation set as well. As such, the validation dataset is really only reserved when the model is at full-production and we want to make a claim about the model's performance. Generally, this should be about 10-20% of the split.

2.4 Model Selection

Model selection was recommended by the Yale Center for Analytical Sciences (YCAS). We adopted their recommendation by running a logistic regression with a L_1 regularization, which is also known as a lasso regression. The idea behind regularization is to avoid over-fitting, especially in cases where there are a lot of parameters involved [12]. Indeed, the basic intuition behind L_1 regularization is to "penalize" unimportant predictors and make their weights (i.e., contribution to the dependent variable) close to 0. The loss function can be expressed as:

$$Loss = Error(Y - \widehat{Y}) + \lambda \sum_{1}^{n} |w_i| \tag{2.1}$$

where λ represents the regularization rate, which controls how much we want to penalize the weights. Note that when λ is zero, this is equivalent to a regular regression with no penalization. Because L_1 regularization has the power to push weights toward 0, it is commonly used for feature selection among a large set of variables.

2.5 Mitigating Bias

2.5.1 The AIF360 Package

Implementation of bias mitigating strategies is done through an open-sourced tool developed by IBM called AI Fairness 360 (AIF360). The goal of this tool is to allow researchers across disciplines to check their models and ensure their models are as bias-free as possible [13].

As such, the underlying framework behind the implementation of fairness algorithms in this thesis will mostly be based on pre-written functions included in the AIF360 package (more information on the package can also be found on the AIF360 website). We will be implementing pre-processing and in-processing algorithms to address bias in our model.

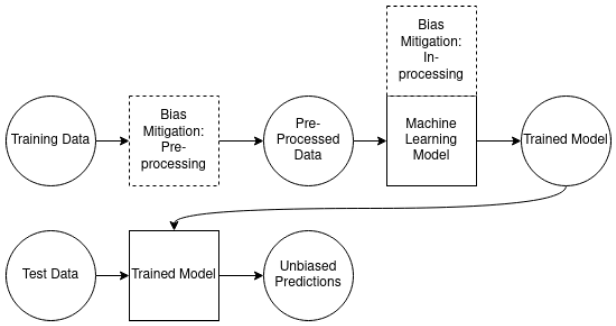

Figure 2.2: Pipeline for mitigating bias [1]

2.5.2 Reweighting

Reweighting is a type of fairness pre-processing algorithm that attempts to coerce the outcomes to be more fair [14]. As the name suggests, reweighting works by weighing unprivileged groups with the "preferred" outcome heavier, and conversely weigh privileged groups with the "preferred" outcome less. In other words, the success and failure are equally weighted within all subgroups

of the population [15]. Mathematically, this points to the idea that group membership g and outcomes should be independent:

$$Pr(\hat{y} \cap g) = Pr(\hat{y}) \times Pr(g) \tag{2.2}$$

The weights are then calculated as follows:

$$W_{\text{positive outcome among privileged}} = \frac{N_{\text{privileged}} \times N_{\text{positive outcome}}}{N_{\text{all}} \times N_{\text{positive outcome among privileged}}} \tag{2.3}$$

$$W_{\text{positive outcome among unprivileged}} = \frac{N_{\text{unprivileged}} \times N_{\text{positive outcome}}}{N_{\text{all}} \times N_{\text{positive outcome among unprivileged}}} \tag{2.4}$$

The intuition here is that if the number of positive outcomes $<$ number of positive outcome among the privileged, the privileged group would be receiving a weighting < 1 and hence be penalized; conversely, if the number of positive outcomes $>$ number of positive outcome among the unprivileged, the unprivileged group would receiving a weighting $>$ than 1.

The advantage of reweighting is that it does not require the removal of sensitive attributes or changing labels to ensure more fairness, which could be a thorny question as the research needs to decide which labels are "privileged" or not. On the flip side, reweighting might lead to criticisms of artificially "forcing" the dataset to conform in a certain way in order to fit a fairness metric.

