Carnegie Mellon University

QA: ML Fairness (theoretical) case study

17-313 Spring 2023

S3D
Software and Societal Systems Department


FINAL FOUR
AZTECS
FINAL FOUR | VS. #9 FLORIDA ATLANTIC | SATURDAY 5:09 PM CT

Mid-Semester Feedback

- What should we start doing?
  - Extra Credit (yes)
  - Digitally submit activities (We will bring paper if you need it)
  - Opportunity to meet with Instructor to discuss homework (We will discuss today, my office hours are Tuesday 1-2pm, also happy to schedule other meeting times)
  - Notifications about when the weekly self assessment is due (added slack bot)
  - Introducing projects to the class as they are released would be helpful (doing today)
  - More lectures like extreme startup (trying to figure out some)
  - Guest Lectures (Trying to organize one)

S3D

Stop Doing

- The first day making everyone share their internship plans, especially given how stressful the current state is (sorry, that was not at all the goal)
- It would be better to use tried and true projects as a focus for the class. It was frustrating that the issues I had with the first project were mainly install issues that I couldn't really control. (There is a tension here)
- Attendance checking (We think participation in vital to your learning)

Keep Doing

- Candy!
- In-class activities
- Extreme startup and similar

Administrativia

- Project 4 - We will discuss today
  - https://cmu-313.github.io/projects/P4/

S3D

Learning goals

- Understand different fairness approaches
- Describe strengths and weaknesses of fairness approaches
- Reason about tradeoffs in fairness

ML Model = Unreliable Function

Object Detection Model

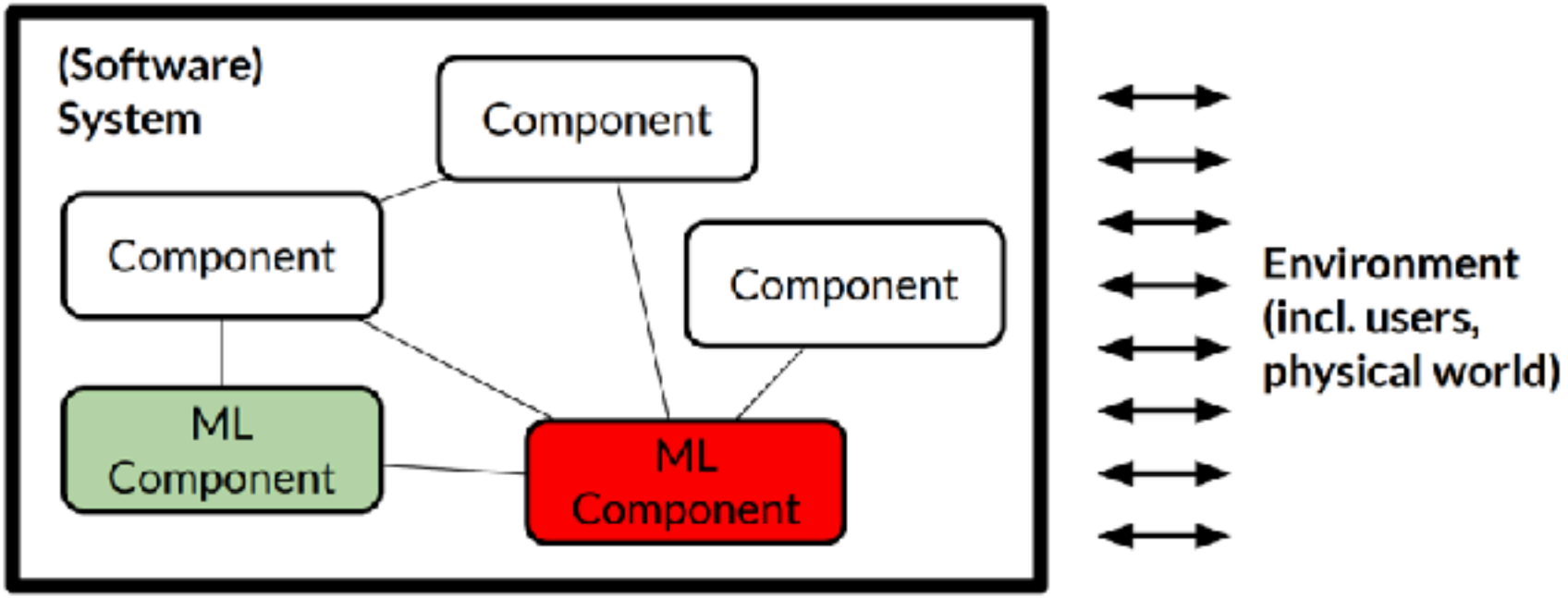
Building 99%
Path 97%
Plants 98%
Flowerpot 41%
Tree 4%

No guarantees, may make mistakes, confidence unreliable

Model often inscrutable, opaque

Evaluated in terms of accuracy, not correctness

What to do when the ML component makes mistake?

