Universitat de Barcelona

Non-Discrimination in Artificial Intelligence: An Application to Fair Cardiovasular Disease Dignosis

by

Analise Burko

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Fundamental Principles of Data Science

Supervisors:
Dr. Karim Lekadir
Dr. Angélica Atehortua
Vien Dang Ngoc

Abstract

The advent of artificial intelligence within medicine continues to make strides in fast and dependable disease diagnosis, removing the burden off of our clinicians and medical professionals and allowing them to focus on more nuanced and specialized tasks. While the potential of these automated frameworks to do good is expansive, we must criticize the limitations and inherent challenges that such processes pose in light of higher stakes applications where patients' lives and livelihoods are directly impacted. At the forefront of these challenges is bias. Because models must learn from historical data and on a predetermined objective function, they remain vulnerable to perpetuating bias relics and disregarding demographic fairness in the name of increased performance. This project aims to mend these demographic biases and develop a Cardiovascular disease classifier that is fair. We take a three-pronged approach to our development life cycle– building an effective machine learning classification model, in-depth bias evaluation, and a framework of mitigation interventions to result in a model that can effectively diagnose cardiovascular disease at baseline while remaining fair to our identified protected attributes. We investigate and implement a variety of preprocessing and postprocessing mitigation methods to both gradient boosted and deep learning models, successfully managing the fairness-accuracy tradeoff to ensure the equitable sharing of benefits of AI for all.

Table of Contents

Abstract ii

Table of Contents 1

Chapter 1 Introduction 2

1.1 Project Proposal . . . 3

1.2 Roadmap . . . 3

Chapter 2 Background 5

2.1 Burden and Epidemiology of CVDs . . . 5

2.1.1 CVD Risk Factors . . . 7

2.2 Bias in Machine Learning . . . 9

Chapter 3 The Dataset 11

3.1 Feature Selection . . . 12

3.2 Exploratory Data Analysis . . . 13

3.2.1 Protected Attribute Distributions . . . 14

3.3 Data Pre-Processing . . . 15

3.3.1 Data Cleaning . . . 15

3.3.2 Imputation . . . 16

Chapter 4 Model Development 18

4.1 Feature Transformation . . . 18

4.1.1 Imbalanced Data . . . 18

4.1.2 Transformation by Data Type . . . 18

4.2 Model Architecture 19
4.2.1 Learning from Tabular Data 19
4.2.2 XGBoost Model 20
4.2.3 Deep Learning Models 20
4.3 Performance Evaluation 22
4.4 Implementation 23

Chapter 5 Bias Evaluation 25
5.1 What is fair? 26
5.2 IBM AI Fairness 360 Toolkit 27
5.2.1 Fairness Metrics 27
5.3 Privileged and Unprivileged Groups 28
5.4 Bias in Original Models 30

Chapter 6 Bias Mitigation 32
6.1 Pre-Processing Methods 33
6.1.1 Disparate Impact Remover 33
6.1.2 Reweighing 34
6.2 Post-Processing Methods 36
6.2.1 Calibrated Equalized Odds 36
6.2.2 Reject Option Classification 37

Chapter 7 Results & Discussion 38
7.1 The Bias-Accuracy Tradeoff 38
7.2 Discussion 39

Chapter 8 Conclusion & Future Work 42

References 44

Chapter 1

Introduction

The adoption of fairness principles in AI has become increasingly crucial as such methods rapidly expand into a vast variety of domains and tasks, especially those that involve high-stakes decisions. Traditionally, the human decision-maker is able to correct for existing biases using personal judgment and expertise, but now research is outpacing the improvement and optimization of AI algorithms while not maintaining fairness and explainability along the way. While algorithmic bias has found a place in regular discussion of model development, discussion of bias relative to fairness has yet to become an industry standard. Unfair biases often exist in data used to train models as well as in a model's own decision-making algorithm, rendering the topic universal in any space where these models are being developed. Specifically, we speak of the term *fairness* referring to the identification and mitigation of such biases in data to results in model prediction that are fair and that do not ethically discriminate.

This project focuses on the topic of fairness applied to cardiovascular disease diagnosis in the UK. In short, we want to create programs that work for everyone. In terms of demographic disparities, we take a close look at sex, age, and race to actively work against perpetuating common biases against subgroups within these demographics that exist in the real world.

1.1 Project Proposal

The bulk of this project has three core components: (1) model development, (2) bias evaluation, and (3) bias mitigation. At the model development stage we aim to investigate both an XGBoost model, a standard MLP deep-learning model, and a more specialized deep-learning model made for tabular data to perform CVD binary-classification at baseline. These models must achieve high-performance to serve as a foundational framework for our proceeding fairness techniques. A model that has good overall performance is essential so that we may attribute unfair biases as inherent to data and architecture as opposed to shortcomings involving specific model ability.

Once we have sound performance results from our models, it is necessary to determine what fairness actually means in our problem space. Considering the three common perspectives of fairness in machine learning, we decide on a set of bias evaluation metrics to appropriately observe original model performance on fairness and establish methods to reapply to observe the impact of our later bias mitigation interventions [1].

Finally, to counteract unfairness after its been recognized we implement a set of interventions in the form of bias mitigation algorithms, which are procedures for reducing unwanted bias in data or directly on ML models. Because one component of the premise of this project is to take an already well-performing model and make it fair, we follow indications from prior art and choose to exclusively implement preprocessing and postprocessing interventions to correct unfair biases identified in step 2.

At the end of this process we aim to not only have a model that performs well under our problem statement and within fairness guidelines, but also produce a framework for understanding which types of bias mitigation interventions are effective and how we should be characterizing fairness for future development.

1.2 Roadmap

This thesis report is structured into the following sections:

- **Background:** highlights trends in cardiovascular disease and origins of bias to support the motives of this project

- **The Dataset:** explains source and structure of our data, EDA, and data cleaning and preprocessing
- **Model Development:** details the process of model selection, creation, training, and evaluation
- **Bias Evaluation:** methodology behind bias evaluation in project, quantifies bias in data and models
- **Bias Mitigation:** details bias mitigation methods along with discussion and analysis of intervention efficacy on the fairness-accuracy tradeoff
- **Results and Discussion:** presents final results and discusses observations, insights, interpretation, and shortcomings of methods
- **Conclusion and Future Work:** assessment of work, identified future areas for continued and improved development

Link to project GitHub

Chapter 2

Background

2.1 Burden and Epidemiology of CVDs

Cardiovascular disease (CVD) refers to conditions that affect the heart or blood vessels and are normally associated with atherosclerosis and an increased risk of blood clots. CVDs are the leading cause of global burden and mortality as estimated by the Global Burden of Disease (GBD) Study of 2019 [2] who found that prevalent cases increased from 271 to 523 million between 1990 to 2019, deaths from 12.1 to 18.6 million, and years lived with disability (YLDs) from 17.7 to 34.4 million. Though CVD is one of the leading causes of death in the UK, they can most often be prevented by leading a healthy lifestyle. Early diagnosis is critical in improving patient quality of life and allowing actionable measures to be taken sooner, ensuring the best outcomes possible.

ARRHYTHMIAS.
DALYs have increased from 3.79 to 8.39 million and prevalence doubled to 59.7 million from 1990-2019 [2]. Age-standardized DALYs, death rates, and prevalence stayed approximately the same. The increase in total values is mainly due to population growth and aging of the population.

CARDIAC ARREST.
Estimated incidence of 6-7 cardiac arrests per 1000 hospital admissions in the USA and 1.5-2.8 per 1000 admissions in Europe [5]. The annual incidence in europes is 67-170 per 100k inhabitants and has increased over the last two decades.

CARDIOMYOPATHIES.

DALYs have increased from 7.06 to 9.14 million from 1990-2019, and deaths from 238-340k [2]. Age-standardized death rates decreased from 8 per 100k people to 5.6/5.8 per 100k people for men/women, and morbidity from 6.5/4.2 per 100k YLDs. Generally, there exists a larger proportion of cases in men than women and commonly occurs from ages 35 to 39 years, and an overall increased trend with age.

CEREBRAL INFARCTION.
Prevalence, deaths, and DALYs have increased to 101, 6.55, and 143 million respectively from 1990-2019 [2]. Age-standardized death rates and DALYs declined over this period indicating high positive impact of preventative measures. Prevalence has increased through parts of Asia and the USA, with significantly greater risk for men to women.

HEART FAILURE.
Overall occurrence is 3.9 cases per 1000 hospital admissions in the UK, where prevalence increases significantly with age. Heart failure results in high rates of hospitalization and ongoing treatment leading to the consumption of 1-2% of health care expenditure of industrialized countries. Men are generally younger than women when first admitted to the hospital.

ISCHEMIC HEART DISEASE.
182 million DALYs and 9.14 million deaths in 2019 [2]. Global increases in IHD are attributed to population growth and aging, with increases demanding more preventive and therapeutic services as they continue. Global burden is increasing across the board, with age-standardized death rates increasing in the USA and UK which indicates that greater preventionary and health care has not improved long term declines of IHD respectively.

MYOCARDIAL INFARCTION (MI).
The incidence of hospitalized MI has been fairly stable over time with 224-486 incidents per 100k with stark trend differences by age, race and sex, where women and the elderly are particularly vulnerable. There were declines in MI incidence among elderly individuals and some age groups of women. Case fatality rate has declined over time.

PERIPHERAL VASCULAR DISEASE.
Prevalence and deaths have doubled to 113 million cases and 74k deaths from 1990-2019 [2]. DALYs, YLLs, and YLDs followed the same trends. Age-standardized deaths, YLLs, YLDs, and DALYs all remained the same or decreased slightly. Women experienced more cases and increased rates compared to men while men

had higher totals of deaths and DALYs.

In a review by the British Heart Foundation Centre on Population Approaches for Non-Communicable Disease Prevention completed in 2016, researchers investigate epidemiological trends of cardiovascular disease (CVD) through the UK [3]. While mortality due to CVD has decreased by 68% from 1980 to 2013, CVS hospital admissions, prescriptions, and operations have increased. Coronary heart disease remains the single biggest cause of death in the UK. Yet the general decline of CVD has shifted them from the main cause of death in the UK to the second leading cause. So while trends are slowly moving in the correct direction, there is a huge amount of burden to overcome.

The GBD study looks at trends according to the Quality and Outcomes Framework (QOF) and compares mortality, disability adjusted life-years, years of life lost, and years lived with the disability in the inspected timeframe 1990-2013. The GBD found that the overall burden of CVD is declining in the UK, with death rates in England falling by 52%, and the two most prevalent subtypes, coronary heart disease (CHD) and stroke, fallin by 60% and 42% respectively. Prior studies show a similar change with myocardial infarction declining by 50% when considering age-standardized mortality [3]. CHD prevalence has remained stable with a slight decline since 2003 according to QOF data, with stroke also remaining relatively stable. However, the observation overall CVD prevalence trends by age and sex show an increase in CVD prevalence for both men and women over the age of 75 and for men aged 65-75; women aged 45-64 experienced a decrease from 10.8% to 8.4%. Bhatnager et al hypothesize that the increase in CVD prevalence can be attributed to the offset resulting from decreased incidence and increased survival.

While mortality reduction trends across CVD, CHD, and stroke have been significant, each remains heavily burdensome throughout the UK with large increases in hospital admissions and treatment in recent years. This truth calls for more advanced and dependable monitoring of such conditions. Bhatnager et al additionally calls for an increased effort in comprehension of the driving forces behind differed epidemiological trends between different subpopulations to further reduce incidence and mortality resulting from CVD, CHD, and stroke.