2.5.3 Prejudice Remover

Prejudice is defined as a statistical dependence between the sensitive variable p and the target variable $\hat{y}$ [16]. We can quantify the degree of prejudice mathematically:

$$\sum_{y,p \in D} Pr[\hat{y}, p] ln \frac{\hat{Pr}[y, p]}{Pr[\hat{y}]Pr[p]} \tag{2.5}$$

A smaller value more strongly constraints the independence between the sensitive variable and target variable.

To address prejudice in our model, we implement an in-processing algorithm that tries to reduce the *degree of disparity* by adding a discrimination-aware regularization term to the unprivileged group. The concept is somewhat similar to the lasso regularization, with the difference being that the penalized term is explicitly specified (as opposed to regularization techniques where the penalized term is automatically determined). In fact, the two regularization techniques can be used in conjunction in the loss function.

Chapter 3

Results

3.1 Measuring Fairness

In the literature review section, we explore several frameworks on defining fairness. Here, we will more formally define fairness by first introducing the disparate impact (DI), which is the probably of restraint given that the patient is in the unprivileged group over the probability of restraint given that the patient is in the privileged group. In probability notation this can be written as:

$$\frac{Pr(Y = 1|D = unprivileged)}{Pr(Y = 1|D = privileged)} \tag{3.1}$$

The choice of sensitive attributes should be clinically informed and backed by literature. Racial disparities in care is well-documented in emergency medicine and mental health care [17], and studies have also shown a "gender-divide" among use of restraint [18]. We also wanted to include whether a patient is a medicaid beneficiary or not, which serves proxy for social economic status and a factor in disparate care.

Using our definition from equation 3.1, we can also visualize bias by plotting out the percent of restraints across various sensitive attributes (Figure

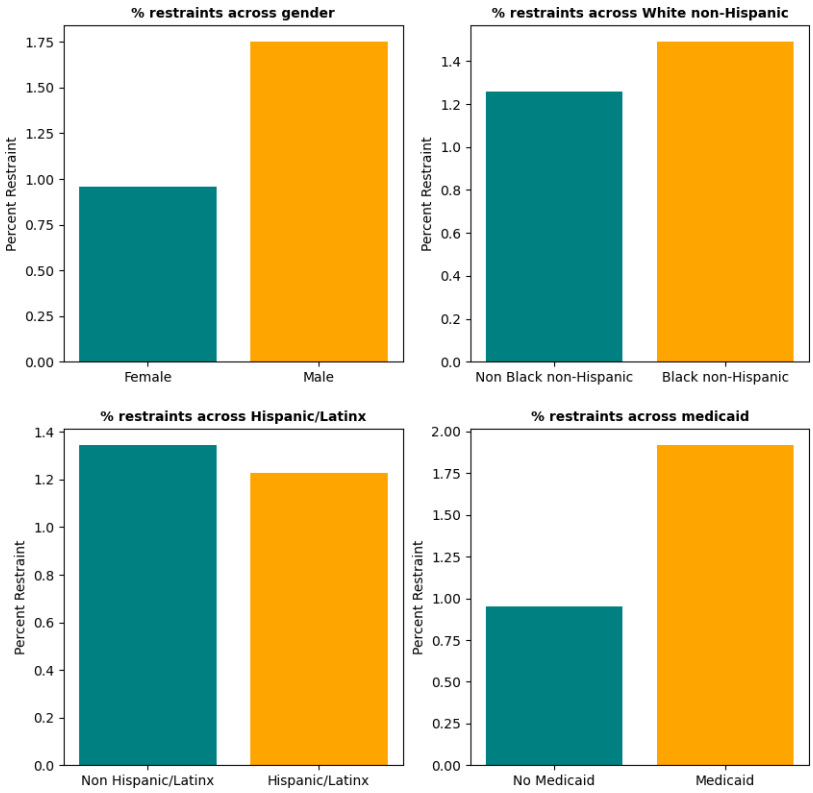

Figure 3.1: Visualizing bias across selected sensitive attributes

3.3). We can clearly see that disparate outcomes are quite pronounced across gender and medicaid status. While it might be tempting to control for all sensitive attributes, we need to be very careful in selecting sensitive variables as there is an inherent trade-off between fairness and accuracy (in general, the more sensitive attributes we add, the more accuarcy we would need to sacrifice).

