Algorithmic fairness with causal models

Zaiga Thomann

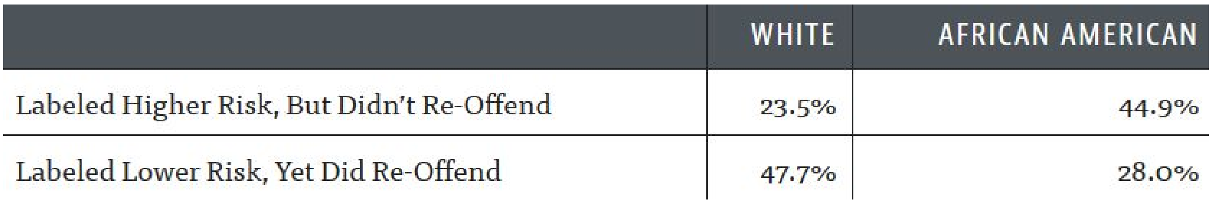
This report seeks to summarise the work I recently conducted as an intern at Silverpond. Broadly the project was interested in exploring ways in which decision-making algorithms could be made fair, whether through adjusting the initial training process or by tweaking an existing algorithm. I narrowed the scope of my project down to examine how causal modelling techniques can be used to adjust existing decision-making algorithms to exhibit fairer behaviour. I have focussed here on the broader motivation and theory but would be happy to share other resources and explanations as requested. I am hoping that if you have questions or are intrigued by this topic that you will access some of the resources mentioned to learn more about causal inference and modelling.

A huge thanks to SIlverpond for not only hosting me but making me feel so comfortable and welcome in their office and their intellectual domain. An even bigger thanks to my supervisor Lizzie Silver, Senior Data Scientist, Silverpond, for showing care and consideration when answering all my questions and for making me feel confident enough to undertake research in a completely new field.

Are decision making processes fair?

In 2016 Propublica, a not for profit organisation in the US that undertakes investigative journalism in the public's interest, released their analysis of an algorithm used to predict recidivism rates of defendants. The risk scores generated by the algorithm, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), are used by courts to help inform sentencing decisions.

The Propublica analysis showed that while the algorithm correctly assigned risk scores at the same rate for black and white defendants, the way that it erred differed according to race. The study highlighted that the algorithm was more likely to incorrectly label black defendants as being at a high risk of recidivism whilst white defendants were more likely to be incorrectly labelled as having a low risk (Angwin and Larson, 2016). This result can be seen in Figure 1.



	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Table *1: Prediction failure rates for defendants by race (Angwin and Larson, 2016)*

Interestingly the COMPAS survey that is used to calculate an individual's risk score does not explicitly collect or consider race as an attribute. However, it is widely accepted that other attributes and data collected are acting as a proxy for race instead. This case study also highlights issues around using historical data to inform current and future algorithms, as these algorithms have the potential to encode and enforce past bias into future decisions.

There are some very interesting discussions around the COMPAS algorithm and the best statistical measures and methods to use in assessing its fairness however this is not the focus of this project. Instead I highlight COMPAS as just one example of a decision-making algorithm that has the potential to have meaningful and significant effects on an individual's life and one in which some serious questions have been raised about its fairness.

The Propublica investigation highlighted the need for us as a society to interrogate and question the algorithms being used around us. Discussions of fairness and bias are occurring more frequently as more of these cases come to light (the Amazon recruitment algorithm is another example (Dastin, 2018)). What these cases have shown is that as a society we care about how decisions are being made and that we want reassurance that automatic decision-making processes in the future are making fair decisions on our behalf. But what does it mean to be fair? And how do we begin to decide what a fair decision should look like?

What is fairness?

There are two parts to this question. The first is what, at a human society level, we want fairness to mean. This in and of itself is a hotly contested philosophical discussion and is beyond the scope of my research project. What I found interesting however is that in order to translate notions of fairness to decision making algorithms we need to be able to define fairness in a way that can be implemented computationally.