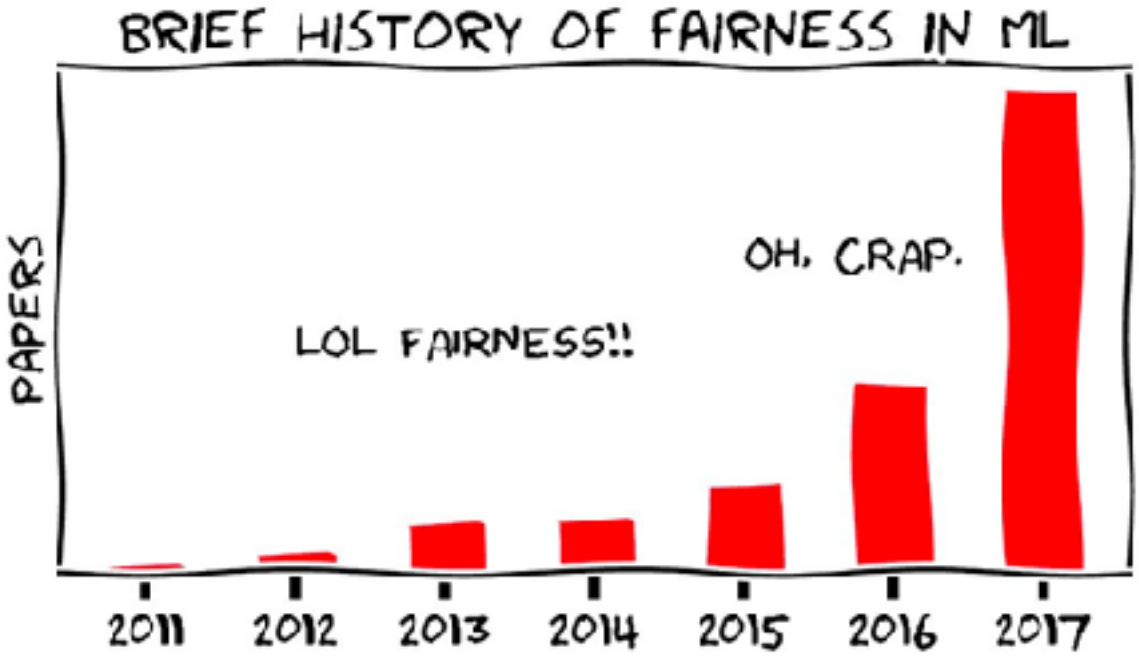


39

Fairness

ML Fairness

- Getting answers is the easy part… Asking the right questions is the hard part.



https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb

Perception:

Life is often not this simple…

Fairness

- Is a deeply technical topic, but we will discuss it at a higher level of abstraction.
- The formulas are important, but knowing which formula to apply is MUCH more important
- This is a special case of how to to test when the desired outcome is hard to measure.

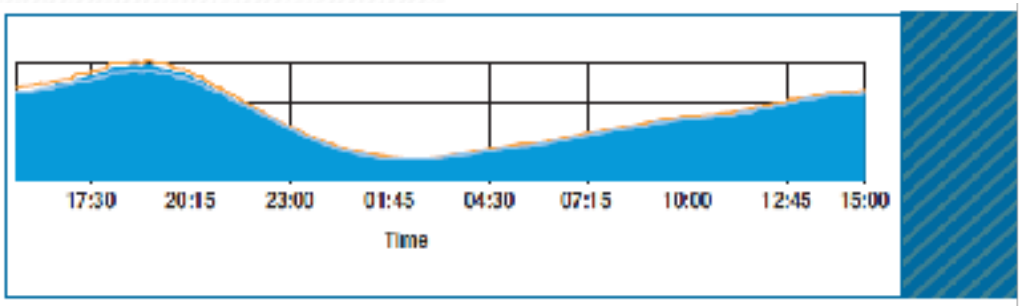


FIGURE 2. A graph of SPS ([stream] starts per second) over a 24-hour period. This metric varies slowly and predictably throughout a day. The orange line shows the trend for the prior week. The y-axis isn't labeled because the data is proprietary.

VS

What does ”fair” mean?

What is Fairness?