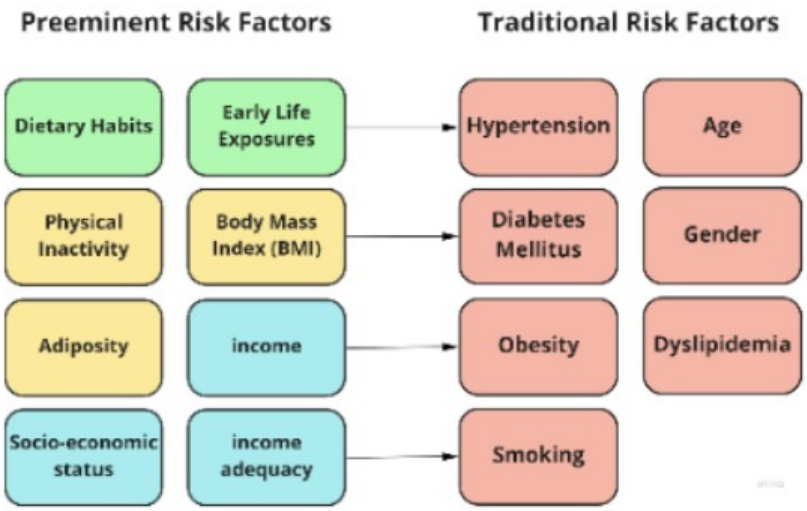
2.1.1 CVD Risk Factors

Though CVD prevalence has trended downward in recent years, there remains a gap in knowledge regarding trends for CVD risk factors related to demographic,

socio-economic and geographic factors that influence these trends. Still it is understood that cardiometabolic, behavioral, environmental, and social risk factors are the main drivers of CVDs.

Lee et al investigated Canadian population health from 1994-2005, identified hypertension, diabetes mellitus, smoking, and obesity as modifiable cardiovascular risk factors [4]. These were chosen as factors of interest as they may be modified by changes in both health policy and lifestyle. Prevalence of these factors is increasing among young people which increases gravity since the presence of risk factors in early and mid life predisposes people to earlier onset of CVD and increased life-years lost.

Traditional cardiovascular risk factors have a number of preeminent risk factors that strongly influence established cardiovascular risk factors. These lifestyle risk factors are pertinent as they also affect novel pathways of risk such as inflammation/oxidative stress, endothelial function, thrombosis/coagulation, and arrhythmia. Therefore basic lifestyle habits should be considered fundamental risk factors for cardiovascular disease. The following outline various examples of these factors and their intermediate and primary effects on CVDs:



Some trends involving CVD risk factors:

- income adequacy – a proxy for disposable household income– has the highest baseline rates of risk factors
- hypertension is more prevalent among people with a higher BMI
- heart disease rates were higher among participants with lower socioeconomic status than among their higher-income counterparts

- presence of risk factors in early and mid life predisposes people to earlier onset of cardiovascular disease and greater potential for life-years lost
- obesity can lead to the development of hypertension and diabetes and is a major predisposing factor for early-onset myocardial infarction

2.2 Bias in Machine Learning

Lack of interpretability acts as the main barrier to the widely accepted adoption of deep learning models in critical domains like healthcare. In addition, due to bias in datasets or models, decisions made by machine learning algorithms are prone to be unfair, where an individual or a group is favored compared with the others owing to their inherent traits. While the term bias refers to any error that has led the model to be unfair, fairness measures the prejudice toward an individual or group based on defined sensitive characteristics or protected attributes. There are a variety of common sources where bias originates, leading a model to be unfair toward a certain group. The following highlight common sources of bias in ML:

HISTORICAL INJUSTICE
Groups have historically experienced injustice and discrimination, whether that be through structural and institutional mechanisms or the unconscious bias in society. Because data is extracted to reflect the population in which it was taken, these biases are reflected as well and integrated into our models since they are fundamentally designed to fit and therefore predict using such training data.

Bias derived from cases of historical injustice can take a variety of forms. A common instance is the increased likelihood of missing or misrecorded data for minority groups. Generally, when the target variable outcome is determined by subjective human intervention, there is historical bias present. Measurement bias is closely linked to historical bias by the defining of labels in the ML process. These labels are meant to act as a proxy for more complex constructs. Measurement bias exists when these proxies poorly substitute the construct they are meant to represent.

PROXY VARIABLES
Proxy variables refer to model features that are highly correlated or associated with the protected attributes. When a proxy variable is used as a model input, the model is effectively making predictions using protected attributes.

UNBALANCED SAMPLES

Unbalanced samples for different subgroups within protected attributes leads to population metrics not being representative of those subgroups where model parameters are skewed toward the majority. Models attempt to find global trends despite subgroups often possessing different relationships between features and target variables, therefore if one group is a majority of the overall population then the model will favor the trend of that group resulting in representation bias– when the development population fails to generalize well to all subgroups.

ALGORITHM CHOICE

The type of AI algorithm used may also impact fairness. Some models, such as deep learning models, are extremely difficult to interpret and therefore difficult to identify and correct for sources of bias. Universally, the objective function that a model trains on determines the behaviors of the resultant model. Most models are trained on a cost function that maximizes accuracy or some other direct error metric across the entire population. Because fairness is not often explicitly optimized, models prioritize overall performance over the trade-off of fairness. This issue is largely attributed to aggregation bias– the model aims to maximize accuracy over an entire population, discounting subgroups in the process.

FEEDBACK LOOPS

Biased models inherently generate more biased data, amplifying the problems already discussed. For instance, if a model is less-accurate against a certain subpopulation that subpopulation will be less-likely to engage with that model, generating even more representation bias than was previously present from the original training data.

Chapter 3

The Dataset

The original dataset is sourced from UKBioBank (UKBB) and contains 502,482 records with 4009 features. Columns were reduced to include only the protected attributes of sex, race, and age, predictor attributes that correspond with CVD risk factors, and target CVDs for diagnosis - *arrhythmia, cardiac arrest, cardiomyopathies, cerebral infarction, heart failure, ischemic heart disease, myocardial infarction, and peripheral vascular disease*. Early model development diverged down two different classification objectives. The first objective was a multi-label classification problem where each of the eight target CVDs are classified independently within the subset of patients determined positive for CVDs (90k records). The second objective is the binary-classification problem that considers the whole population and classifies if a person is positive or negative for any CVD in general. All sections of this report carry forward with the second objective- the binary classification problem. For now the first multi-label objective has only made it through model development, where the performance of the same model architectures to be discussed later are used as ensembles for multi-label classification (further information can be found in the project github). Each of the target CVDs largely share the same risk factors and our predictor features only cover a small subset of the nuanced problem space, this fact paired with highly unbalanced data have served as barriers to reaching a model architecture that may perform reliably. We move forward with the binary problem because it is a more robust and dependable model to serve as a foundation for the fairness evaluation and mitigation components of our development pipeline.

3.1 Feature Selection

Table 3.1: Predictor Features for Model Development

Risk Factor Category	**Feature**	
Physical Measures	Hypertension	Waist Circumference
	Hip Circumference	Diastolic Blood Pressure
	Systolic Blood Pressure	Body Mass Index (BMI)
	Body Fat Percentage	Whole body fat mass
	Whole body fat-free mass	Pulse Rate
	Impedence of whole body	
Sociodemographics	Sex	Qualifications
	Current Employment	Ethnic Background
	Age completed education	
Lifestyle/Environment	Sleep duration	Insomnia
	Current tobacco smoking	Past tobacco smoking
	Cooked vegetable intake	raw vegetable intake
	Fresh fruit intake	Dried fruit intake
	Oily fish intake	Non-oily fish intake
	Processed meat intake	Poultry intake
	Beef intake	lamb intake
	Pork intake	Cheese intake
	Coffee type	Alcohol status
	Variation in diet	Spread type
	Daytime dozing/sleeping	Water intake
	Major dietary changes	Non-butter spread
	Alcohol consumed	Daily alcohol consumption
Mental Health	Freq of Depressed Mood	Anxiety
	Seen a psychiatrist	
Blood Assays	Vascular	APOB
	Cholesterol	CRP
	Glucose	HDL
	LDL	LP-a
	Triglyceride	IGF-1
	Testosterone	HbA1c

These inputs all correspond with supported preeminent and traditional risk factors for CVDs. After selecting for features that do not exceed a 65% threshold of

missing data, these are the final 61 inputs to be used in all model training. These attributed only represent a subset of risk factors and are not comprehensive. Notably, we lack direct inputs related to early life exposures, physical inactivity, income, and socioeconomic status. Including this information for future development would aid in better capturing the problem space.

3.2 Exploratory Data Analysis

First we investigate the distributions of our outcome variables and CVDs as a whole. Figure 3.1 shows that both are severely imbalanced, with overall CVDs accounting for only 9.2% of the entire dataset, the eight CVD outcomes also vary greatly in prevalence with most records indicating Heart Disease or Arrhythmia and CVDs such as cardiomyopathies and cardiac arrest trailing far behind. Recognizing data imbalance will be pertient later in model development for data resampling and in choosing appropriate performance metrics.

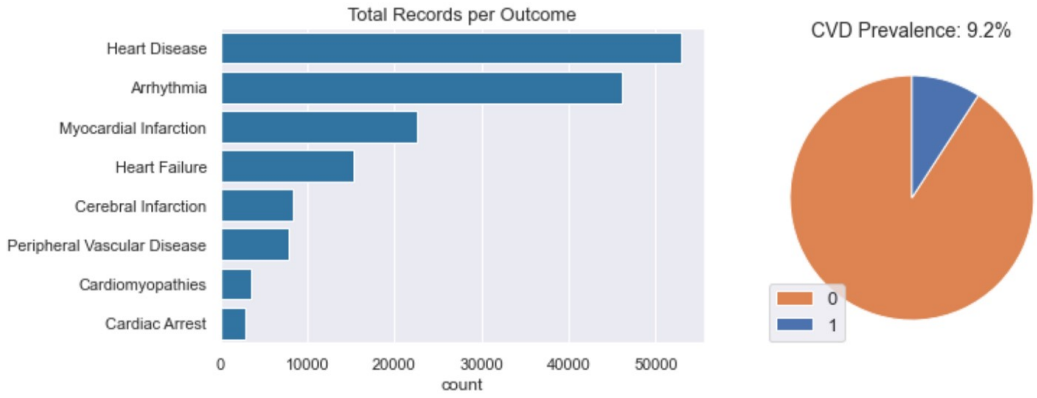


Figure 3.1: Total counts for each of the eight identified CVDs (left). On average each individual positive for at least one CVD is diagnosed with 1.68 of the CVD outcomes. Overall binary CVD prevalence in dataset, approx. 9.2% records are positive

Next we look at a feature correlation heatmap in figure 3.2 and the top features that have the greatest absolute correlation with a positive CVD outcome. We can see that there is very little correlation among our feature space, meaning that there is not an excess of data redundancy and unecessary complexity in our input space. Additionally, we see on the right that the top features correlating with CVDs are all well supported by our background knowing the true risk factors for CVDs where hypertension, age, adiposity, and BMI are all strongly correlating.

Confirming these correlations serves as a good sanity check for our dataset before we move into development.

