Reducing Bias between Demographic Groups in a Binary Classification Model for Predicting 10 Year Coronary Heart Disease

Charlie Galvin
Dep. Computer Science
Durham University
Durham, Durham
lbsr71@durham.ac.uk

Abstract: This paper aims to adopt a fair algorithm on a biased dataset in order to reduce it. The introduction covers the reasons for the project, then we go through data processing, reducing the bias through a fair algorithm and analysing the results.

INTRODUCTION

This paper aims to assess the 10 year risk of getting Coronary Heart Disease (CHD) within patients. This is not only a personal issue, but it is a disease that affects many adults throughout Europe and America. According to the National Heart, Lung and Blood Institute, it is the "leading cause of death in the US" [1] and people that have cardiovascular problems are noticed as more common patients of COVID 19. They also have a greater risk of mortality from the virus and it continues to cause "about one-third of all deaths in people older than 35 years". [2] Although this disease is incurable, it is treatable to the extent once diagnosed, through "lowering your risk factors and losing your fears, you can live a full life"[3]. This means that through using predictive algorithms, we can accurately inform people with the potential of having it within 10 years to change their lifestyle and lower their risk factor. With reduced discrimination and bias, thousands of lives can be dramatically improved if not saved every year from the model, but with discrimination this could seriously damage the lives of those in minority groups. As a personal motivation, I have always loved biology and had wanted to be an army doctor when I was younger, due to my religion and interest in science. I am fascinated by how important computer science is within medicine for reasons like predictive algorithms and when I learned I had the freedom to do a project in AI using a human-centric dataset , I knew I wanted to do something in medicine.

PROJECT MOTIVATIONS

I believe "assessing the 10 year risk of a patient having Coronary Heart Disease is very relevant to bias in AI due to:

- How human centric it is: All the features that are given are related to the human in the dataset and thus it is a human-centric issue.

- When looking at medical data in general, in many datasets there are underrepresented groups especially in gender or age. This type of misrepresentation can lead to some bias which again we have learned how to deal with.

- Through further reading, I have learned how sensitive predictive models can be in medical diagnosis. An example can be seen when the University of Pittsburgh designed an algorithm that suggested patients with pneumonia and asthma hade better outcomes [6]. If this was actually implemented it is likely it would have caused many deaths showing how careful we need to be when using it for diagnosis.

I also think that having a medical dataset, especially one in CHD, could lead to some potentially difficult tasks such as working with an imbalanced target variable. This is something I do not yet know how to solve but is especially prominent in hospital records where having the disease is rare. This means that although we will get very high accuracy, we may be predicting patients who do have the disease as negative. Of course we want to tell those that have CHD that they do, meaning we want as few false negatives as possible, going against what the algorithm would want as it wants to achieve high accuracy. This negotiation is a task I am interested in solving. The tasks I want to focus on are:

a) resolving the issue above
b) suspecting discrimination of underrepresented groups and justifying my thoughts through graphs.
c) resolving this bias to decrease discrimination at the minimal cost to accuracy.
d) Doing all this fairly, without trying to change the distributions of the dataset too much with any sampling or reweighting that I may do.

I plan on using Logistic Regression in this Binary classification problem. After reading a "Comparison of machine learning algorithms for clinical ev ent prediction (risk of coronary heart disease)" [7], I learned it is comparable to much more complex algorithms. The actual algorithm is not where I want to focus my idea of debiasing and so I just want to choose one that can generate comparable results.

The software tools that I am planning on using are:

(a) Python, version 2.7.3, for data preparation and algorithmic training and evaluation. I will use the sklearn library for my logistic regression model, sklearn.LogisticRegression as well as to split my data into training and testing will be sklearn.train_test_split. I may use GridsearchCV for hyperparameter tuning. I will use matplotlib.pyplot for data visualisation such as bar charts and heat maps as well as pandas for working with dataframes.

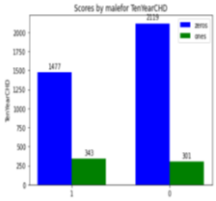
(b) Deepnote to display my extended report diary as well as the graphs that I show in python with matplotlib.

DATA ANALYSIS

I am planning on using the "Framingham Study" dataset [8] for this project. Within my comparison paper [7] , 'Table 1'

within this shows the missing table data. This dataset is quite healthy in the % of missing results in each column, but handling the data is still an important choice and can result in bias no matter what we do. As the paper states "there is no perfect way to compensate for missing values" and after looking at the "5 ways to handle missing values data" [9], I have decided to replace missing values with the mean. Not only does this prevent getting rid of missing values, as we already have a small dataset so losing records would be a shame, but I don't think this will affect the distribution of results too much and affects the distributions less than using the mode, which the comparison paper has done. I sacrificed repeatability in doing this for a better chance of maintaining distributions. Furthermore, I did some binning on the "age" feature, using pd.cut to split the age groups into 8 different age groups. This made it clearer to see which age groups were underrepresented and easier to find bias within the dataset.