- Law
  - fairness includes protecting individuals and groups from discrimination or mistreatment with a focus on prohibiting behaviors, biases and basing decisions on certain protected factors or social group categories.
- Social Science
  - “often considers fairness in light of social relationships, power dynamics, institutions and markets.”3 Members of certain groups (or identities) that tend to experience advantages.

What is Fairness? continued

- Quantitative Fields
  - (i.e. math, computer science, statistics, economics): questions of fairness are seen as mathematical problems. Fairness tends to match to some sort of criteria, such as equal or equitable allocation, representation, or error rates, for a particular task or problem.
- Philosophy:
  - ideas of fairness “rest on a sense that what is fair is also what is morally right.” Political philosophy connects fairness to notions of justice and equity.

Fairness as QA

How can we define “fair”

- For the purposes of creating an oracle
- We must have a better definition than infamous 1964 Supreme Court obscenity test:
  - I shall not today attempt further to define [obscene material], and perhaps I could never succeed in intelligibly doing so. But *I know it when I see it*, and the motion picture involved in this case is not that.[

We don’t need to start from scratch…

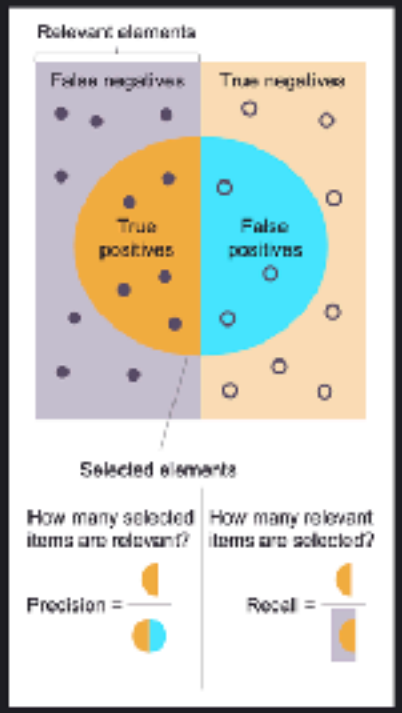
What can we do?

What can we do?

- We can evaluate with different criteria (e.g., different admissions score thresholds).
- We can observe the outcome of changing thresholds, and we can set different thresholds for different groups. (e.g., different SAT scores for in-state or out-of-state admissions)
- We can observe the impact of these different thresholds across a variety of metrics for each group.

First, some definitions:

Fairness Metrics



True Positives (TPs): 16	False Positives (FPs): 4
False Negatives (FNs): 6	True Negatives (TNs): 974
Precision = $\frac{TP}{TP+FP} = \frac{16}{16+4} = 0.800$	
Recall = $\frac{TP}{TP+FN} = \frac{16}{16+6} = 0.727$	

Female Patient Results	
True Positives (TPs): 10	False Positives (FPs): 1
False Negatives (FNs): 1	True Negatives (TNs): 488
Precision = $\frac{TP}{TP+FP} = \frac{10}{10+1} = 0.909$	
Recall = $\frac{TP}{TP+FN} = \frac{10}{10+1} = 0.909$	

Male Patient Results	
True Positives (TPs): 6	False Positives (FPs): 3
False Negatives (FNs): 5	True Negatives (TNs): 486
Precision = $\frac{TP}{TP+FP} = \frac{6}{6+3} = 0.667$	
Recall = $\frac{TP}{TP+FN} = \frac{6}{6+5} = 0.545$	

Source: Google ML Crash Course https://developers.google.com/machine-learning/crash-course/fairness/evaluating-for-bias

https://dattaraj-public.s3.ap-south-1.amazonaws.com/AI-Fairness_26th+April_DR.pdf

Varieties of fairness (names vary)

- **Group unaware**
  - Ignore group data (one group could get excluded)
- **Group thresholds**
  - Different rules per group (rules differ by group)
- **Demographic parity**
  - Same percentage in pool as outcomes (might result in random selection)
- **Equal opportunity**
  - Equal chance out positive outcomes regardless of groups (focus on individual, rules differ per group)

S3D

Group unaware

- We use some criteria that is independent of the categories we are considering for fairness.
- Guarantees about outcomes: None. One group may be completely excluded

Group thresholds

- We create different criteria per group
- Guarantees about outcomes: candidates inside a group are evaluated by the same standard as others inside the same group.
- By definition, groups are evaluated to a different standard (e.g., different fitness standards by gender in US Military)