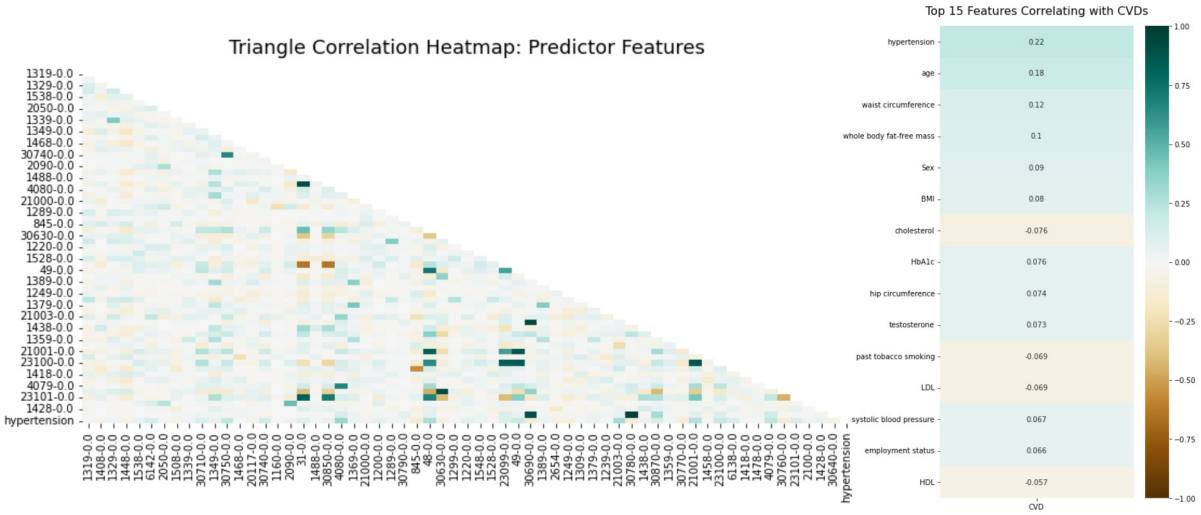


Figure 3.2: Feature correlation triangle heat map (left) and top 15 features that correlate with CVDs (right)

3.2.1 Protected Attribute Distributions

Next we investigate the distributions of each protected attribute, sex, race, and age for the second and third steps of our development pipeline. Figure 3.3 shows that of patients with CVDs there is a slight imbalance toward males which follows societal trends even despite there being slightly more females in the overall dataset. Looking at the entire dataset for race, there is an extreme imbalanced where white participants account for >90% of records.

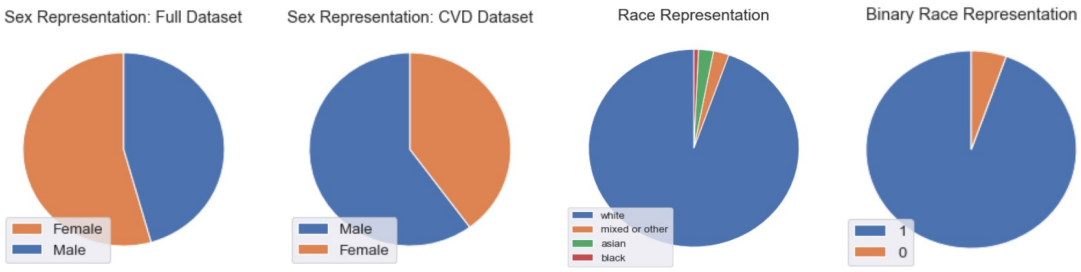


Figure 3.3: Prevalence of Sex and Race subgroups. Sex plots show both for the full dataset(far left) and for sample positive for CVDs (middle left). Race plots show for multi-race grouping (middle-right) and for binary-race grouping (far right) both for the entire dataset.

For age, it is shown in figure 3.4 that CVD prevalence skews toward older individuals as expected. Surprisingly, it is revealed that the entire dataset consists mainly of individuals in their 40-60s. Constraining age in this manner sets a limit on how we may group age as a protected attribute– while young people are often observed separately within CVD epidemiology, we will not get to compare diagnosis for young people in this project.

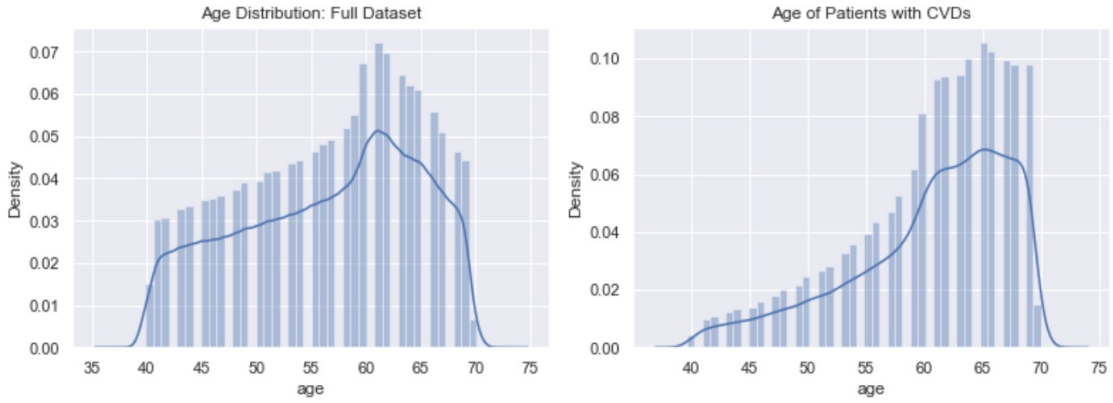


Figure 3.4: Age distribution, full dataset on left, CVD population on right. All records are from the upper 30s to early 70s, CVD patients have a large right skew.

3.3 Data Pre-Processing

3.3.1 Data Cleaning

Once the dataset was trimmed to the necessary records and features, imputation was performed to fill in NaN values. UKBB uses negative values (categorical: -1/-3/-818, numerical: -1/-2/-10) to encode uninformative labels such as Prefer not to answer and Unsure. To refill these instances such responses were first converted to NaN for inclusion into the imputation process. Columns with greater than 65% NaN values were dropped from the dataset, these were ‘Someone to take to doctor when needed’ , ‘Frequency of drinking alcohol’, and ‘Amount of alcohol drunk on a typical drinking day’.

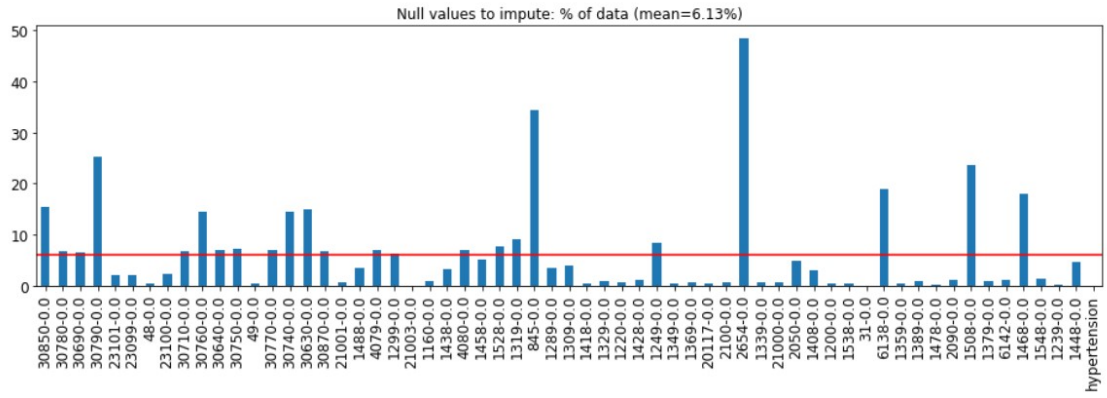


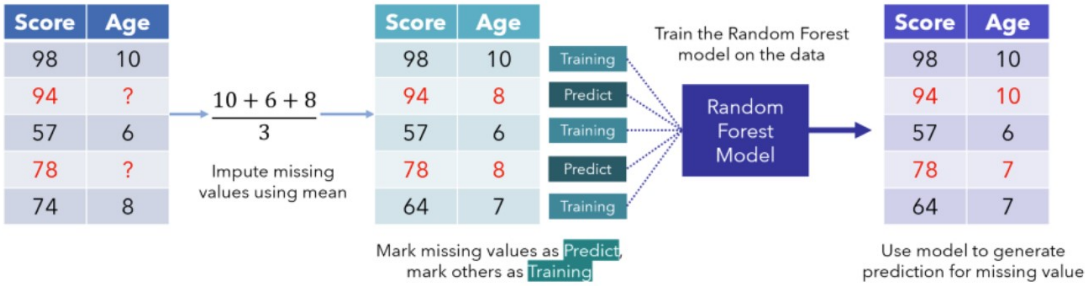
Figure 3.5: Data Preprocessing. Null value rates for each feature before implementation of MissForest imputation. All features with >=65% null values were discarded.

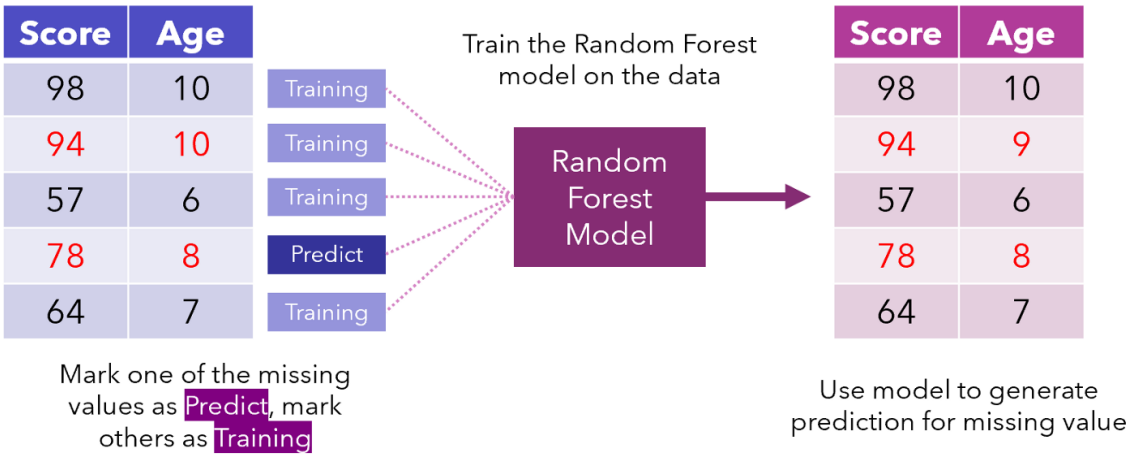
After preparing the data for imputation the graph above shows the percent of NaN labels per feature that need to be imputed. The red line visualizes the mean across all features, showing that on average 6.5% of values need to be imputed per feature

3.3.2 Imputation

MissForest Imputation was the chosen imputation method because of its robust iterative learning scheme and practicality in handling both numeric and categorical data, performing better in prior art against other methods such as KNN-impute and inserting the column mean or average. MissForest is an ML-based learning technique that builds upon Random Forest algorithms to generate values for missing data. In this case, MissForest imputation was performed using the `missingpy` package in python.

Missforest first imputes missing values using the column mode or median. Then other rows are used as training rows into a Random Forest classification or regression model depending on datatype. In our case, data was considered fully imputed after five iterations.


