COMPAS, ProPublica, and Beyond

First things first...

- Are there things that AI/ML algos **shouldn't** be used for?
- What if the tech is **better** than humans?
- What if the tech is the **same** as humans?
- AI/ML vs "hand coded"?
- Can algos be or have **moral(s)**?

Northpointe and COMPAS

- Northpointe (now Equivant) built a criminal recidivism risk tool/algo: COMPAS
- First developed in 1998
- Asseses risks and needs in 43 categories
- Fair amount of sociology/theory; let's look at "practitioner's Guide"
- So what are the inputs to this algo/model?

ProPublica

- Non-profit investigative journalism org
- Founded 2007
- 100+ journalists
- Many awards
- Conducted investigation of COMPAS in 2016

ProPublica vs. COMPAS

- PP obtains COMPAS scores ($\hat{y}$'s) for $n \approx 11K$ defendants in Broward County FL via public records request (x's?)
- Joins with public records of criminal history, race, gender, age and:

  recidivism (y's)

ProPublica Dataset

- Demographics: name, gender, age, race
- Criminal history: juvenile & adult counts
- Charge description
- Various procedural vars
- Recidivism outcomes
- COMPAS scores

ProPublica Findings

- Distribution of COMPAS risk scores: skews low for whites, more uniform for blacks
- Q: Is this problematic?
- PP then uses ML to build a predictive model for COMPAS scores

PP Findings, cont'd

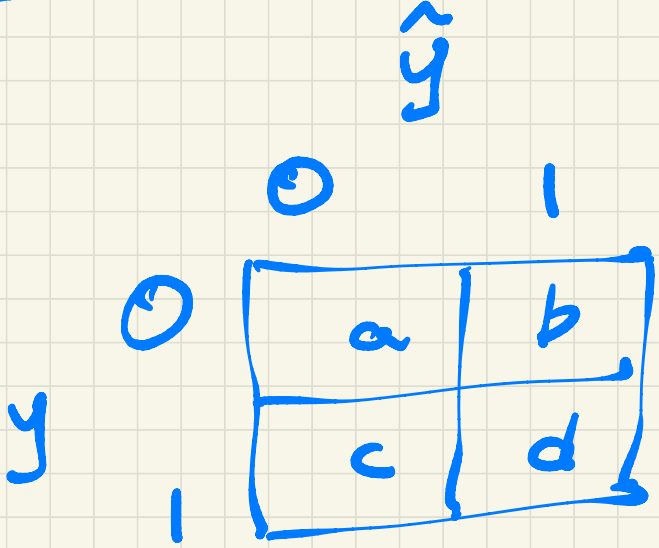
- Then analyze confusion matrix for a classifier for recidivism:

$$\hat{y} = \begin{cases} 1 \text{ if COMPAS high} \\ 0 \text{ if COMPAS low} \end{cases}$$

- Look at black & white matrices separately

- Focus on false positive rates

Confusion Matrices



- error rate $\hat{\varepsilon} = \frac{b+c}{a+b+c+d}$
- false positive rate $= \frac{b}{a+b}$

$= P_{\hat{}}[\hat{y}=1 | y=0]$

- false negative rate $= \frac{c}{c+d}$

$= P_{\hat{}}[\hat{y}=0 | y=1]$

- true positive rate $= \frac{d}{c+d}$

$= Pr[\hat{y}=1 | y=1]$

- true negative rate $= \frac{a}{a+b}$

$= Pr[\hat{y}=0 | y=0]$

- FPR + TNR = 1.0
- FNR + TPR = 1.0

Thresholding Scores

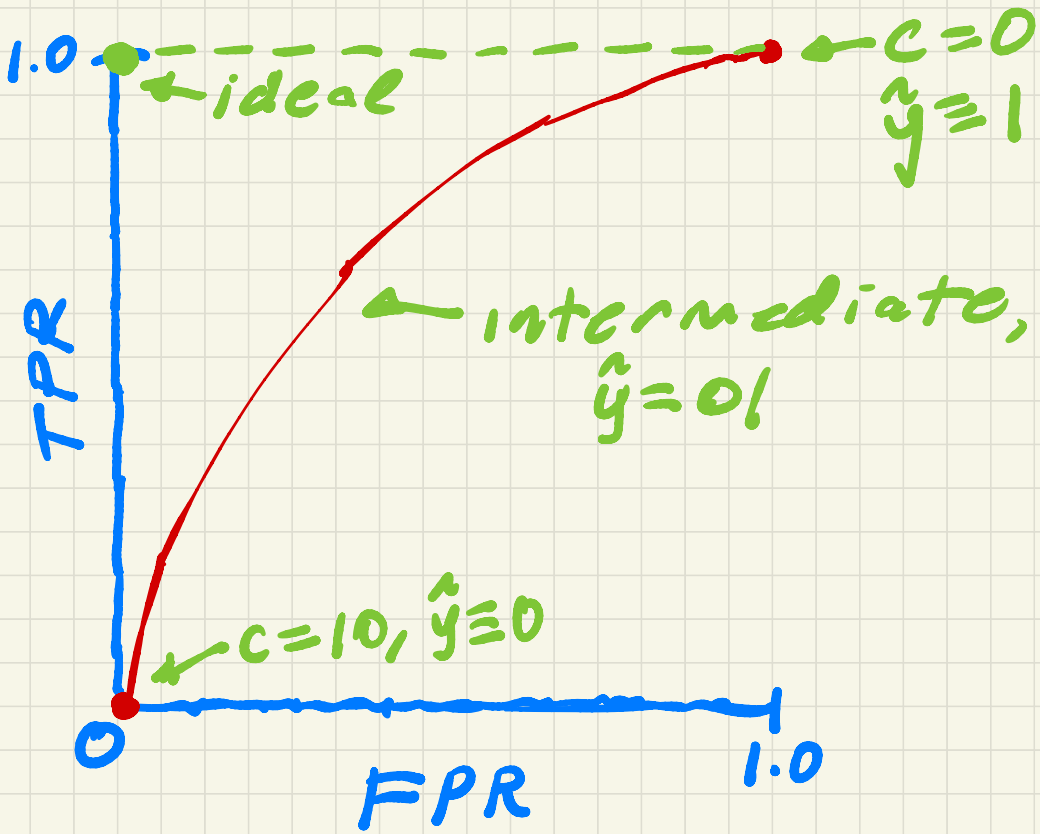
- Have some "risk" score r, say $0 \leq r \leq 10$
- Prediction $\hat{y}$ based on r:

$$\hat{y} = 1 \iff r \geq c$$

- $c = 0$: FPR = 1.0, FNR = 0
  TPR = 1.0, FPR = 0
- $c = 10$: FPR = 0, FNR = 1.0
  TPR = 0, TNR = 1.0

Area Under Curve (AUC)

- As we increase c, we trace out a curve in FPR/TPR space:



Northpointe Response to PP

- Plot separate curves for black & white pops
- Claim that by picking single threshold, get different points
- Apples & Oranges
- Further claim AUCs are equal

Summary

- PP: Your algo is unfair, black FPR >> white FPR
- Northpointe: Wrong, it's fair, black AUC = white AUC

Who's "right"?

Why choose?