Even in the field of computer science and machine learning there are myriad definitions of what constitutes a fair decision. Verma and Rubin (2018) and Mehrabi et al. (2019) do a great job at summarising a range of definitions and I won't go over them all. I will however outline the three approaches that are most pertinent to my research.

To help explain these it will be useful to have a hypothetical scenario to consider. As a loan provider you are developing an algorithm to help automate the process of approving credit applications. To do so you train it on data of individuals who have already applied, and the outcomes of their application processes. You are aware that in this historical data there may have been a bias related to gender and the granting of loans and you are concerned that this may affect the fairness of your algorithm. In this case gender is the attribute that you want to 'protect' from bias, as such we will call it our **protected attribute**. There are three approaches/definitions that you are considering:

Fairness through unawareness (FTU)

The approach taken by FTU is to address this bias by simply leaving gender out of the attributes that are provided to the machine learning algorithm. However, this doesn't necessarily prevent unfair decision-making. In fact, it is highly likely that other attributes may act as a proxy for the protected attribute, potentially further obfuscating the presence of bias.

Equal Opportunity (EO)

If you choose the equal opportunity approach you will address this potential bias by ensuring that the probability of a decision for an individual is the same regardless of their gender, conditional on the non-protected attributes. For example, two candidates with the same occupation, salary and debt, but different genders, should have the same probability of getting a loan. (Note that if gender is not passed to the decision algorithm this is equivalent to FTU.)

Counterfactual fairness (CF)

The aim of counterfactual fairness is to ensure that the same decision would have been made whether the candidate had been female or male. It does this by recognising that protected attributes can have a causal effect on other attributes, for example gender may have an effect on an individual's salary. So for counterfactual fairness to be achieved the probability of an individual receiving a certain decision must be the same regardless of the value their gender takes *and the corresponding effect this change may have on their other attributes.*

What kind of fairness do we care about and why?

This project began with the understanding that we wanted to experiment with ways of achieving counterfactual fairness. This was for a number of reasons. Firstly, the power and importance of causal techniques is becoming better understood and appreciated and an increasing number of academics and data scientists are exploring their possible uses. Secondly, my supervisor for this project has a background in causal modelling and so the project could leverage her expertise. Finally, and perhaps more importantly, counterfactual fairness provides a definition that allows and encourages us to not only recognise bias but also to account for the way in which societal bias can permeate throughout a multitude of factors in an individual's life.

What has been done already in the field of causal techniques for fair decision-making?

As mentioned above, causal techniques are gaining recognition and popularity in the statistics and data science sphere. Using these techniques to achieve fairer outcomes is also a growing area of research in the machine learning field and new approaches and interpretations are being shared on a regular basis. I would like to thank the FairML reading group (Wallace, 2020) for helping to keep me up to date with this research and a lot of these papers have come from their reading list. The two listed below are the main ones I consulted for this work.

Counterfactual Fairness (Kusner et al., 2017)
This paper is credited with being one of the first to offer a definition for the notion of counterfactual fairness. In the lead up it also does quite a nice job of summarising different notions of fairness. It also offers an algorithm for training a counterfactually fair algorithm. However, the algorithm can only be used when the protected attribute does not influence attributes that we need to take into account to make the decision. It can't be applied to the loan-granting case, for example, if gender influences salary, and salary information is required to make loan decisions.

Equal Opportunity and Affirmative Action via Counterfactual Predictions (Wang et al., 2019)
This paper builds on the original *Counterfactual Fairness* paper by providing an approach to adjust an existing decision-making algorithm to make it counterfactually fair. It does this by 'smoothing' the decisions made by the original algorithm over the probability distributions for the protected attribute and the space of all possible counterfactual realities for a particular individual. This approach is unique as its emphasis is on 'upcycling' an existing algorithm which means that the original algorithm can already exist or be trained by non-causal approaches.

What did we do?