Score
Age
98
10
94
?
57
6
78
?
74
8
10 + 6 + 8
3
Impute missing values using mean
Score
Age
98
10
94
8
57
6
78
8
64
7
Training
Predict
Training
Predict
Training
Train the Random Forest model on the data
Random Forest Model
Score
Age
98
10
94
10
57
6
78
7
64
7
Mark missing values as Predict, mark others as Training
Use model to generate prediction for missing value


Score
Age
98
10
94
10
57
6
78
8
64
7
Training
Training
Training
Predict
Training
Train the Random Forest model on the data
Random Forest Model
Score
Age
98
10
94
9
57
6
78
8
64
7
Mark one of the missing values as Predict, mark others as Training
Use model to generate prediction for missing value

Chapter 4

Model Development

This section details the justification and implementation of the models used through the life of the project. In short, three different model architectures are tested for binary classification on our tabular dataset. An XGBoost model, Multi-layer perceptron, and TabNet model are each tested to compare diagonostic efficacy.

4.1 Feature Transformation

Preparation of training data to optimize model performance.

4.1.1 Imbalanced Data

Because the target CVD column is highly unbalanced with the positive class accounting for only 9.2% of outcomes, it is necessary to resample the training data so that the model may learn in a balanced manner on both outcomes. We tested three different sampling methods: ADASYN resampling, oversampling, and undersampling using the `imblearn` package. Undersampling proved to be the most effective.

4.1.2 Transformation by Data Type

Standardization was performed on all numeric variables using `scikit-learn` `QuantileTransformer`, this transformation forces each numerical variable into a

Gaussian distribution by smoothing the relationship between observations and then mapping those observations onto a normal distribution. This process stabilizes our numerical outputs by scaling values to a standard range and also controls for outliers.

Categorical features were originally provided ordinally encoded from UKBB, whose values can be cross-referenced on the UKBB website. Because all categorical features were low cardinality, it was not necessary to account for sparsity using embeddings. Some categorical features possessed a ordinal relationship *ie 0-never, 1-rarely, 2-sometimes, 3-often*, these variables were all left untouched. Categorical features that possessed nomial relationships were one-hot encoded. Lastly, the feature for hypertension was altered from a datetime datatype to binary to indicate only if that record was positive.

4.2 Model Architecture

To tune the hyperparameters, we split the datasets into training, validation, and test sets. Specifically, we performed a random stratified split of the full training data into train set (80%) and validation set (20%) We selected the set of hyperparameters corresponding to the smallest loss on the validation set for the final configuration. For all models, early stopping is applied using the validation set.

4.2.1 Learning from Tabular Data

Gradient boosted machine learning models have been established as a gold-standard for learning on tabular data due to their superior performance. These algorithms grow successive trees that systematically reduce the error of the model at each iteration. When adding new models, it uses a gradient descent algorithm to minimize loss. On the contrary, deep learning models have not become an established architecture solution for structured data. Deep learning models for classification are particularly challenging on heterogeneous tabular data. The heterogeneity contains dense numerical and sparse categorical features as well as features being more weakly correlated. Heterogeneous data contains a variety of attribute types, such as continuous or discrete numerical attributes from different distributions resulting in missing or complex irregular spatial dependencies [5]. Additionally, the concepts of fairness and explainability are fundamentally interlinked in the topic of algorithmic trustworthiness, and deep-learning produces black-box models which

serves as a direct challenge to explainability.

However, there have been more recent developments to improve DL options on tabular data while providing options for local and global explainability. A primary contender in this space is the TabNet architecture. Understanding the composition of prior-art in regards to learning on tabular data, we investigated an XGBoost model because it is standard practice, a MLP model as a standard deep learning model that appears commonly in fairness studies, and the TabNet model as an alternative DL choice to compare.

4.2.2 XGBoost Model

XGBoost is a gradient-boosted algorithm that is technically an ensembling technique, training decision tree models in succession. Each model is trained to correct the errors of the previous one which proves to be advantageous to the traditional ensemble that trains each model in isolation. Gradient boosting refers specifically to training models to predict the residual of prior models.

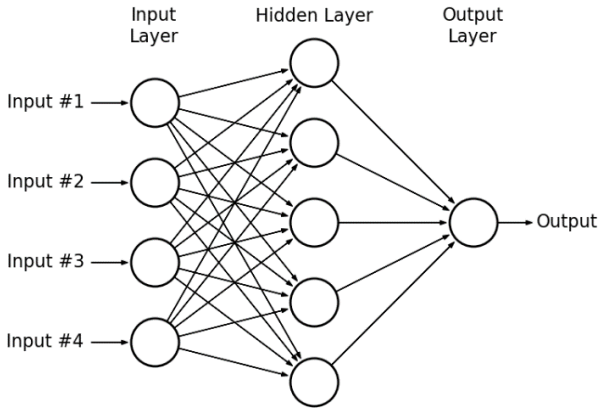
Because of the inherent large complexity of this ensemble model, XGBoost is particularly prone to overfitting. We control for overfitting in this model through the `eta` and `gamma` parameters. eta reduces model complexity by reducing residual weights and gamma works by specifying the minimum reduction in the loss required to make a further partition on a leaf node of the tree. Hyperparameters are tuned using `sklearn's GridSearchCV`.

4.2.3 Deep Learning Models

Multi-Layer Perceptron (MLP)

A Multilayer Perceptron has input and output layers, and one or more hidden layers with many neurons stacked together. MLP is a feedforward algorithm because a linear combination of inputs combined with initial weights and subjected to an activation function is propagated to the following layer– in this way each layer feeds the next. The mechanism of backpropagation allows the MLP to learn by iteratively adjusting network weights to minimize the cost function.

Our MLP model was built in tensorflow and consists of 3 dense hidden layers with tanh activations, followed by a sigmoid activation output for binary classi-



fication. The model's objective loss-function is binary-crossentropy. For training our model, we use the Adaptive moment estimation (Adam) optimizer which acts as a replacement for stochastic gradient descent for training deep learning models. The main benefit of Adam is that it is an optimization algorithm that is able to handle sparse gradients on noisy problems.

Batch normalization layers are introduced between all dense layers for preventing input-related sensitivity common to DL on tabular data and the prevention of overfitting. Batch normalization works by normalizing the output of a previous activation layer by applying zero mean and unit variance, addressing covariate shift and allowing the layers of the network to learn independently.

TabNet

TabNet is a deep learning end-to-end model that has been shown to perform competitively against XGBoost models on tabular data. It includes an encoder, in which sequential decision steps encode features using sparse learned masks and select relevant features for each row using the mask (with attention). Using sparsemax layers, the encoder forces the selection of a small set of features. The advantage of learning masks is that feature selection need not be all-or-nothing. Rather than using a hard threshold on a feature, a learnable mask can make a soft decision, thus relaxing classical(non-differentiable) feature selection methods. This project uses the `tabnet-pytorch` implementation and observes competitive results against the other model options. Because performance of the TabNet model did not greatly exceed the other DL option, we do not move forward with TabNet beyond this point.

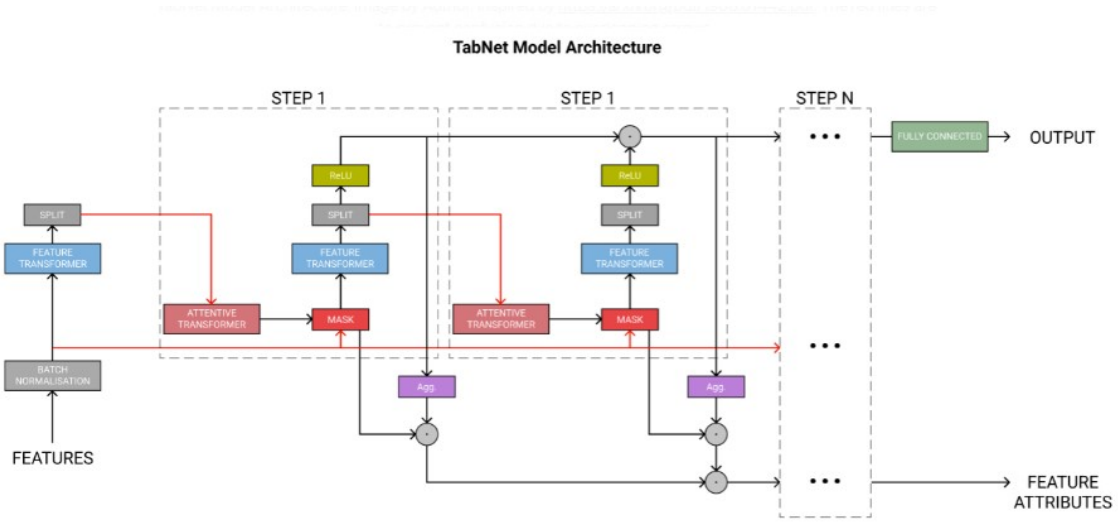


Figure 4.1: TabNet architecture schematic

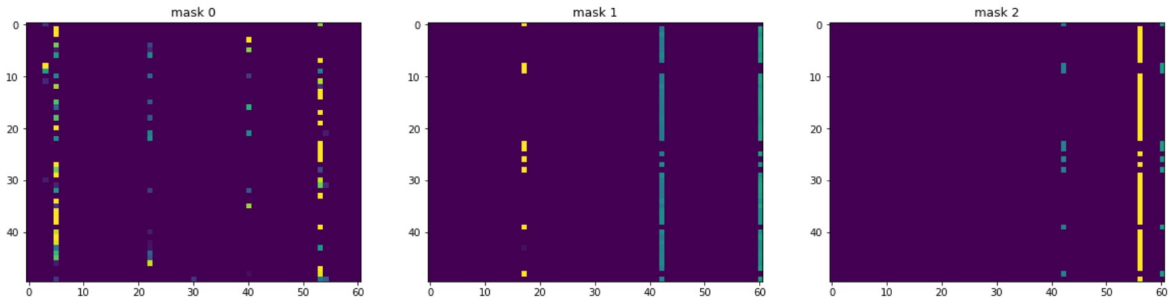


Figure 4.2: TabNet local explainability masks. Each mask represents the feature importance for model learning at its corresponding layer. Graphs indicate weights on each of our 61 input features (X-axis) for each of the individual rows (y-axis)

4.3 Performance Evaluation

Since this problem is a case of medical diagnostics, we prioritize correctly diagnosing the people who actually have CVDs so that they can get treatment, therefore we are willing to allow a few more false positive cases to ensure the best ultimate health outcomes for everyone. Therefore we prioritize sensitivity as our main performance metric. Because our dataset is highly unbalanced where the positive class accounts for only a small percentage of overall records we also prioritize balanced accuracy to ensure fair representation between classes.