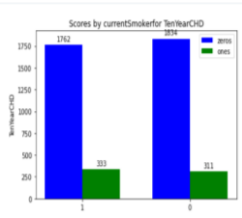
Within the demographic subgroups, I have noticed that a higher proportion of men are predicted to have 10 Year CHD compared to women. The bias that I have seen can be displayed using either Statistical Parity Difference or Disparate Impact.



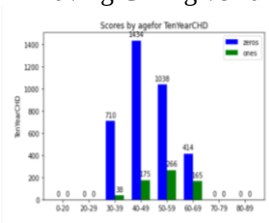
Disparate impact analysis using the four-fifths rule:

Group	Number in group	Positive TenYearCHD	Positive Rate	Disparate Impact
Male	1820	343	18.8 %	66.0 %
Female	2420	301	12.4 %	YES

The disparate impact is 66 %. Since this is less than 80 %, showing there is evidence for disparate impact.



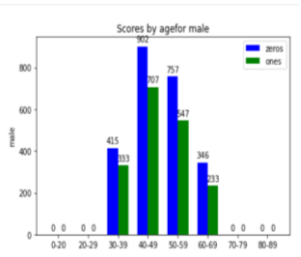
Although the demographic parity check seems to be met with this data, it is well known that smoking does increase the chances of having CHD. Indirect bias can be introduced through other features, such as the fact more males on average have more cigs per day than compared to women as well as there are more male current smokers than female. This could be due to social context at the time as most patients are 30+ in age, so at the time more men did smoke than women. This could explain why more men have CHD. When we look at age, the probability of having CHD given that you are in a certain age group is:

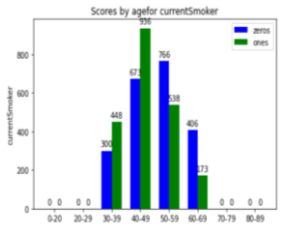


Age group	Chance of positive outcome (%)
30-39	5.08%
40 - 49	10.87%
50 - 59	20.40%
60 - 69	28.50 %

This shows the dataset underrepresents younger people and so the algorithm may show bias in predicting older people as positive with CHD compared to youth.

Looking at the results, it seems that smoking did not have the biggest effect with age as the most prevalent age groups for smoking were the younger groups such as 30-39 and 40-49. The relation between smoking and age could be that for younger patients there are more women then men, whilst the reverse for older patients. If more men smoke than women then this could lead to more indirect bias as this could increase the chances of them having CHD. Although this is an unjustified statement, I think it is a reasonable hypothesis.





As well as the combined features above, I believe that the bias occurs for the obvious reason that if more males have a positive outcome, and the system will predict positive outcome 100% of the time due to the skewed target variable, then the system will achieve higher accuracy and appear more bias to males just from this fact alone. Although I am not sure if this is considered bias as it treats both genders the same, it's just more males have more positive outcomes, I believe because more males have more positive outcomes due to other bias's such as representational bias, this is enough to notice that there could be bias in the dataset. However, we will redistribute the target variable to have an equal number of positive and negative cases and see if this removes the bias or whether it still exists.

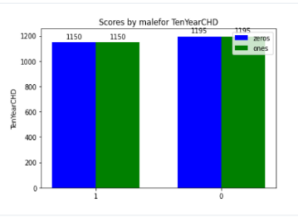
IMPLEMENTATION

At the beginning of the project I knew that we would have to do some debiasing in order to balance our target variable data, which had 85% of data as positive outcomes (getting CHD in next 10 years). When running our algorithm, we obtain our results through Logistic Regression. This is a type of classification algorithm which is most effective in binary classification, making it a good choice to use. It is used to predict the probability of a categoric dependent variable, and so predicts P(outcome = 1) as a function of X, which contains all of the relevant feature information. It uses a sigmoid function to describe the relationship between variables and will return a probability of a patient having "10YearCHD" based on their features, as a probabilistic outcome between 0 and 1. We then use a threshold, which is set at 0.5, to classify the value into a 0 or 1. I decided to use this model as it was fairly easy to implement and achieves comparable results to more complex algorithms. It is well-known for binary classification and is much better than linear regression for example, as well as it is faster than SVM for the size of the dataset we have.