For this project we always wanted to use causal modelling techniques (I'll go into what these are a bit later) and the counterfactual definition of fairness as the basis of any approach. Initially we identified three areas of potential interest:

1. Exploring and demonstrating methods to choose useful and accurate causal models for a dataset
2. Demonstrating methods to train a decision-making algorithm to ensure the counterfactual fairness of its decisions

3. Demonstrate methods to adjust an existing algorithm so that it makes fairer decisions according to the counterfactual fairness definition

To focus the scope of the project we decided to focus on the third area of interest, and in particular adjusting an algorithm to make it fair according to the methods outlined by Wang et al. (2019).

How did we do it?

Here I will provide a brief overview and some background information necessary to understand the approach.

Data

Real Data

The first step was to find a good dataset. In particular the dataset needed:

- Individual level data
- Measurement of at least one protected attribute (e.g. gender or race)
- A decision made about those individuals or the ability to reasonably construct a decision
- Numerical attributes or the ability to convert the data to numeric data
- Continuous or ordinal attributes (this condition came a little later on in the process to aid computation)

The UCI Adult data set (Dua and Graff, 2017), with a few adjustments, generally fit these requirements well. It is also a dataset that is commonly used in the fairness space and that was a bonus as it meant that readers (you!) may have come across it before.

The data does not include income, only an indicator of whether the individual's income was above or below $50k/year. We wanted to use linear models, so we constructed a real-valued income column using the indicator column as a constraint, and using other related features (hours, occupation, gender, race) as predictors.

Synthetic Data

The next step was to create a synthetic data set that behaved similarly to the real data. The causalgraphicalmodels package (Barr, 2020) was used to create this dataset. The advantage of creating a synthetic dataset is that the causal structure is immediately known. This is incredibly important as the algorithm requires a causal model to be supplied. In the case of the synthetic dataset we knew the true nature of this model and as such we could more clearly evaluate the effect of the adjustment. In the case of real data however, it is harder to be sure of the true nature of the causal model and this potential error may affect the adjustment.

Decision Data

The one key attribute missing from the UCI Adult data set was a decision made about each of the individuals. As such I had to construct synthetic decisions for the data so that we could train our biased decision-making algorithm. We intentionally created biased "training data" in order to train a biased decision algorithm, so we could test how well the debiasing tool worked.

The scenario that best fit the available data was that of a bank deciding whether to grant or deny a loan. As such all that was required for each individual was a binary decision. To do this I normalised all the data and then took the weighted average of an individual's normalised attributes plus an error term. This factor was then used as the probability for a binomial distribution from which the decision was then sampled. Since gender contributed directly to the "decision", it is inevitable through this process that the decisions will be biased.

In the case of the real data, the dataset was split and only a portion was used as historical decision data, whilst the rest was retained to be able to assess the efficacy of the fairness algorithm. In the case of the synthetic data this was not important as more data could be generated.

A Decision Algorithm

In both cases, real and synthetic data, a simple classifier was trained on the synthetic biased decisions described above. This algorithm was intended to be an example of an algorithm that a bank may use to automate loan approvals. Since it was trained on biased decision data, the algorithm also produced decisions that were biased with respect to sex.

Causal Models

Once the data was generated it was time to implement the fairness algorithms. The technique described in Wang et al. (2019) relies on causal models for its analysis and so I thought it pertinent to share some basics.

Causal modelling is a statistical technique that models how different attributes have a causal effect on one another. An example model is shown below.

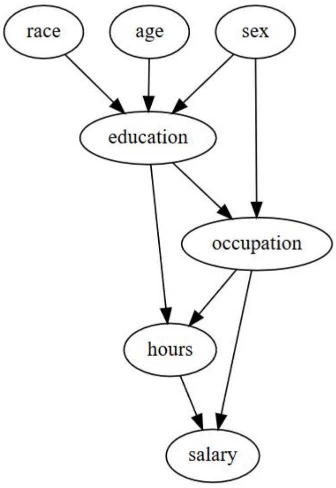


Figure 1: A possible causal model of a proposed structure for the UCI Adult Dataset.