Performance Metrics:

- **Accuracy:** $\frac{tp+tn}{tp+tn+fp+fn}$
- **Balanced Accuracy:** $\frac{sensitivity+specificity}{2}$
- **Sensitivity:** $\frac{tp}{tp+fn}$
- **Specificity:** $\frac{tn}{tn+fp}$

4.4 Implementation

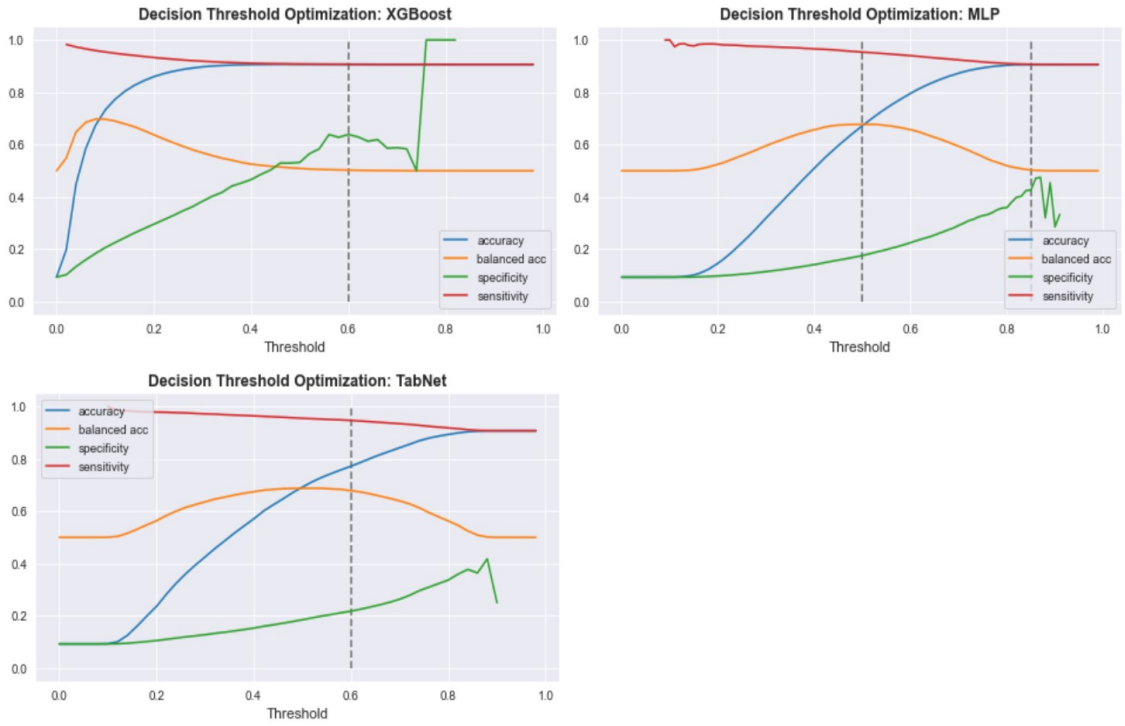


Figure 4.3: Optimization of classification decision threshold for XGBoost, MLP, and TabNet models (metrics: balanced accuracy, accuracy, specificity, and sensitivity), Chosen: XGBoost–0.6, MLP–0.5, TabNet–0.5

Table 4.1: Model Performance at Chosen Thresholds

Model	Threshold	Accuracy	Bal Accuracy	Sensitivity	Specificity
XGBoost	0.6	0.91	0.51	0.91	0.64
MLP	0.5	0.67	0.68	0.95	0.18
TabNet	0.5	0.69	0.69	0.96	0.18

By optimizing for balanced accuracy and sensitivity we determine the optimal classification decision threshold and corresponding baseline metrics as shown in proceeding table. We find that the two deep learning models perform the same on these metrics at 0.68 and 0.95 respectively. Though the XGBoost model performs poorly in balanced accuracy we move forward with the model to observe the effect of intervention techniques on balanced accuracy and as it performs well on all other metrics.

Chapter 5

Bias Evaluation

This section outlines the methodology behind how we determine fairness in this project. A set of standard fairness metrics is presented as well as a framework for interpreting these metrics to relate them specific protected attributes in the real world.

Terminology:

- **Favorable Label:** a label whose value corresponds to an outcome that provides an advantage to the recipient
- **Protected Attribute:** an attribute that partitions a population into groups whose outcomes should have parity
- **Privileged Value:** a protected attribute value indicating a group that has historically been at a systemic advantage
- **Fairness Metric:** a quantification of unwanted bias in training data or models
- **Discrimination/unwanted bias:** undesirable bias or discrimination, which is when specific privileged groups are placed at a systematic advantage and specific unprivileged groups are placed at a systematic disadvantage

In general we can consider fairness through the following lenses: **Data vs Model** fairness may be quantified in the training dataset or in the learned model

Group vs Individual
group fairness partitions a population into groups defined by protected attributes

and seeks for some statistical measure to be equal across all groups. Individual fairness seeks for similar individuals to be treated similarly

WAE vs WYSIWYG
we are all equal or *what you see is what you get*–WAE says that fairness is an equal distribution of skills and opportunities among the participants in an ML task, attributing differences in outcome distributions to structural bias and not a difference in distribution of ability. WYSIWYG says that observations reflect ability with respect to a task

5.1 What is fair?

The concept of fairness is extremely nuanced and subjective to a problem domain. For the application of this project, understanding the biological mechanisms that drive CVD prevalence is essential in striking a balance of understanding on how the problem exists in the real world while simultaneously understanding how data bias influences model learning and results in unfairness. We take a group fairness approach to compare subgroups within each protected attribute as a whole *ie male vs female as opposed to looking at individual females independently of individual males.*

Firstly, sex and age are two of the primary risk factors of CVD and provide great insight for diagnosis. There are substantial differences in the prevalence and burden of different CVDs according to sex [6]. For both sexes, heart disease is the leading contributor of CVD mortality, however the absolute number of women dying from CVD and stroke exceeds men. Despite increased mortality, men experience higher CVD prevalence. These facts are interlaced with complex cofactors, such as women having a longer life expectancy in general and therefore there are more older women, where increased age is a risk factor for CVDs. On the contrary, there are not underlying biological mechanisms that drive differences in CVD diagnosis between races. Race is commonly inaccurately represented in clinical data due to a variety of structural influences. In this dataset, non-white races are severely underrepresented partly due to the composition of race in the UK, which is over 90% white. This attribute imbalance means that our models are likely to prioritize correct diagnosis of white patients if fairness remains unchecked. Because the differences between sexes and people of different ages are legitimate, our model should still reflect those differences when making predictions. Therefore we prioritize metrics that aim for equal predictive power for each

of the subgroups of race and age independently, not metrics that make the same predictions regardless of sex and age. These metrics are average odds difference and equal opportunity difference. Alternatively, in the case of sex, we lean toward a WAE approach because we do not want race to influence our predictions.

5.2 IBM AI Fairness 360 Toolkit

Both this evaluation stage and the following bias mitigation stage utilize the **IBM AIF360** toolkit [7]. This package includes a comprehensive set of fairness metrics for datasets and models, explanations for these metrics, and algorithms to mitigate bias in datasets and models

One should note that the apparent importance of a sensitive feature does not reliably reveal anything about the fairness of a model. This idea is well explained by the concept of *redlining*. Redlining is a term originally coined for the denial of certain services, such as loans, to inhabitants of certain racially determined areas [8]. So while banks were not discriminating against race directly, there were indirectly via the proxy variable of location. Fittingly, the term redlining also refers to the ineffectiveness of the direct removal of the protected group from the input data to mitigate algorithmic bias. Removing the protected attributes would not solve the problem due to the red-lining effect, but rather aggravate it, as the discrimination still would be present, only it would be better hidden. Knowing this we must establish evaluation and mitigation methods that consider these more complicated relationships.

5.2.1 Fairness Metrics

The following fairness metrics from AIF360 are used to evaluate fairness for all protected attributes.

- **Average Odds Difference (AOD):** measures the bias by using the false positive rate and the true positive rate

  $$AOD = \frac{1}{2}[(FPR_{D=unprivileged} - FPR_{D=privileged} + TPR_{D=privileged} - TPR_{D=unprivileged}]$$

- **Disparate Impact (DI):** compares the proportion of individuals that receive a favorable outcome for two groups, a majority group and a minority

group

$$DI = P(\hat{Y} = 1|A = minority)/P(\hat{Y} = 1|A = majority)$$

where $\hat{Y}$ are the model predictions and A is the group of the sensitive attributes

- **Equal Opportunity Difference (EOP):** measures the deviation from the equality of opportunity, which means that the same proportion of each population receives the favorable outcome

$$EOP = P(\hat{Y} = 1|A = minority) - P(\hat{Y} = 1|A = majority; Y = 1)$$

- **Statistical Parity Difference (SPD):** measures the difference that the majority and protected classes to receive a favorable outcome

$$SPD = P(\hat{Y} = 1|A = minority) - P(\hat{Y} = 1|A = majority)$$

- **Theil Index:** measures an entropic distance the population is away from the 'ideal' state of everyone having the same outcome

$$Theil\ Index = \frac{1}{n}\Sigma_{i=1}^{n}\frac{b_i}{\mu}ln\frac{b_i}{\mu},\ \ where\ \ b_i = \hat{y}_i - y_i + 1$$

Table 5.1: Fairness Metric Thresolds

Metric	**Optimal Value**	**Acceptable Range**	**Interpretation**
DI	1	0.8 to 1.0	<1 favors privileged group >1 favors unprivileged group
SPD	0	-0.1 to 0.1	<0 favors privileged group >0 favors unprivileged group
AOD	0	-0.1 to 0.1	<0 favors privileged group >0 favors unprivileged group
EOD	0	-0.1 to 0.1	<0 favors privileged group >0 favors unprivileged group
Theil Index	0	-	bias increases as score increases

5.3 Privileged and Unprivileged Groups

While we have historical intuition for which groups are likely privileged and unprivileged for each protected attribute, we must first confirm statistical significant

differences between the predictions for each subgroup as well as determine which groups are privileged and unprivileged in both models predictions.

