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/algorithm: COMPAS
- First developed in 1998
- Assesses 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 ($y$'s) for ~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:

\[ y = \begin{cases} 1 & \text{if COMPAS high} \\ 0 & \text{if COMPAS low} \end{cases} \]

• Look at black & white matrices separately

• Focus on false positive rates
Confusion Matrices

\[
\hat{y} \\
\begin{array}{cc}
0 & 1 \\
0 & a & b \\
1 & c & d \\
\end{array}
\]

- **Error rate**
  \[
  \hat{\varepsilon} = \frac{b+c}{a+b+c+d}
  \]

- **False positive rate**
  \[
  \beta = \frac{b}{a+b} = P(\hat{y}=1|y=0)
  \]

- **False negative rate**
  \[
  \gamma = \frac{c}{c+d} = P(\hat{y}=0|y=1)
  \]
- true positive rate \[ = \frac{d}{c+d} \]
  \[ = \Pr[\hat{y}=1|y=1] \]

- true negative rate \[ = \frac{a}{a+b} \]
  \[ = \Pr[\hat{y}=0|y=0] \]

- \( \text{FPR} + \text{TNR} = 1.0 \)
- \( \text{FNR} + \text{TPR} = 1.0 \)
Thresholding Scores

- Have some "risk" score \( r \), say \( 0 \leq r \leq 10 \)
- Prediction \( \hat{y} \) based on \( r \):
  \[ \hat{y} = 1 \iff r \geq c \]

- \( c = 0 \): \( \text{FPR} = 1.0, \text{FNR} = 0 \)
  \( \text{TPR} = 1.0, \text{FPR} = 0 \)

- \( c = 10 \): \( \text{FPR} = 0, \text{FNR} = 1.0 \)
  \( \text{TPR} = 0, \text{TNR} = 1.0 \)
Area Under Curve (AUC)

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

- Ideal case: $\hat{y} = 1$
- Intermediate, $\hat{y} = 0$
- $c = 10, \hat{y} = 0$
- $c = 0, \hat{y} = 1$
Northpointe Response to PP

• Plot separate curves for black and white pops

• Claim that by picking single threshold, get different points

• Apples & Oranges

• Further claim AVCs 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?