Causal models allow us to recognise and track the effect that an attribute has on the rest of the attributes in a dataset. For example, this model states that age, race and sex all have a causal effect on the level of education that an individual may complete. The aim of a causal model such as this is to map the causal relationships that underpin the attributes of data collected, with the aim of being able to answer counterfactual questions. For example whether or not you would receive a loan had you been a male instead of a female. If you want to read more about causal modelling I recommend a number of Judea Pearl's books. *The Book of Why* (Pearl and Mackenzie, 2018) provides a nice introduction to the field of causality for readers with a broad range of mathematical experience, whilst *Causal inference in statistics: a primer* (Pearl et al., 2016) and *Causality: models, reasoning and inference* (Pearl, 2009) provide more details on the mathematical processes and theory involved.

Choosing a causal model

Synthetic Data
To begin with a causal model was chosen to help construct the synthetic dataset. Figure 1 shows this model. Common sense and intuition was used to choose the model. Since this model was used to construct the synthetic data we knew that it represented the true underlying causal structure of this data. Since initial tests were conducted on the synthetic data, it did not matter if this was the true causal structure of the real data.

Real Data
Initially when applying the procedure developed to the real data the causal model in Figure 1 was used. This returned odd results which was not surprising as a huge assumption was being made about the causal model being correct. To address this assumption, the py_causal package (University of Pittsburgh/Carnegie Mellon University Center for Causal Discovery, 2019) was used to learn a model structure from the data. It is important to note that the model, while more closely reflecting the underlying structure of the data, may still not be the true causal structure.

EO Fairness

The step of calculating EO fairness is quite straightforward. Essentially to make an existing algorithm EO fair you take a weighted average of the decisions that are made if the protected attribute of the individual is changed. The weights are the probability of that value for the attribute occurring in the population.

One assumption that was made in the implementation of this algorithm was that there was only one protected attribute being considered. In the case of multiple protected attributes the algorithm would need to be adjusted slightly to include all possible combinations of values of the protected attributes and if the protected attributes were not independent then the joint probability distribution would have to be calculated as well.

Calculating counterfactual values

The core concept behind counterfactual fairness is the notion of counterfactuals. This notion is best described with an example.

Bob, whose gender is male, is applying for a bank loan. His application is denied.

A counterfactual question might read: If Bob had of been a woman, would his application have been approved?

In this instance the counterfactual is the alternate reality where Bob was indeed a woman. Whilst this is impossible to test IRL, it is possible to use causal models to make predictions on how the loan outcome may have been different had Bob had a different gender.

For the purposes of calculating the counterfactual values it is important to ensure that the information that is known about the individual is used. For example, if Bob has a high salary as a male, it is reasonable to expect that as a female he will also earn a relatively high salary. Working out how to do this was one of the parts of this project that I found most difficult. In order to simplify the problem for the real world scenario I had to make a large assumption.

Assumptions

In the case of the synthetic data I constructed the data so that the functions describing the relationships between attributes in the data (the structural functions) were linear. This meant that the counterfactual values for an individual were unique and solvable as per the procedure described in Pearl et al., (2016, p.94,p.96,p.98) and Pearl (2009,p.37,p.206,p.209).

In order to apply this same process to the UCI Adult dataset, I had to make the assumption that the structural functions were linear. This is an important assumption to note as it is likely not true for all of the attributes and as such is a potential source of error.

CF Fairness

The principle behind counterfactual fairness is that the probability of an algorithm making a certain prediction or decision for an individual is the same as the prediction the algorithm would make if the individuals protected attributes were changed and the causal effects of those changes were also reflected in the rest of the individual's data. What is important, and different about this approach, is that the effect of changing the protected attribute is permeated throughout the rest of the data by calculating the counterfactual values for an individual.

Figure 3 below depicts an example of this process. Firstly a prediction is generated for an individual. Secondly the protected attribute of sex is changed from male to female and the effects of this are updated in the data. This is depicted by the red lines. Once the data has been updated a prediction is generated. If a predictor/algorithm is counterfactually fair then the probability of the predictions in these two cases will be equal.