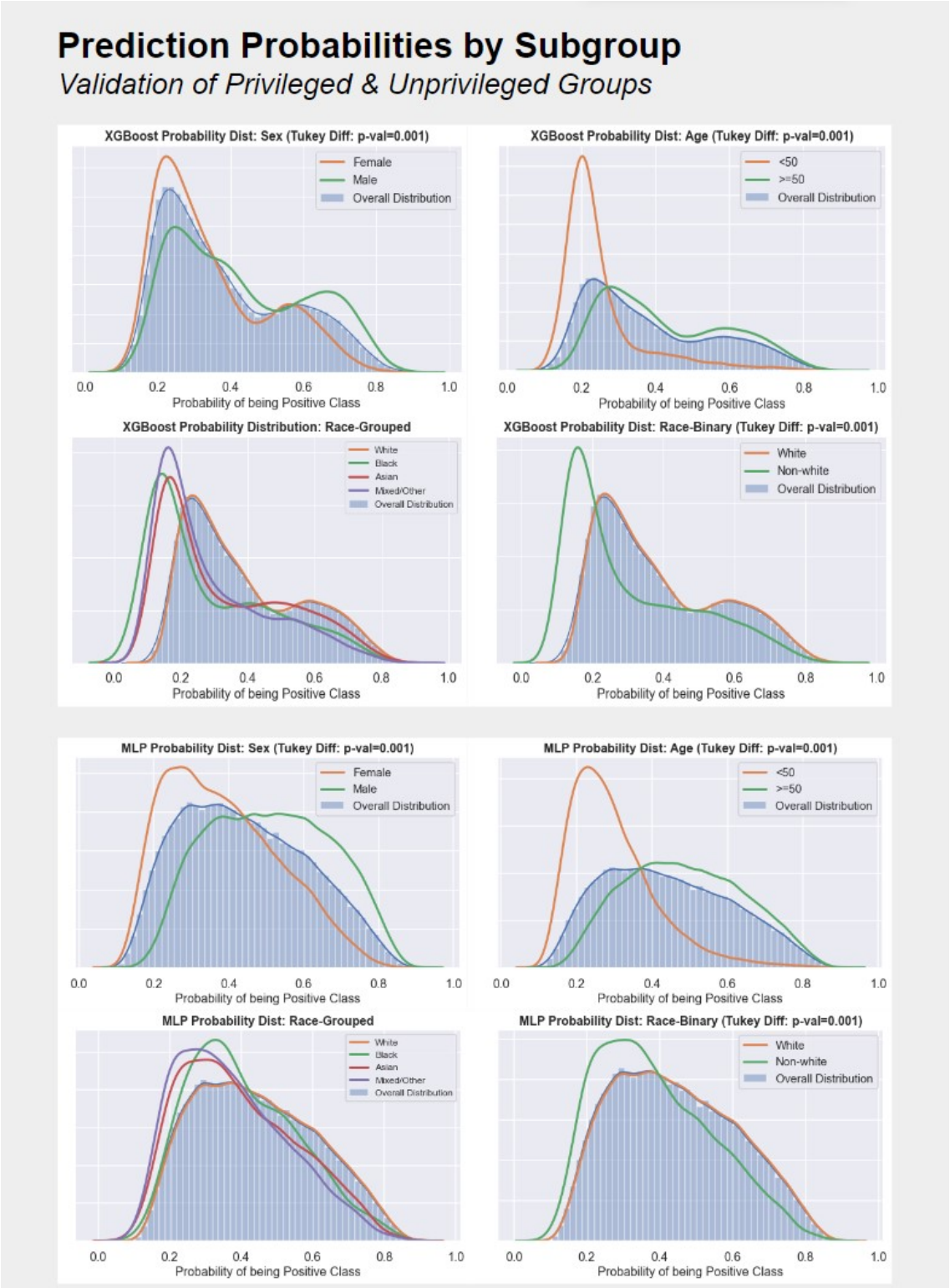


Figure 5.1: XGBoost and MLP prediction probabilities for each protected attribute separated by subgroup.

Tukey's HSD (honest significant difference) test was used on the predicted probability distributions of each subgroup to determine if the pair-wise means are significantly different from each other. Tukey's HSD test is an post-hoc test based on the studentized range distribution, indicating which specific group's means are different in a pairwise fashion. The relevant statistic is:

$$q = \frac{\bar{x_{max}} - \bar{x_{min}}}{s.e.} \quad where \;\; s.e. = \sqrt{MS_w/n}$$

Critical values are found in the *Studentized Range q Table* based on selection for the largest pairwise contrast *(a=0.05)* and number of groups , *k*. If it is found that $q > q_{erit}$ then the two means are determined significantly different. We tested pairwise difference for each of the plots shown in the figure previously and found that each binary split of protected attributes rejected the null hypothesis (p=0.001<0.05) and are significantly different. Tukey's HSD test for the race-grouped plots ,which split race into four groups (white, black, asian, and mixed/other), support the null-hypothesis between asian and black subgroups for the MLP model and black and mixed/other for the XGBoost model. Because differences between all four of these groups are not supported by either model, we move forward with race split into binary white/non-white subgroups.

Table 5.2: Privileged and Unprivileged Classes

Protected Attribute	**Privileged**	**Unprivileged**	**Tukey: mean difference**
Sex	Male	Female	0.0854
Age	>50 yrs old	<50 yrs old	0.1783
Race	White	Non-white	0.0577

5.4 Bias in Original Models

The final fairness metrics shown in figure 5.2 reveal considerable bias for both original models. The XGBoost model is biased for all metrics for sex, all but disparate impact for age, and only disparate impact for race– remembering that the Theil index is on a absolute sliding scale and exerts some bias for all. The MLP model is considered biased for all metrics under all protected attributes. As expected, all biases favor our predetermined privileged subgroups– male, white, and >50.

Figure 5.2: XGBoost and MLP original fairness results– plots where metric falls within the bias range shown in red. Metrics from left to right: disparate impact, average odds difference, statistical parity difference, equal opportunity difference, theil index

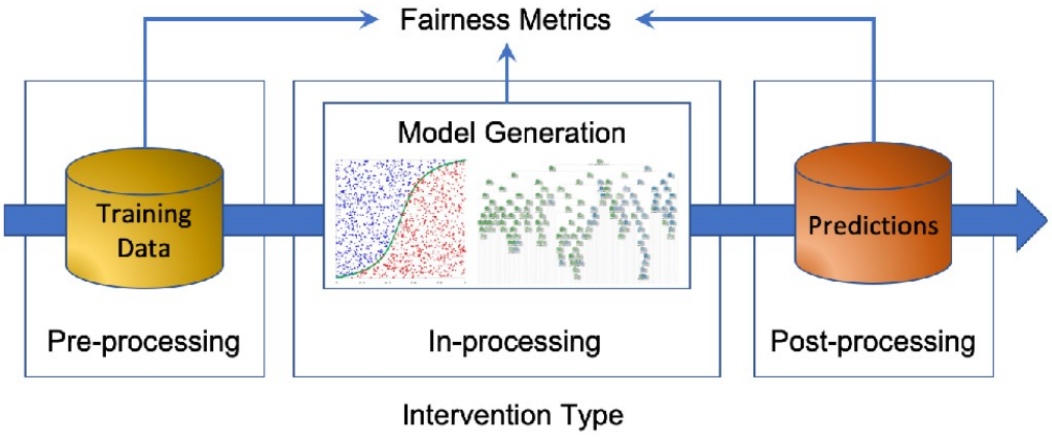
Chapter 6

Bias Mitigation

This section details the bias mitigation intervention algorithms used in step 3 of the project. There are three possible points of intervention to mitigate bias, preprocessing, inprocessing, and postprocessing. Because of the structure of the project where we want to take a well performing model from step 1 and perform evaluation and mitigation to achieve fairness, we focus on preprocessing and post-processing. This would allow the techniques to be reapplied to the models without needing to access their internal functions.

Additionally, we want to achieve fairness by deprecating model performance as little as possible. Following prior art as to which methods result in the best bias-accuracy tradeoff, we test preprocessing mitigation interventions on the XGBoost model and postprocessing interventions on the MLP model.

Bias Mitigation Pipeline:



6.1 Pre-Processing Methods

6.1.1 Disparate Impact Remover

The first preprocessing intervention taken is applying a Disparate Impact Remover to our data. A DI Remover edits values to increase fairness between specified privileged and unprivileged groups with an aim of removing the model's ability to distinguish between group membership. The main specification of the algorithm is the `repair_level` which indicates the distribution overlap between the two groups. Figure 6.1 shows a DI Remover applied with a repair level of 1.0, which provides full overlap. A key assumption of the DI Remover is in-group ranking which assumes that within a group individuals have similar experiences, therefore their in-group ranking should be preserved before and after repair.

Because be do not see a decline in performance as we increase the repair level, we set the repair level for each protected attribute's DI Remover to 1.0.

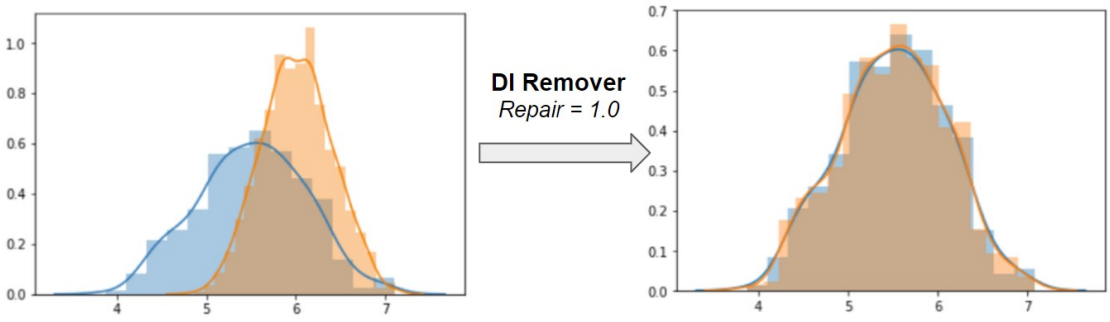


Figure 6.1: Disparate Impact Remover example with a repair level of 1.0

Figure 6.2 displays the effect of the DI Remover on each protected attribute. Interestingly, we see an improvement of balanced accuracy for all three protected attributes. As discussed earlier, sex and age are both risk factors for CVDs and therefore we should not aim to diagnose subgroups at the same rate. Contrarily, we should prioritize disparate impact by race because race should not be a driver in our decision-making process. The plots reflect that this perspective is reasonable, as we observe the reduction of DI across thresholds for race, reaching the acceptable range for fairness. Plots for sex and age only slightly reduce and slightly increase DI respectively, therefore showing that the masking of sex and age from decision making does not prevent these subgroups from having different outcomes.

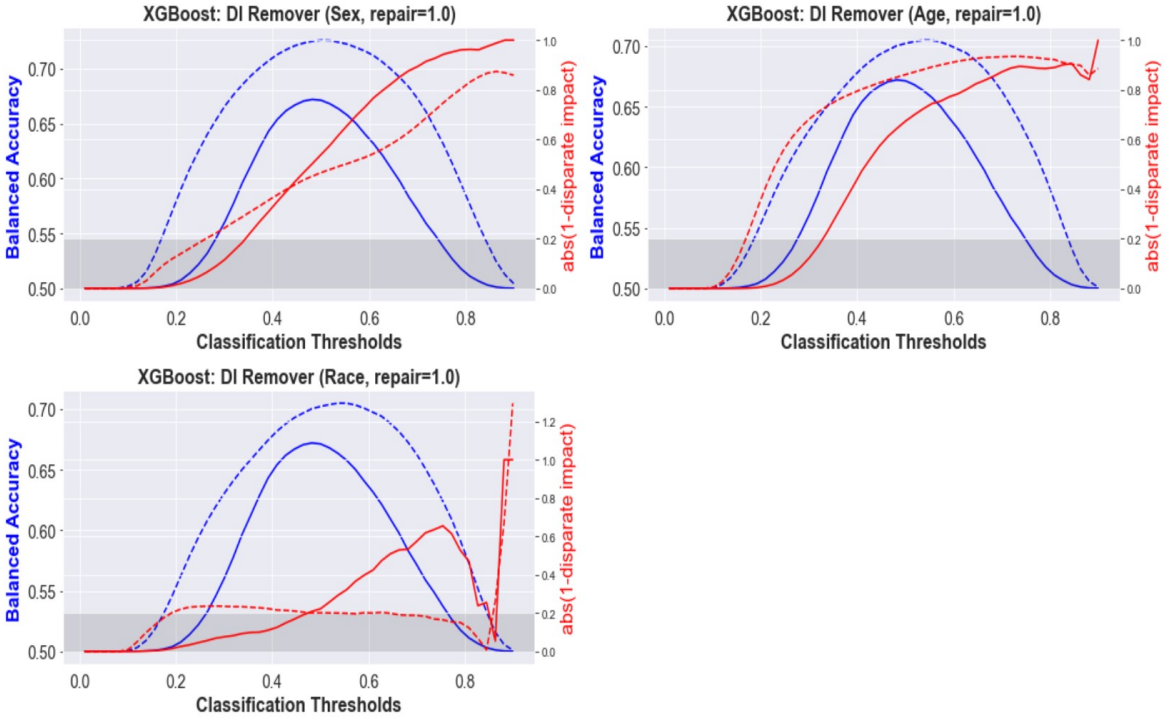


Figure 6.2: XGBoost balanced accuracy and disparate impact results before and after Disparate Impact Remover (DIR) transformation. Solid lines indicate performance before DIR and dashed lines indicate performance after DIR transformation. Disparate impact is represented as abs(1 - DI) to represent net bias, where a negative value favors the unprivileged subgroup. An abs(1-DI) value is considered fair between zero and 0.2