Another metric that we can use to measure fairness is the average odds difference, which is the average of difference in the False Positive Rate (FPR) and True Positive Rate (TPR) for unprivileged and privileged group. Mathematically, this can be expressed as:

$$\frac{(FPR_{unpriv} - FPR_{priv}) + (TPR_{unpriv} - TPR_{priv})}{2} \tag{3.2}$$

An average odds average close to 0 implies that the measure is fair.

From here on out, we will define a few strategies to mitigate bias and assess their performance while trying to satisfy our fairness metric.

3.2 The Do-Nothing Strategy

Before we evaluate the different strategies to mitigate bias, we should evaluate how the model performs in terms of accuracy and fairness.

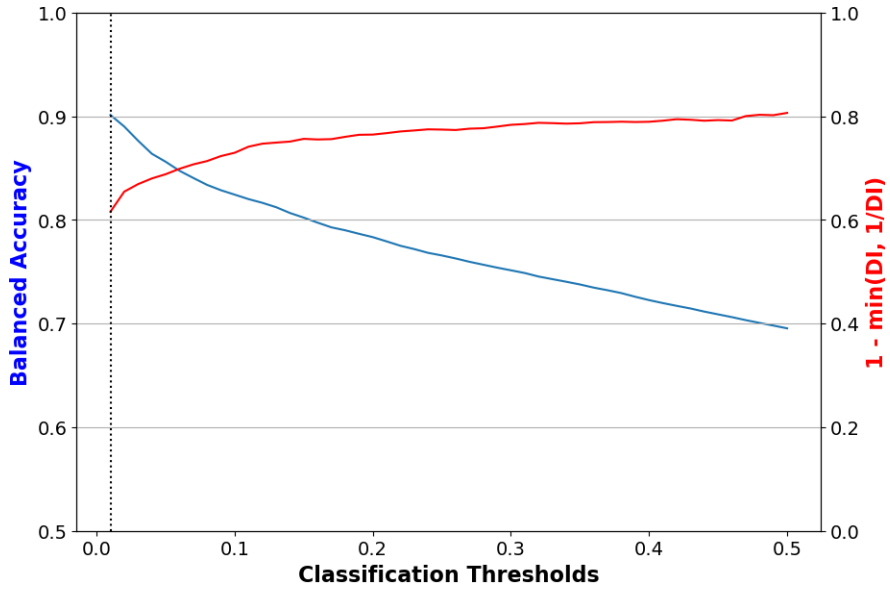

Figure 3.2: Accuracy and fairness trade-off for "Do-Nothing" strategy

Note that here we slightly redefine the disparity index as $1 - min(DI, \frac{1}{DI})$ to bound our fairness metric between 0 and 1. In general, we want the disparity index to be less than 0.2 to be considered fair.

As we can see in Figure 3.2, the model itself yields quite a high accuracy (0.8993) but a very high disparity score (0.6268), suggesting that the model is quite unfair. We can also illustrate the same point with the average odds difference fairness metric:

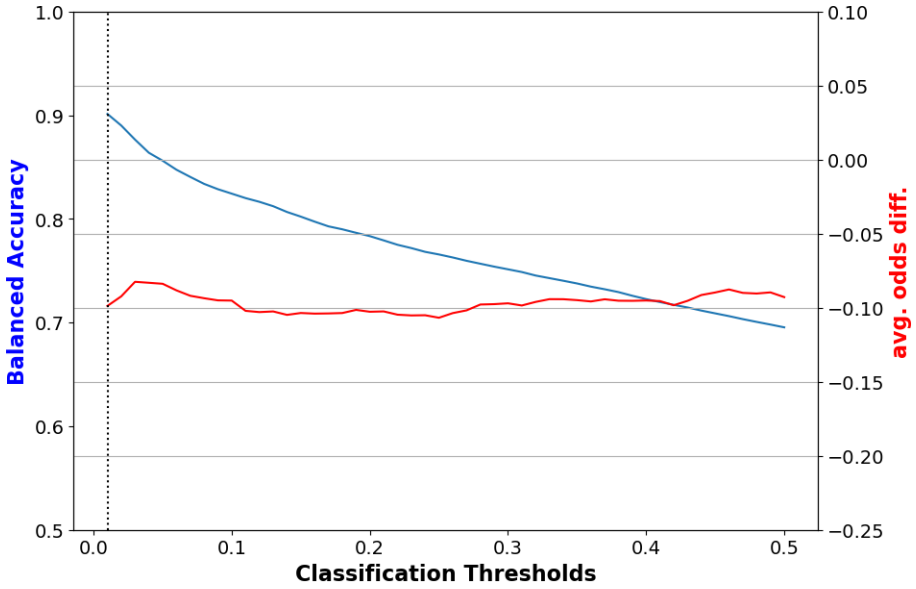

Figure 3.3: Average odds ratio difference for "Do-Nothing" strategy

3.3 Reweighting Weights

As introduced in section 2.5.2, reweighting is a form of pre-processing fairness technique that gives greater weighting to unprivileged groups, while decreasing the weighting for privileged groups. We call the AIF360 package to reweighting method to change the instance weights, and then re-run our fairness metrics.

First off, we see significant improvement in the lower limits of our disparity score in 3.4, which is slightly lower than our "fairness threshold" of 0.2

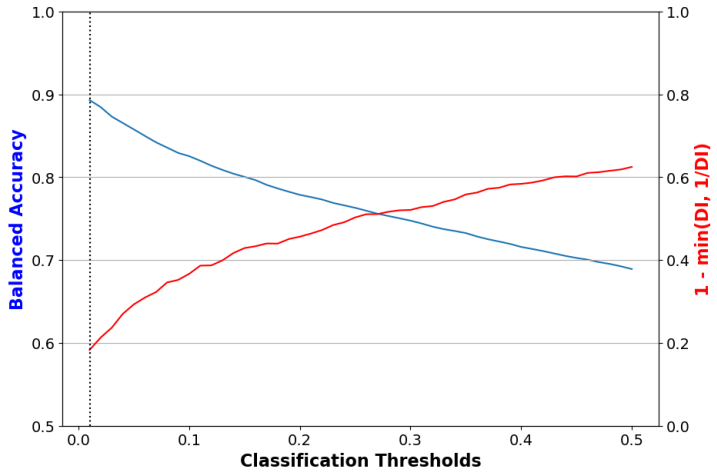

Figure 3.4: Disparity fairness metric after reweighting

for the model to be considered fair. Another important point to note is that our accuracy score is still quite high (0.8934) considering that our model is "fair". This is in contrast to the do-nothing strategy, where we were able to obtain a very high accuracy but our model was also very unfair.

We can also see a similar story when we use a different fairness metric, the average odds difference (Figure 3.5). Again, we see that we were able to obtain high accuracy and maintain an odds difference close to 0. A more subtle point from this graph is the trade off between accuracy and fairness: as we try to move closer to an odds ratio of 0, we see that some accuracy must be sacrificed.