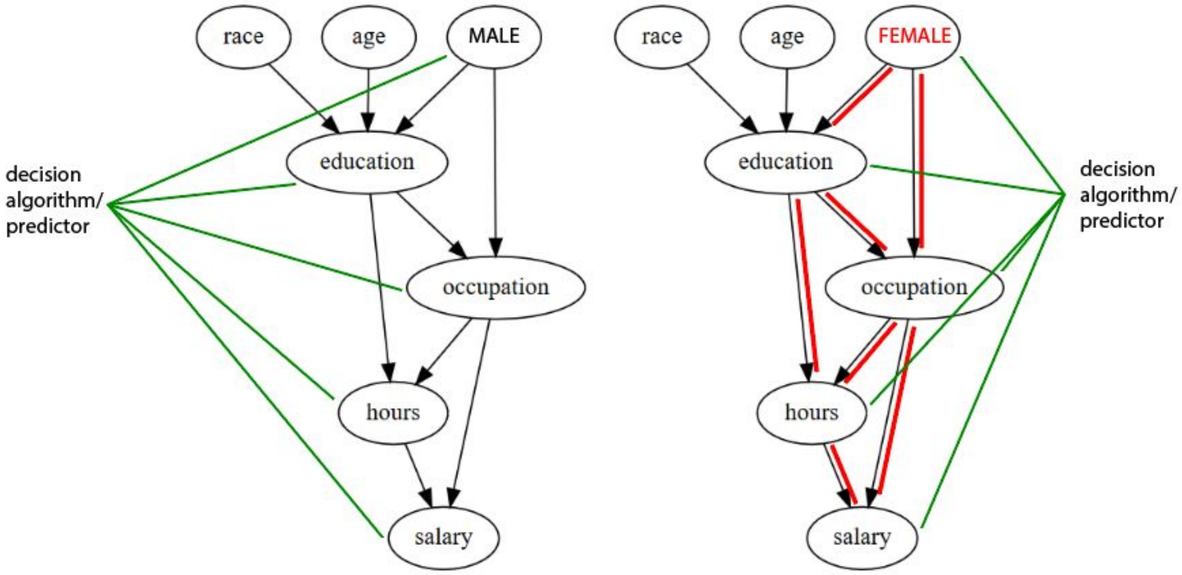
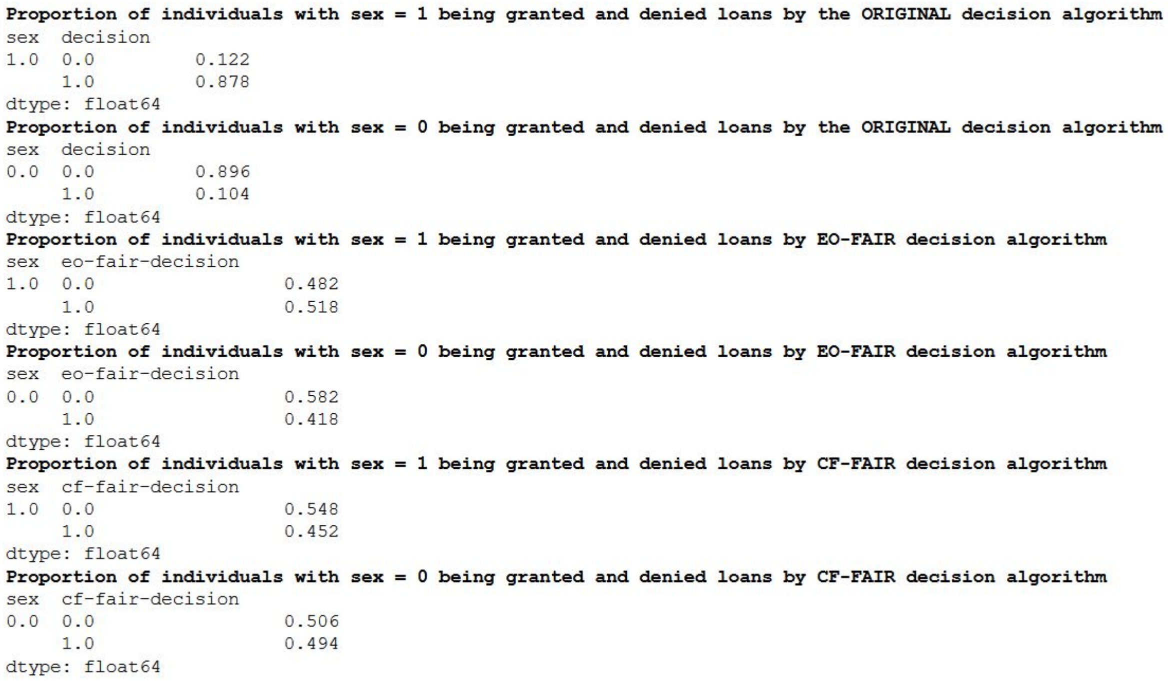