6.1.2 Reweighing

The next preprocessing technique implemented is reweighing. Reweighing takes training data and weights the examples in each group-label combination differently to ensure fairness before classification. This means that unfavorable groups with a favorable outcome are given larger weights to motivate the model to adjust its decision boundary away from preexisting bias, . We can define weights W in the following formula, where DI=1:

$$W(Y = 1, X_{sex} = 0) = \frac{P_{exp}(Y = 1, X_{sex} = 0)}{P_{obs}(Y = 1, X_{sex} = 0}$$

Results from reweighing are seen in Figure 6.3. We observe increased balanced accuracy for all protected attributes

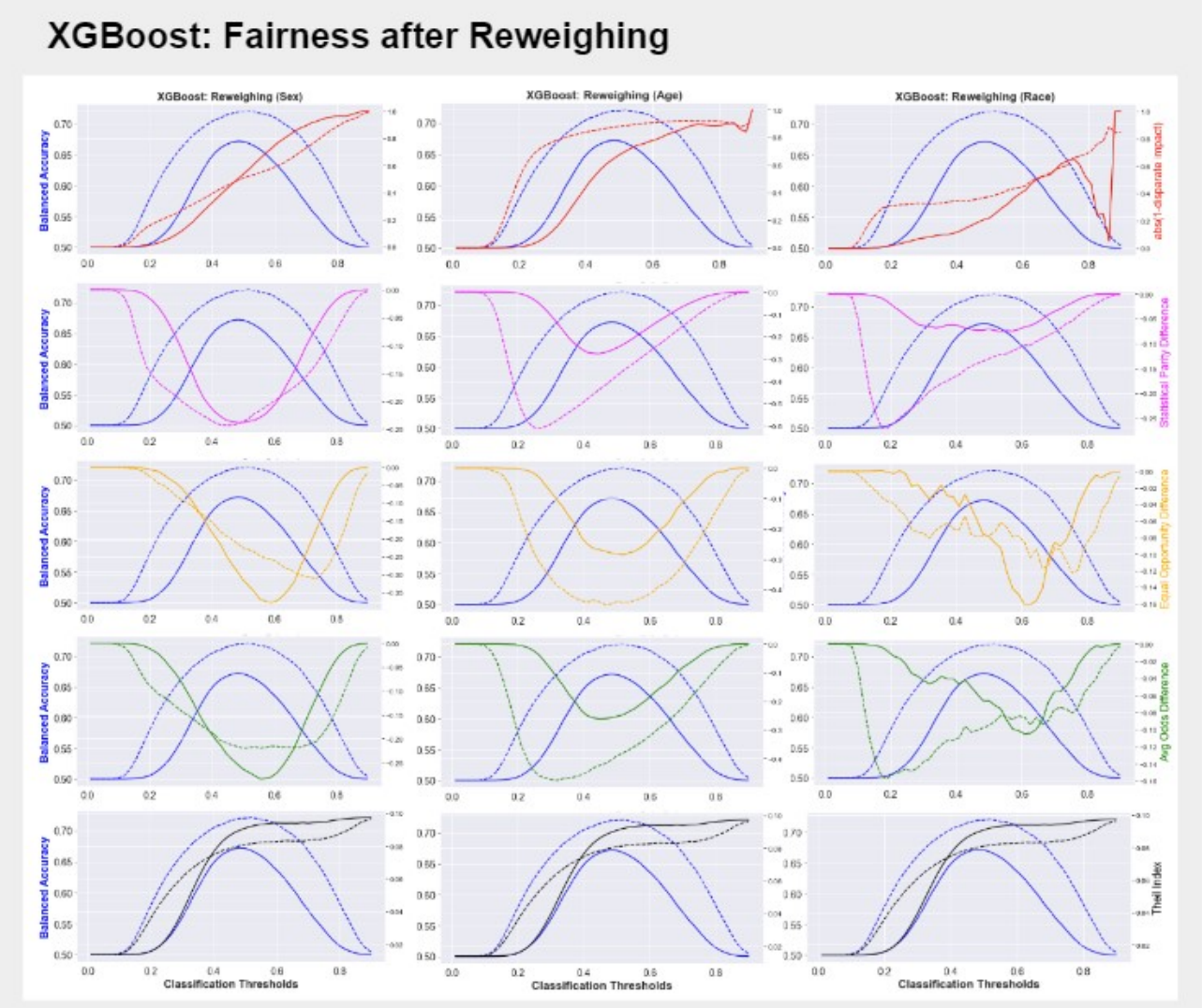


Figure 6.3: XGBoost fairness metrics and balanced accuracy vs classification decision thresholds before and after implementation of AIF360 Reweighing preprocessing

Aside from improving balanced accuracy, we do not see great results for reweighing when evaluating our fairness metrics. There is a slight improvement of the Theil Index, and fairness metrics for Race are in general slightly improved within the already acceptable range. For sex, DI, EOP, and AOD are all slightly improved but not within acceptable limits. For age we do not see improvements in fairness.

6.2 Post-Processing Methods

6.2.1 Calibrated Equalized Odds

Calibrated equalized odds uses an equalized odds objective to optimize over calibrated classifier score outputs to find probabilities to alter output labels. This post-processing technique is of special interest to this project because the EOP objective is the most relevant fairness metric to all protected attributes since we want to predict all subgroups at fair performance rates.

This intervention has a `cost_constraint` parameter of either the false positive rate (fpr), false negative rate (fnr), or a weighted combination of both. We use fnr as our constraint because we are interested in optimizing for sensitivity.

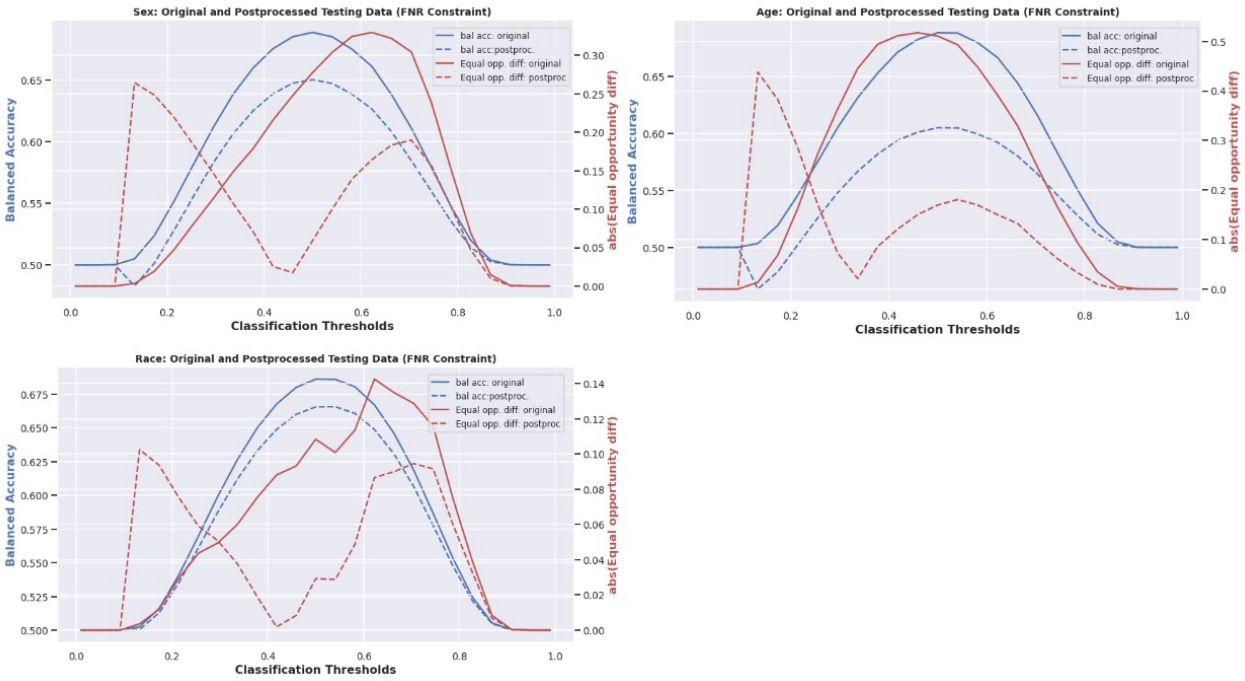


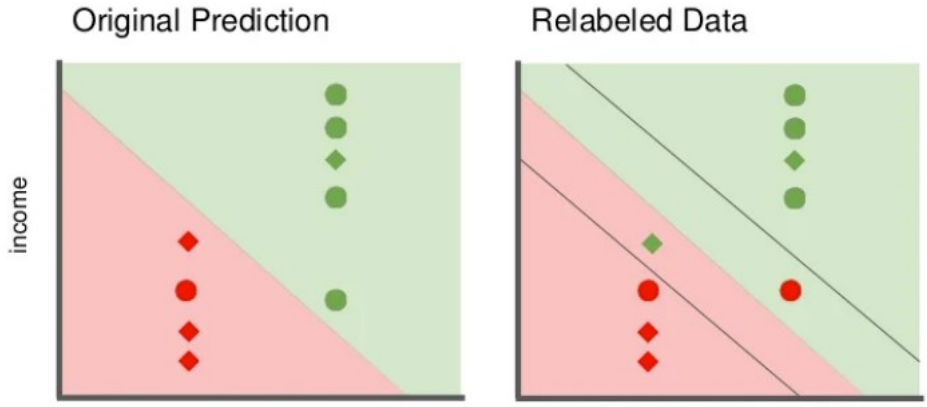
Figure 6.4: MLP balanced accuracy and equal opportunity results before and after Calibrated Equalized Odds postprocessing transformation. Solid lines indicate performance before transformation and dashed lines indicate performance after transformation.

Figure 6.4 shows the impact of calibrated equalized odds postprocessing on the absolute equal opportunity difference. For all protected attributes we see a slight degradation of balanced accuracy, most significantly for age. Both race and age are brought within the acceptable range for fairness and sex remains 0.05 above

acceptable at a classification threshold of 0.6.

6.2.2 Reject Option Classification

Lastly, we implement Reject Option Classification (ROC) whose basic principle is to implement positive discrimination. ROC builds on the theory that discrimination happens most often when there is not a clear distinction between candidates. ROC correct bias by giving favorable outcomes to unpriviliged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty .



When building a ROC object with AIF360 there are we first specify the metric used for optimization (statistical parity difference, equal opportunity difference, or average odds difference). We test ROC optimizing for both statistical parity difference and equal opportunity difference at a classification threshold of 0.5, defined earlier for MLP, and an ROC margin of θ=0.1

Defining the Rejection Margin θ:

$$\hat{Y} \leftarrow \hat{Y} + \theta \;\; if \;\; X_{protected\_attribute} = 0 \;\; and \;\; 0.5 - \theta \leq \hat{Y} \leq 0.5 + \theta$$

$$\hat{Y} \leftarrow \hat{Y} + \theta \;\; if \;\; X_{protected\_attribute} = 1 \;\; and \;\; 0.5 - \theta \leq \hat{Y} \leq 0.5 + \theta$$

Chapter 7

Results & Discussion

After determining a framework for bias evaluation and implementing a set of mitigation interventions, we observe final fairness results side-by-side in Figure 7.1-2 and final model performance results in Table 7.1.