My approach for training the model and seeing how it generalizes was first to tune all the hyperparameters for the data, using existing code I found online [10]. I knew GridSearch was a powerful hyper tuning tool and that I would use it within my project. Although I have not used the optimum parameters, I decided to use "newton-cg" as my solver parameter as it consistently achieved high accuracy in all results. The results predicted a negative outcome for 98.4% of testing data, achieving an accuracy ~85%, which although may appear impressive and is higher than the comparison paper's, it means 15% of our results are false negatives. One aim is to reduce the number of false negatives as giving a patient hope that they will not get CHD and then they do (a false negative), is far worse than a false positive. Our motivation is thus to achieve a high sensitivity at the cost of low specificity. I would say no obvious overfitting or underfitting has occurred.
Through testing with a more balanced dataset, we see more true negatives (closer to a 50:50 split). We have more true negatives and less false positives, due to the fair dataset, as well as the number of false negatives has increased which may show how the system is biased towards predicting positives each time due to how it has been trained. The desire to predict positives for data has shown by the increase in this value. However, I don't think this bias says much about demographic groups but rather the distribution of the target variable, so I have rebalanced the data using resampling and have done this experiment again.

After running the algorithm on a newly balanced dataset, the number of false negatives is still quite high at 15.7%. However I would say that is has generalised similarly to before. Although the model does perform slightly better, with a lower percentage of both false positives and false negatives, the benefit is not as drastic in comparison to last time which could mean that balancing the outcome has maybe lessened the extent of bias or increased the overall distribution of demographic groups within the dataset.

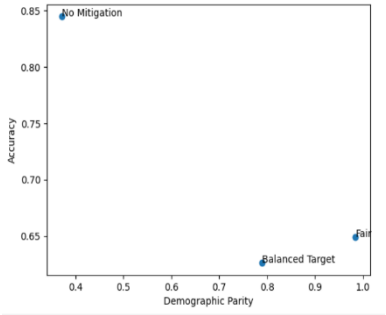
REDUCING BIAS WITHIN THE DATASET
I have decided that I am going to reweight my dataset though sampling data as a way of mitigating bias in pre-processing. The earlier bias is reduced, the less likely it will occur in the overall workflow so that is why I chose to do it in pre-processing. My algorithm reweighs the data for both male and female, in order to get the expected outcome for having bias as 0.5 in each, meeting the demographic parity condition. I decided not to under-sample as I thought I would lose valuable information that may aid in increasing accuracy for my predicting model. Instead I oversampled data for both the outcome of males who did not have CHD and females who did have CHD. The only concern with this is overfitting although I do not think this has occurred.



After running both the majority group as my test set and minority group in my test set, I have seen that the accuracy between them is similar, with the minority having lower precision and higher recall compared to the majority. One of the easiest ways, although not always true, in seeing if debiasing has occurred is to see how accurately the algorithm predicts the majority and minority classes. If there is a major difference this can be seen as an indication of bias. When I ran the majority and minority subsets on the original data, the accuracy differed by 0.6. On the minority my accuracy was 0.647 and on the majority, 0.652. This differs by 0.006 which is much closer. The overall statistical parity is met completely, which is what the resampling algorithm was designed to do. When looking at Disparate Impact:

Group	Number in group	Positive TenYearCHD	Positive Rate	Disparate Impact
Male	2025	1024	50.56 %	100.9 %
Female	2315	1182	51.06 %	NO

This is incredibly close to one which is a good indication that the bias has been reduced. Furthermore when looking at "Average Odds Difference" (AOD), which is the average difference between False Positive Rates and False Negative Rates, this should be equal to 0 for a fair result. Our results suggest an AOD equal to 0.030645. I think all of these indicators combined suggest that the fair algorithm I have developed has been a success. Furthermore the ROC curves are very similar for both which indicates that it obeys equalised odds. In comparison with my research paper, which achieved 84% accuracy, it had a specificity of 0.99 and a sensitivity of 0.05. Whilst I had along the way achieved an accuracy of 89%, I have achieved an accuracy of ~65%, but have a specificity and a sensitivity of 0.65 This balance indicates that my algorithm does not just predict a positive outcome each time, which means it may be more accepted in society and also more helpful in comparison with the one implemented in the research paper.



Lastly, this reduction in accuracy is always a trade-off with reducing bias. This graph shows that although we do have a decrease in accuracy, the demographic parity gets closer to 1 and thus the discrimination is massively reduced so I think it is worth it.

ACKNOWLEDGMENT

I would like to acknowledge Dr Ehsan Toreini for being an excellent lecturer.

REFERENCES

[1]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4958723/

[2] https://www.narayanahealth.org/blog/coronary-artery-disease-life-expectancy-and-prognosis/

[3] https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease

[4]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6489351/#:~:text=Background,in%20predicting%20coronary%20artery%20diseases.

[5]https://www.pulmonologyadvisor.com/home/topics/practice-management/the-potential-pitfalls-of-machine-learning-algorithms-in-medicine/

[6]https://www.pulmonologyadvisor.com/home/topics/practice-management/the-potential-pitfalls-of-machine-learning-algorithms-in-medicine/

[7]https://www.sciencedirect.com/science/article/pii/S1532046419301765

[8]https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset

[9] https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/

[10]https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/