Figure 3: Making a decision using an individual's counterfactual values

Results

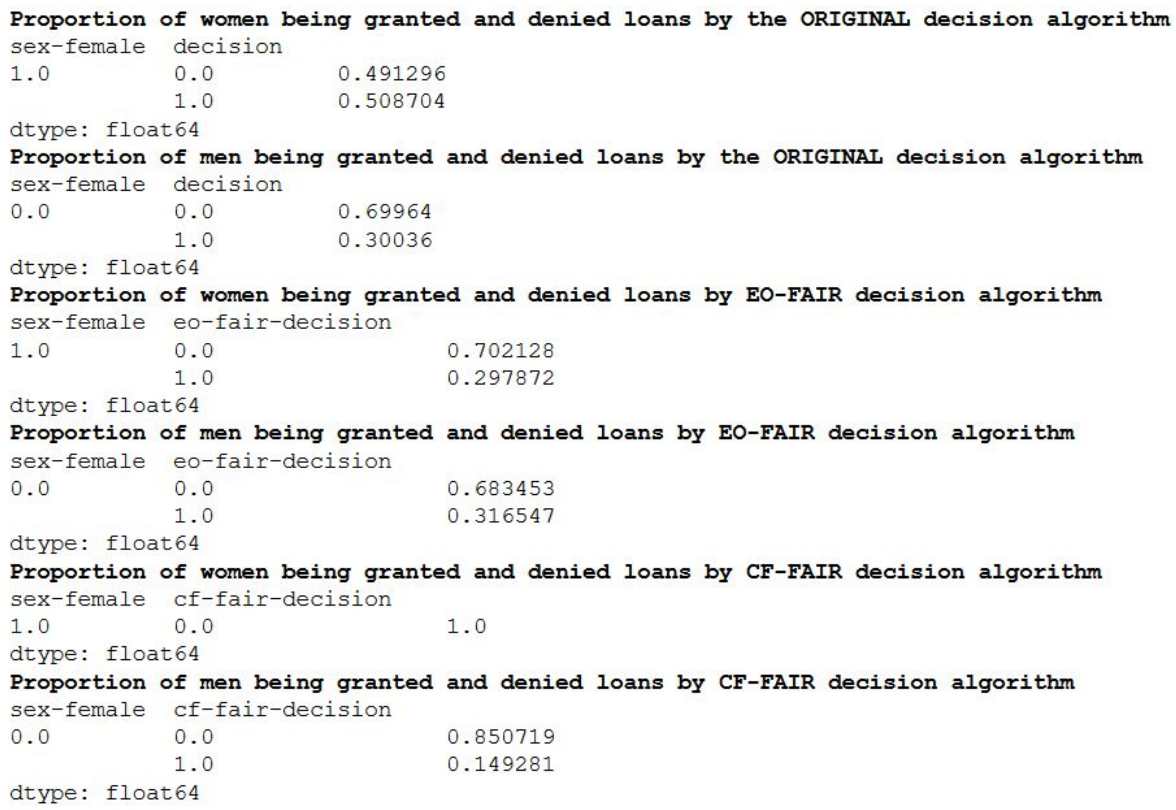
In the end we saw great results with the algorithm implemented when using the synthetic data. The proportions of loans granted and denied to males and females evened out when comparing the original decisions, to the decisions made by the equal opportunity and counterfactually fair algorithms. This can be seen in the Figure 4 as the proportions all tend towards 0.5 as the successive algorithms are applied.



```
Proportion of individuals with sex = 1 being granted and denied loans by the ORIGINAL decision algorithm
sex  decision
1.0  0.0         0.122
     1.0         0.878
dtype: float64
Proportion of individuals with sex = 0 being granted and denied loans by the ORIGINAL decision algorithm
sex  decision
0.0  0.0         0.896
     1.0         0.104
dtype: float64
Proportion of individuals with sex = 1 being granted and denied loans by EO-FAIR decision algorithm
sex  eo-fair-decision
1.0  0.0                 0.482
     1.0                 0.518
dtype: float64
Proportion of individuals with sex = 0 being granted and denied loans by EO-FAIR decision algorithm
sex  eo-fair-decision
0.0  0.0                 0.582
     1.0                 0.418
dtype: float64
Proportion of individuals with sex = 1 being granted and denied loans by CF-FAIR decision algorithm
sex  cf-fair-decision
1.0  0.0                 0.548
     1.0                 0.452
dtype: float64
Proportion of individuals with sex = 0 being granted and denied loans by CF-FAIR decision algorithm
sex  cf-fair-decision
0.0  0.0                 0.506
     1.0                 0.494
dtype: float64
```