7.1 The Bias-Accuracy Tradeoff

Table 7.1: Performance after Bias Mitigation

Model	Intervention	Attribute	Balanced Accuracy	Sensitivity	Specificity
XGBoost	original	-	0.63	0.88	0.38
	DI Remover	sex	0.71	0.82	0.61
	DI Remover	race	0.70	0.77	0.62
	DI Remover	age	0.70	0.77	0.62
	Reweighing	-	0.82	0.82	0.60
MLP	original	-	0.68	0.79	0.56
	Cal. odds	sex	0.65	0.83	0.45
	Cal. odds	race	0.66	0.83	0.49
	Cal. odds	age	0.59	0.91	0.27
	ROC - SPD	sex	0.68	0.67	0.69
	ROC - SPD	race	0.69	0.70	0.69
	ROC - SPD	age	0.69	0.67	0.67
	ROC - EO	sex	0.68	0.70	0.67
	ROC - EO	race	0.69	0.69	0.70
	ROC - EO	age	0.67	0.73	0.60

7.2 Discussion

While we generally expect to observe a fairness-utility tradeoff, the most stark result here is that we see significant model performance improvement from both pre-processing techniques. However, pre-processing in terms of achieving fairness has a mixed result. Both sex and race see across-the-board improvements for each metric. In contrast, age fairness is worse for both DI Removal and reweighing for all metrics. My hypothesis for why we see both performance improvement and, in the case of race, fairness decline has to do with the underlying impact that each protected attribute has on CVD diagnosis in the real world. Both of these methods attempt to balance DI, which is fundamentally unreasonable for age- a leading risk factor in CVD diagnosis. By obfuscating pertinent information related to CVD diagnosis prior to training, where age is the second-leading feature for our model's feature importance, and we are weakening the model's decision making capacity for both subgroups.

Results from postprocessing on the MLP model show quite different results. We achieve or nearly achieve fairness for each record using any of the methods available (calibrated equalized odds and reject option classification constrained by SPD or EOD). For calibrated odds we observe a 4-12% improvement in sensitivity while compromising balanced accuracy by 2-9%. Balanced accuracy for both ROC methods remains withing a 1% range of the original model and sensitivity decreases between 7-12%. Notably, ROC results also show significant improvement for specificity by 11-14%.

The contrast between between the success of preprocessing versus postprocessing methods for improving model fairness across all metrics may be explained by the methods of each type of intervention. The preprocessing interventions, DI Remover and Reweighing, both try to manipulate outcomes for subgroups of a protected attribute to balance favorable and unfavorable outcomes between groups. This is effective for a protected attribute that should not directly influence the outcome such as race, but ineffective when subgroups of the protected attribute should not be diagnosed equivalently in real life, in the case of age. The sex attribute sees results that are between these two outcomes because it itself falls between these two characteristics. Contrarily, postprocessing interventions are effective because they do not make the assumption that subgroups should be labeled equivalently. Calibrated Equalized Odds has an EOP objective, which prioritizes achieving the same classification rates between subgroups while ROC works within a confidence band around uncertainty, a much more specific approach to model correction.

Figure 7.1: XGBoost Results (left) for each intervention split by protected attribute– Disparate Impact Remover and Reweighing. All metrics excluding the Theil Index are considered fair from -0.1 to 0.1, indicated by the light grey region. MLP Postprocessing Results (right) for each intervention split by protected attribute– Calibrated Equalized Odds and Reject Option Classification constrained by statistical parity difference and equal opportunity. All metrics excluding the Theil Index are considered fair from -0.1 to 0.1, indicated by the light grey region.

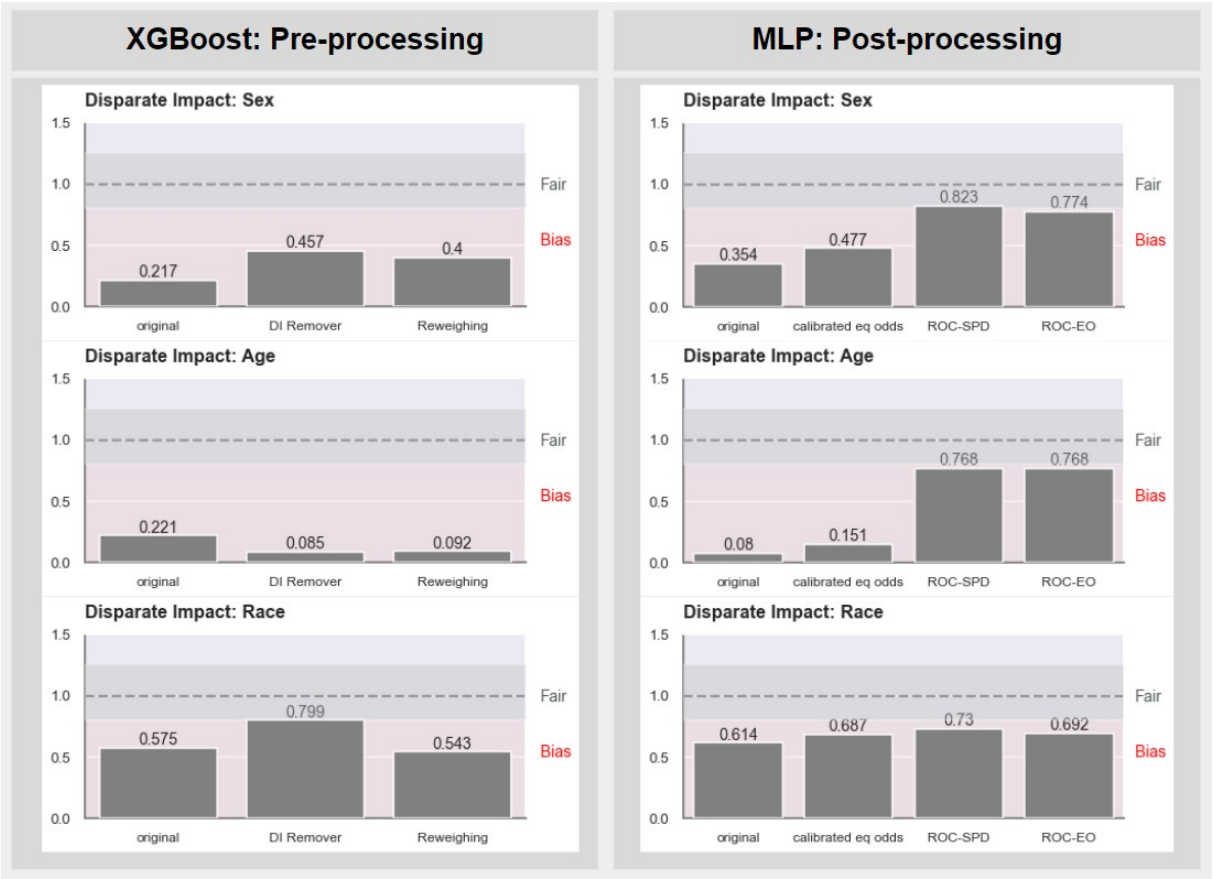


Figure 7.2: MLP Post-processing and XGBoost Mitigation Results– Disparate Impact

Chapter 8

Conclusion & Future Work

In this work we investigated how to approach fairness in artificial intelligence and methods to correct for unfair biases- all through the lense of developing an effective approach to automated cardiovascular disease diagnosis. Taking a tabular dataset from UKBioBank containing physical, environmental, and sociodemographic risk factors for CVD we developed an XGBoost and MLP models to evaluate for fairness then apply bias mitigation interventions for three protected attributes: sex, age, and race.

Both sides of our results show exciting opportunity for future development. This project limited itself to an inferred viable subset of intervention methods, but there exist a number of other methods both within the IBM AIF360 framework and beyond. Directly shooting off from this work there is potential to investigate fairness on the TabNet model and also to implement more complex mitigation interventions. Many studies show high efficacy of in-processing methods so that would be an excellent place to start. Alternatively, one could investigate the effects of combining intervention methods to optimize results- for instance, preprocessing methods in the project were able to improve performance while post-processing methods served best at bias mitigation, perhaps the combination could produce both positive qualities.

There also exists much room for model improvement. Sourcing more data to better capture the nuances of the problem space could allow for a more effective revisiting of the multi-label CVD problem. Both the age and race protected attributes were constrained by data composition, where age was only represented in a small range and race was severely misbalanced, therefore sourcing data that is fuller and that better represents our subgroups would be beneficial.

Finally, while this project focused on investigating fairness, further development into a framework of trustworthiness may be conducted to provide greater confidence in the entire AI pipeline. For a model to be trustworthy three pillars are commonly agreed upon- fairness, explainability, and robustness. Continuing this project into the larger story of trustworthiness is surely worthwhile and has the potential to make great strides for societal trust and perspective of artifical intelligence algorithms as a whole.

References

[1] A. N. Carey and X. Wu, "The fairness field guide: Perspectives from social and formal sciences," *CoRR*, vol. abs/2201.05216, 2022. arXiv: 2201.05216. [Online]. Available: https://arxiv.org/abs/2201.05216.

[2] G. A. Roth, G. A. Mensah, and C. O., "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study," *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982–3021, 2020, ISSN: 0735-1097. DOI: https://doi.org/10.1016/j.jacc.2020.11.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0735109720377755.

[3] K. Smolina, F. L. Wright, M. Rayner, and M. J. Goldacre, "Determinants of the decline in mortality from acute myocardial infarction in england between 2002 and 2010: Linked national database study," *BMJ*, vol. 344, 2012, ISSN: 0959-8138. DOI: 10.1136/bmj.d8059. eprint: https://www.bmj.com/content/344/bmj.d8059.full.pdf. [Online]. Available: https://www.bmj.com/content/344/bmj.d8059.

[4] D. S. Lee, M. Chiu, D. G. Manuel, *et al.*, "Trends in risk factors for cardiovascular disease in canada: Temporal, socio-demographic and geographic factors," *CMAJ*, vol. 181, no. 3-4, Ed., E55–E66, 2009, ISSN: 0820-3946. DOI: 10.1503/cmaj.081629. eprint: https://www.cmaj.ca/content/181/3-4/E55.full.pdf. [Online]. Available: https://www.cmaj.ca/content/181/3-4/E55.

[5] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *CoRR*, vol. abs/2106.03253, 2021. arXiv: 2106.03253. [Online]. Available: https://arxiv.org/abs/2106.03253.

[6] L. Mosca, E. Barrett-Connor, and N. Wenger, "Sex/gender differences in cardiovascular disease prevention what a difference a decade makes," *Circulation*, vol. 124, pp. 2145–54, Nov. 2011. DOI: 10.1161/CIRCULATIONAHA.110.968792.

[7] R. K. E. Bellamy, K. Dey, M. Hind, *et al.*, *Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*, 2018. DOI: 10.48550/ARXIV.1810.01943. [Online]. Available: https://arxiv.org/abs/1810.01943.

[8] E. Puyol Anton, B. Ruijsink, S. Piechnik, *et al.*, "Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation," Jun. 2021.

[9] M. Du, F. Yang, N. Zou, and X. Hu, "Fairness in deep learning: A computational perspective," 2019.