Demographic parity

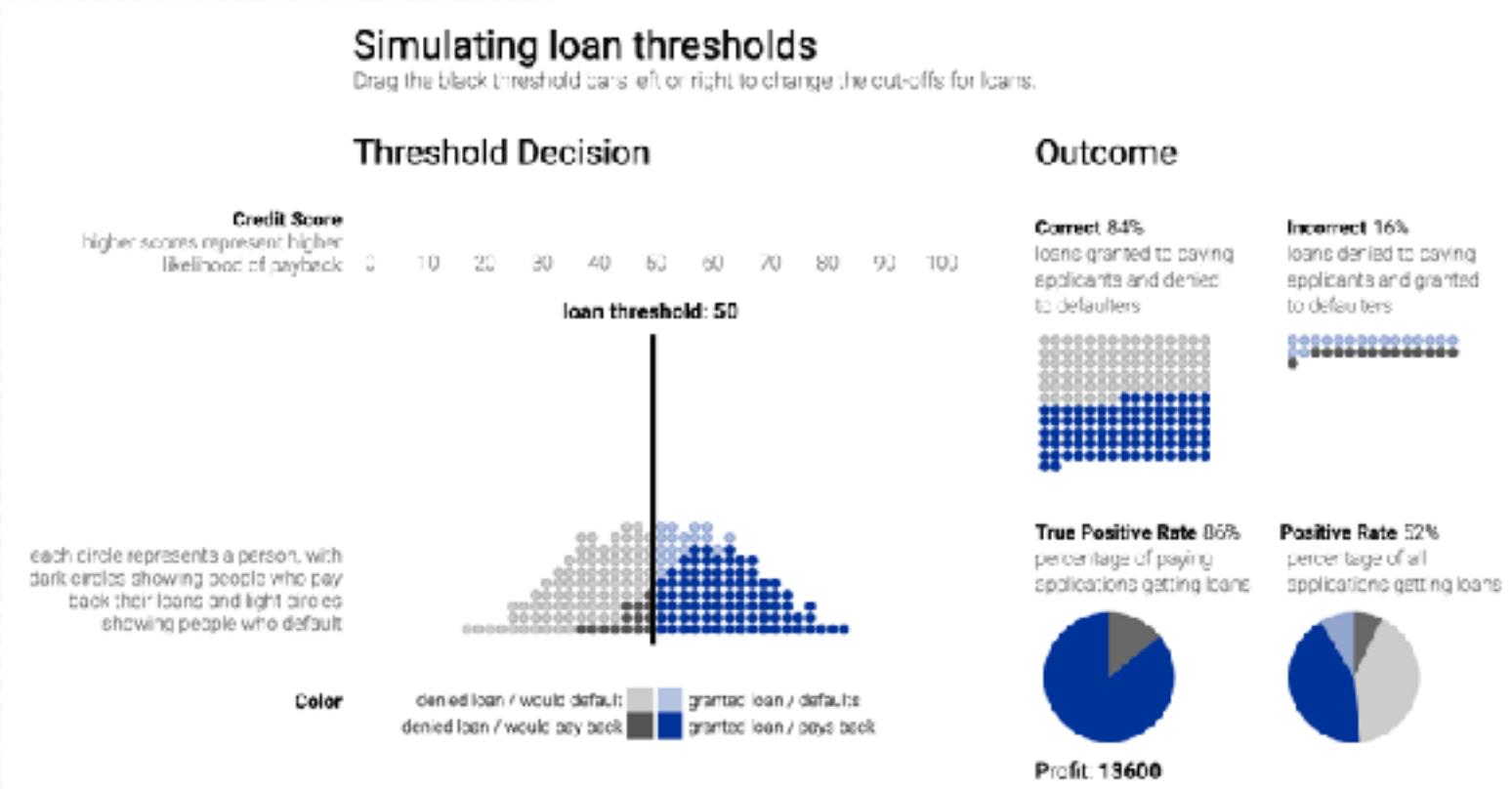
- We create different criteria per group, with a goal of similar outcomes in a certain dimension.
- Guarantees about outcomes: The same percentage of each group will have a positive outcome. e.g., 25 % accepted from group A, 25% accepted from group B.
- However, can result in different true positive rates, (e.g., more “worthy” candidates denied in group A than group B.

Equal opportunity

- We create different criteria per group, with a goal of similar outcomes for similar individuals across groups.
- Guarantees about outcomes: The same number of true positives per group. e.g., 80% true positives in group A, 80% true positives in group B.
- However, can result in different positive rates across groups.

S3D

Explainability



https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Activity

Consider the different approaches to fairness. Can you come up with different scenarios where each fairness approach might or might not be appropriate?

Remember the fairness approaches are:

- Group unaware
- Group thresholds
- Demographic parity
- Equal opportunity

Resources

- Fairness Textbook:
- https://fairmlbook.org/testing.html