3.4 Prejudice Remover

Prejudice remover (see 2.5.3) is a form of in-processing fairness technique that penalizes the privileged group. Once again, we see that the prejudice remover strategy outperforms the "do-nothing" strategy, as we are able to maintain a low disparate impact score (0.03) without sacrificing accuracy

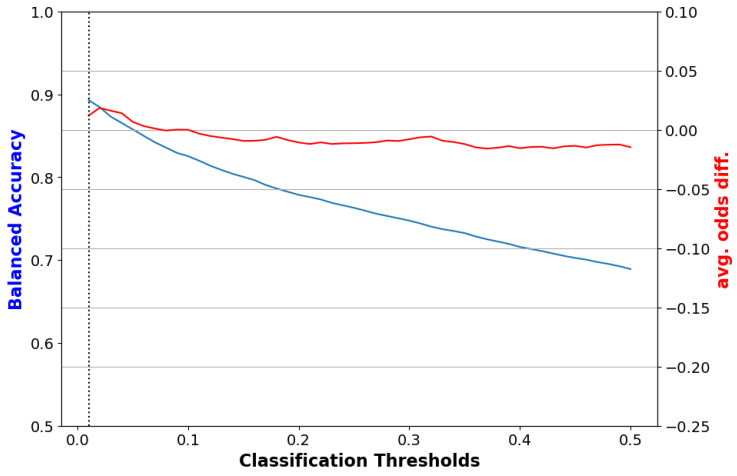

Figure 3.5: Disparity fairness metric after reweighting

(0.8843).

Similarly, we also see that the prejudice remover strategy performs quite well with our second fairness metric, the average odds ratio difference. This validates our results with the disparate impact metric and demonstrates that the prejudice remover is also an optimal strategy to remove bias in our model.

