An Introduction to Algorithmic Fairness

Women and Non-Binary People in Mathematics Presentation

Dom Owens

University of Bristol

November 11th 2020

Contents

Introduction

Definitions and Measures

Causes

A Timeline

State of the Art

Introduction: Algorithmic Fairness

- We now make many decisions aided by statistical (AI, ML...) algorithms

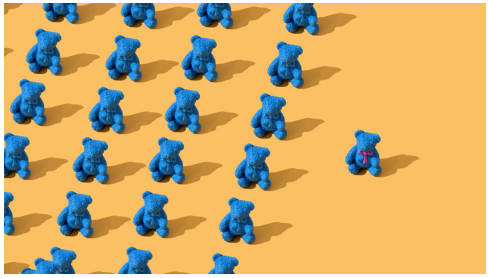
Introduction: Algorithmic Fairness

- We now make many decisions aided by statistical (AI, ML...) algorithms
- Some of these affect our lives
- Parole decisions, hiring, credit scores...

Introduction: Algorithmic Fairness

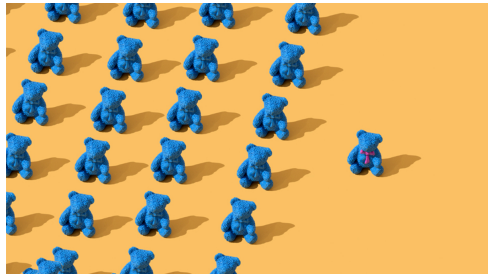
- We now make many decisions aided by statistical (AI, ML...) algorithms
- Some of these affect our lives
- Parole decisions, hiring, credit scores...
- Algorithms *should* make decisions better than humans: More data, much faster, **less biased?**

Definitions: What are we dealing with?



Disparate Treatment: *intentionally* treating an individual differently based on their membership in a protected class

Definitions: What are we dealing with?



Disparate Treatment: *intentionally* treating an individual differently based on their membership in a protected class

Disparate Impact: negatively affecting members of a protected class more than others even if by a *seemingly neutral* policy

Measures: How can we quantify it?

- We make a yes/no decision (e.g. should this person be given this job?)
- With fair decision making $Y \in \{0, 1\}$,
- With our model we decide $\hat{Y} = f(\mathbf{x}) \in \{0, 1\}$,
- Subjects have membership $S \in \{0, 1\}$ of a protected class (Minority ethnicity? Gender non-binary?...)

Measures: How can we quantify it?

- We make a yes/no decision (e.g. should this person be given this job?)
- With fair decision making $Y \in \{0, 1\}$,
- With our model we decide $\hat{Y} = f(\mathbf{x}) \in \{0, 1\}$,
- Subjects have membership $S \in \{0, 1\}$ of a protected class (Minority ethnicity? Gender non-binary?...)

Group Fairness ratio

$$R = \frac{P(\hat{Y} = 1|S = 0)}{P(\hat{Y} = 1|S = 1)}$$

If $R \leq 1 - \epsilon$, we have evidence of discrimination

Measures: How can we quantify it?

- We make a yes/no decision (e.g. should this person be given this job?)
- With fair decision making $Y \in \{0, 1\}$,
- With our model we decide $\hat{Y} = f(\mathbf{x}) \in \{0, 1\}$,
- Subjects have membership $S \in \{0, 1\}$ of a protected class (Minority ethnicity? Gender non-binary?...)

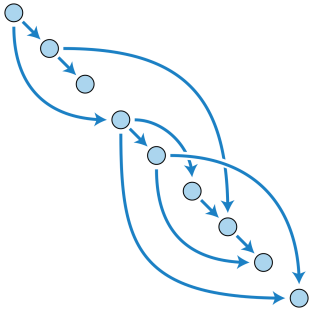
Group Fairness ratio

$$R = \frac{P(\hat{Y} = 1|S = 0)}{P(\hat{Y} = 1|S = 1)}$$

If $R \leq 1 - \epsilon$, we have evidence of discrimination

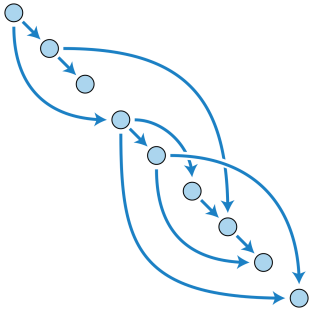
Outcome Test $P(Y = 1|\hat{Y} = 1, S = 0) = P(Y = 1|\hat{Y} = 1, S = 1)$
Is our model equally precise for those in/out of the protected group?

Causes: Where does this come from?