Figure 4: Results of applying eo-fair and cf-fair algorithms to synthetic data

However, this behaviour was not observed in the case of the real dataset. In this case the proportions of individuals both receiving and being denied a grant became less even as the equal opportunity and counterfactually fair algorithms were applied. One thing to note in this instance is that the original decision algorithm seems less biased than in the previous case. However there is still a higher proportion of men being denied a loan and we were expecting to see this proportion decrease. Instead this proportion increased and the algorithm seemed to tend towards denying everyone a loan as can be seen in Figure 5. This may well be due to the assumptions of causal structure and linear structural functions that were made.



```
Proportion of women being granted and denied loans by the ORIGINAL decision algorithm
sex-female  decision
1.0         0.0         0.491296
            1.0         0.508704
dtype: float64
Proportion of men being granted and denied loans by the ORIGINAL decision algorithm
sex-female  decision
0.0         0.0         0.69964
            1.0         0.30036
dtype: float64
Proportion of women being granted and denied loans by EO-FAIR decision algorithm
sex-female  eo-fair-decision
1.0         0.0                 0.702128
            1.0                 0.297872
dtype: float64
Proportion of men being granted and denied loans by EO-FAIR decision algorithm
sex-female  eo-fair-decision
0.0         0.0                 0.683453
            1.0                 0.316547
dtype: float64
Proportion of women being granted and denied loans by CF-FAIR decision algorithm
sex-female  cf-fair-decision
1.0         0.0                 1.0
dtype: float64
Proportion of men being granted and denied loans by CF-FAIR decision algorithm
sex-female  cf-fair-decision
0.0         0.0                 0.850719
            1.0                 0.149281
dtype: float64
```

Figure 5: Results of applying eo-fair and cf-fair algorithms to semi-synthetic UCI Adult dataset

Next Steps

There are two main areas of improvement that I think would be important to implement in the near future.

The first is to implement the metric formulas also outlined in the paper (Wang et al., 2019). This would help with assessing the performance of various 'fair' algorithms as well as helping to identify errors within the code.

Secondly, and possibly more importantly, is the need to handle structural functions beyond linear relationships.