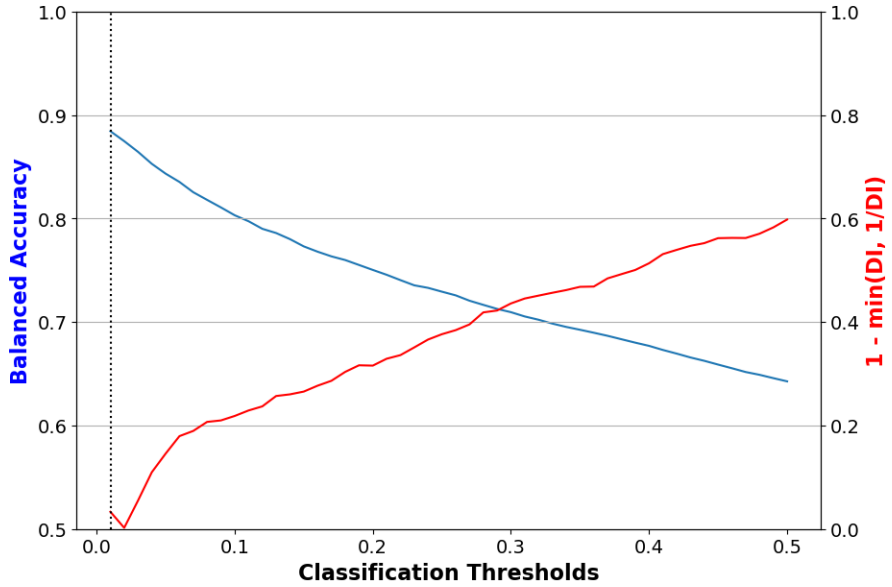

Figure 3.6: Disparity fairness metric after applying prejudice remover

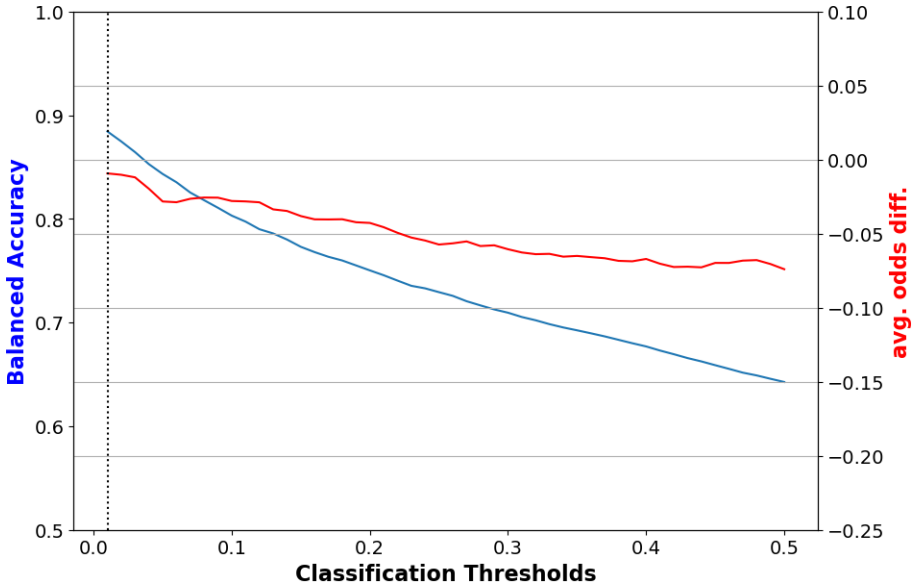

Figure 3.7: Average odds ratio difference after applying prejudice remover

Chapter 4

Conclusions

4.1 Summary of Findings

Overall, we found that both the prejudice remover and reweighting strategies were effective strategies in mitigating bias. Even under two separate metrics (i.e., disparate index and average odds ratio difference), both models satisfied the respective thresholds for fairness while maintaining high accuracy.

Bias Mitigator	**Balanced Acc**	**Disparate Impact**	**Avg. Odds Diff**
Do-Nothing	0.9012	0.6155	-0.0988
Reweighing	0.8910	0.1958	0.0178
Prejudice Remover	0.8843	0.0333	-0.0093

Table 4.1: Fairness metric and model performance across different strategies

4.2 Future Work

To ensure robustness in our work, we would like to test our model with other fairness metrics. Furthermore, it might be useful to examine other regression strategies (e.g., Ridge regression, ElasticNet regression) and compare which strategy would perform better with the addition of more sensitive attributes.

Bibliography

[1] Pablo Mosteiro, Jesse Kuiper, Judith Masthoff, Floortje Scheepers, and Marco Spruit. Bias discovery in machine learning models for mental health. *Information*, 13(5):237, 2022.

[2] Kee Yuan Ngiam and Wei Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273, 2019.

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[4] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[5] Michael Sun, Tomasz Oliwa, Monica E Peek, and Elizabeth L Tung. Negative patient descriptors: Documenting racial bias in the electronic health record: Study examines racial bias in the patient descriptors used in the electronic health record. *Health Affairs*, 41(2):203–211, 2022.

[6] Ambrose H Wong, Halley Ruppel, Lauren J Crispino, Alana Rosenberg, Joanne D Iennaco, and Federico E Vaca. Deriving a framework for a systems approach to agitated patient care in the emergency department. *The Joint Commission Journal on Quality and Patient Safety*, 44(5):279–292, 2018.

[7] Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and behavior*, 36(1):21–40, 2009.

[8] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems*, 30, 2017.

[9] Jie Xu, Yunyu Xiao, Wendy Hui Wang, Yue Ning, Elizabeth A Shenkman, Jiang Bian, and Fei Wang. Algorithmic fairness in computational medicine. *EBioMedicine*, 84:104250, 2022.

[10] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[11] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[12] Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient l˜ 1 regularized logistic regression. In *Aaai*, volume 6, pages 401–408, 2006.

[13] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[14] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

[15] Aileen Nielsen. *Practical fairness*. O'Reilly Media, 2020.

[16] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference,*

ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23, pages 35–50. Springer, 2012.

[17] Kristina Schnitzer, Flannery Merideth, Wendy Macias-Konstantopoulos, Douglas Hayden, Derri Shtasel, and Suzanne Bird. Disparities in care: the role of race on the utilization of physical restraints in the emergency setting. *Academic Emergency Medicine*, 27(10):943–950, 2020.

[18] A Kotze, D Ho, V Bisht, and M Nazir. Physical restraint–the gender divide. *European Psychiatry*, 30:762, 2015.