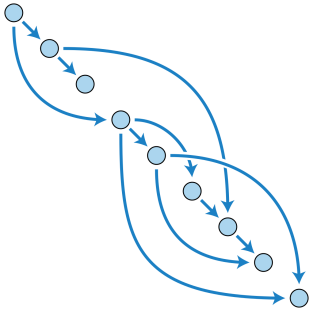
- **"Optimality"** - algorithm aims for accuracy for majority groups

Causes: Where does this come from?



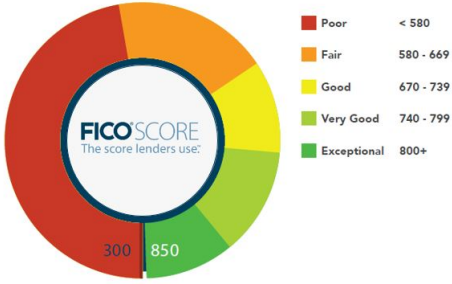
- **"Optimality"** - algorithm aims for accuracy for majority groups
- **Reconstructing** protected characteristics from correlated features

Causes: Where does this come from?



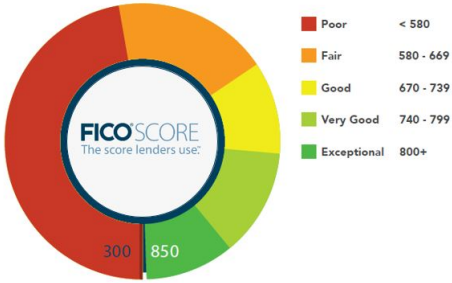
- **"Optimality"** - algorithm aims for accuracy for majority groups
- **Reconstructing** protected characteristics from correlated features
- **Biases in data set** (bad measurements, historical decisions...)

Case Study: FICO Credit Scoring (1989)



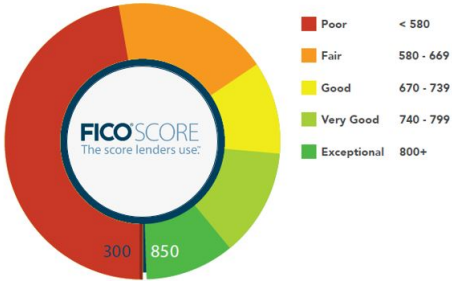
- Credit scoring: should we approve a loan to this person?

Case Study: FICO Credit Scoring (1989)



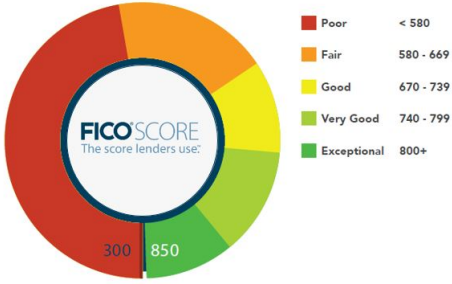
- Credit scoring: should we approve a loan to this person?
- Uses bank data **without protected characteristics**

Case Study: FICO Credit Scoring (1989)



- Credit scoring: should we approve a loan to this person?
- Uses bank data **without protected characteristics**
- Evidence shows algorithm still discriminates

Case Study: FICO Credit Scoring (1989)



- Credit scoring: should we approve a loan to this person?
- Uses bank data **without protected characteristics**
- Evidence shows algorithm still discriminates
- **Reconstructs these** from address, university, web activity...

Case Study: Technical Hiring at Amazon (2014-2018)

- Amazon filtered applications for technical jobs

Case Study: Technical Hiring at Amazon (2014-2018)

- Amazon filtered applications for technical jobs
- Algorithm looked for **similarities to previous hires**

Case Study: Technical Hiring at Amazon (2014-2018)

- Amazon filtered applications for technical jobs
- Algorithm looked for **similarities to previous hires**
- Learned e.g. gender through choice of language on CV

2020: A-level Grades

- A-level exams cancelled

2020: A-level Grades

- A-level exams cancelled
- Ofqual proposed grades from algorithm fed with **historical data on school achievement**

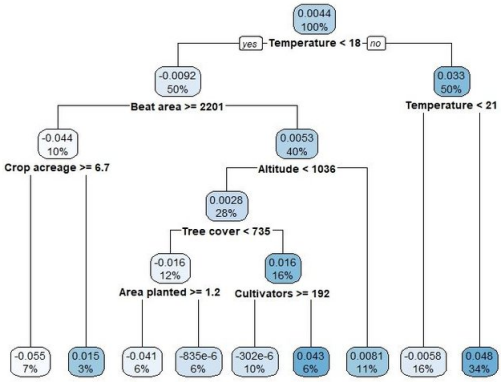
2020: A-level Grades

- A-level exams cancelled
- Ofqual proposed grades from algorithm fed with **historical data on school achievement**
- Were schools on a fair footing? Different resources, class sizes, student backgrounds...