Finally, it would be interesting to do an analysis of how much error is introduced by incorrect assumptions regarding structural functions and causal models. This could potentially be undertaken on a series of synthetic datasets where the true structural functions and models are known but alternate ones are used. Combined with the metric formulas this may provide one way of measuring the error introduced by these assumptions.

If these avenues were explored then it would be easier to identify why the results from the real data test were seen.

Conclusion

A synthetic data set and a semi-synthetic dataset were created to model individuals who had applied and been denied or granted loans. This data was used to create a biased decision algorithm with the aim of applying causal methods to debias it. The debiasing method requires a causal model, a dataset and a biased algorithm.

In the case of the synthetic dataset, where the true causal model and linear nature of the structural functions was known, the debiasing method was observed to even out the proportion of males and females receiving a loan. However, in the case of the semi-synthetic data the debiasing method produced counter-intuitive results, with the new algorithms tending towards denying everyone loans. This indicates that the assumed causal model and structural functions were not correct and that further work needs to be conducted with different models.

The method undertaken in this project has wide reaching and important applications. If these techniques were to be applied in the cases of the COMPAS or Amazon recruitment algorithms the bias displayed by these algorithms may have been identified or even addressed earlier. However, it is important to note that the techniques outlined in this paper still require potential sources of bias to be identified and mapped to protected attributes. As such it is important that the issue of algorithmic fairness is not considered to be a purely technical one. I highly recommend this article as a fantastic demonstration of the complexities of understanding, identifying and addressing algorithmic bias.

What has been demonstrated by this project is the need for accurate causal models when applying these techniques. As such it is essential that individuals and organisations invest in developing a deeper understanding and expertise in the field of causation if they want to undertake efforts to ensure the fairness of their decision algorithms.

Glossary

Attributes – the types of data collected for an individual (for example sex, salary, age are all attributes)

Protected attribute – the attributes that we don't want an algorithm to exhibit bias against

Node – when drawn as a causal model each attribute becomes a node of the model

Paths – the line or causal connection that shows that two nodes have a causal relationship

Children – the children of a particular node are nodes that are causally affected by the node in question

Parent node – a node that has a causal effect on a particular node

Structural Functions – a mathematical equation that describes the relationship between attributes directly connected by a path in a causal model

Resources

Angwin, J., Larson, J., 2016. Machine Bias [WWW Document]. ProPublica. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 2.13.20).

Barr, I., 2020. ijmbarr/causalgraphicalmodels.

Dastin, J., 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.

Dua, D., Graff, C., 2017. UCI Machine Learning Repository: Adult Data Set. University of California, Irvine, School of Information and Computer Sciences, Irvine, CA.

Kusner, M.J., Loftus, J., Russell, C., Silva, R., 2017. Counterfactual Fairness, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4066–4076.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2019. A Survey on Bias and Fairness in Machine Learning. ArXiv190809635 Cs.

Pearl, J., 2009. Causality: models, reasoning, and inference. Cambridge University Press, Cambridge, U.K. ; New York.

Pearl, J., Glymour, M., Jewell, N.P., 2016. Causal inference in statistics: a primer. Wiley, Chichester, West Sussex.

Pearl, J., Mackenzie, D., 2018. The book of why: the new science of cause and effect, First edition. ed. Basic Books, New York.

University of Pittsburgh/Carnegie Mellon University Center for Causal Discovery, 2019. bd2kccd/py-causal. bd2kccd.

Verma, S., Rubin, J., 2018. Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness - FairWare ’18. Presented at the the International Workshop, ACM Press, Gothenburg, Sweden, pp. 1–7. https://doi.org/10.1145/3194770.3194776

Wallace, L., 2020. Fair ML Reading Group in Melbourne [WWW Document]. URL https://github.com/summerscope/fair-ml-reading-group (accessed 1.12.20).

Wang, Y., Sridhar, D., Blei, D.M., 2019. Equal Opportunity and Affirmative Action via Counterfactual Predictions. ArXiv190510870 Cs Stat.