2020: A-level Grades

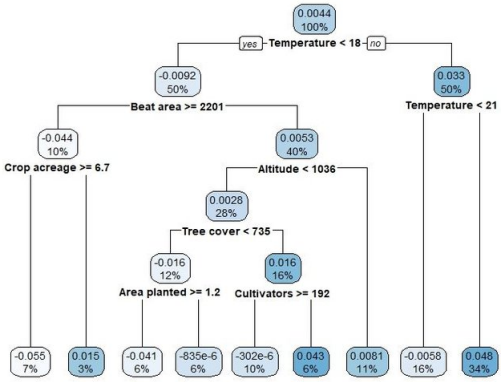
- A-level exams cancelled
- Ofqual proposed grades from algorithm fed with **historical data on school achievement**
- Were schools on a fair footing? Different resources, class sizes, student backgrounds...
- Small classes not graded by algorithm

State of the Art: Causal Models



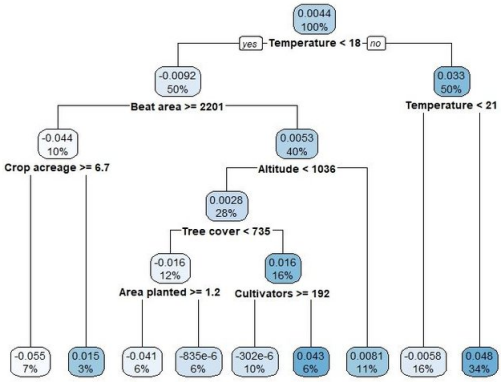
- Selecting a definition of fairness is hard, and depends on your application
- Some definitions are mutually incompatible

State of the Art: Causal Models



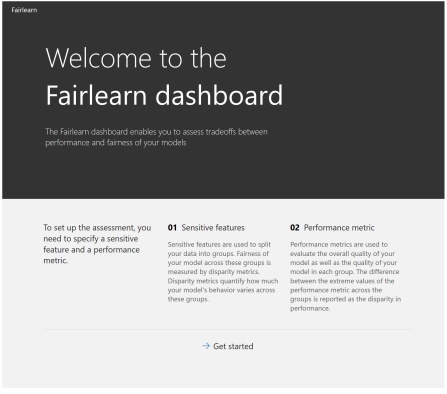
- Selecting a definition of fairness is hard, and depends on your application
- Some definitions are mutually incompatible
- Instead ask "Does the protected attribute have a causal effect on our prediction?"

State of the Art: Causal Models



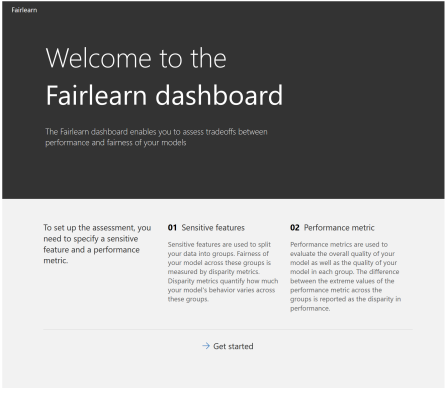
- Selecting a definition of fairness is hard, and depends on your application
- Some definitions are mutually incompatible
- Instead ask "Does the protected attribute have a causal effect on our prediction?"
- 30 years ago, this would require a Randomised Control Trial
- Now, we can ask counterfactual questions on a massive scale (e.g. Causal Tree models)

State of the Art: User Tools



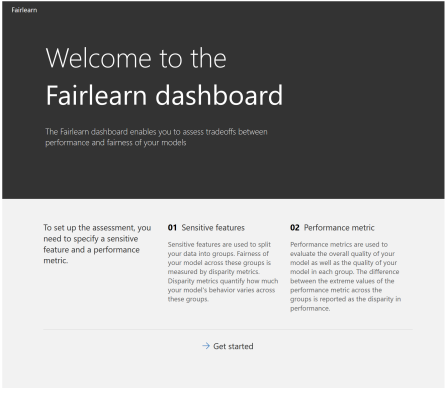
- How can cutting-edge research help practitioners?

State of the Art: User Tools



- How can cutting-edge research help practitioners?
- Tools (Fairlearn etc.) being developed and released
- Integrate fairness decisions into model deployment

State of the Art: User Tools



- How can cutting-edge research help practitioners?
- Tools (Fairlearn etc.) being developed and released
- Integrate fairness decisions into model deployment
- Smart speakers, facial recognition, chatbot responses...

Conclusion

- Used unwisely, algorithms can discriminate

Conclusion

- Used unwisely, algorithms can discriminate
- This can have potentially very unfair outcomes

Conclusion

- Used unwisely, algorithms can discriminate
- This can have potentially very unfair outcomes
- **Don't despair!** Ongoing research, increased awareness, shiny new tools for end users