PLSC 476: Empirical Legal Studies

Christopher Zorn

April 15, 2021

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

COMPAS

- *Correctional Offender Management Profiling for Alternative Sanctions*
- Created by Northpointe Inc. (now Equivant) in 1989
- Inputs = 137 offender and offense characteristics
- "Core Risk Scales":
  - Risk of New Violent Crime ("Violent Recidivism")
  - Risk of General Recidivism
  - Pretrial Risk
- Specifics of their predictive algorithm are proprietary...
- Used in NY, CA, WI, other jurisdictions
- Use was challenged (and upheld) in Loomis v. Wisconsin 881 N.W.2d 749 (WI 2016)

The ProPublica Analysis

- FOIA request $\rightarrow N = 18610$ offenders from Broward Co., FL (all offenders scored in 2013 & 2014)
- Pretrial detention only $\rightarrow N = 11757$
- Used public records to gather data on demographic variables and criminal histories...
- ProPublica: "Recidivism" = arrest on a new charge within two years (=1, 0 if not)
- Focus: **Racial bias in COMPAS scores**
- Data (and R code!) are publicly available at: `https://github.com/propublica/compas-analysis`

Details of their analysis are here.

Recidivism: Predictors

- **Sex** (male vs. female)
- **Age** (in years)
- **Race/Ethnicity** (Black, Asian, white, Hispanic, Native American, Other)
- **Juvenile Felonies** (number)
- **Juvenile Misdemeanors** (number)
- **Prior Arrests** (number)
- **Felony Charge** (=1; misdemeanor / other = 0)

Logistic Regression: Recidivism

	Outcome: Recidivism
Sex: Male	0.368^{***}
	(0.058)
Age	-0.040^{***}
	(0.002)
Race: Asian	-0.407
	(0.350)
Race: White	-0.111^{**}
	(0.050)
Race: Hispanic	-0.303^{***}
	(0.083)
Race: Native American	-0.168
	(0.398)
Race: Other	-0.307^{***}
	(0.105)
Juvenile Felonies	0.028
	(0.050)
Juvenile Misdemeanors	0.068
	(0.054)
Prior Arrests	0.116^{***}
	(0.005)
Felony Charge	0.018
	(0.048)
Constant	0.053
	(0.093)
Observations	10,331
Log Likelihood	$-6,052$
Akaike Inf. Crit.	12,128
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

In-Sample Predictions

```
> df$Recid.Probs<-predict(Recid.fit,type="response")
> df$Recid.Preds<-ifelse(df$Recid.Probs>0.5,1,0)

> conf<-xtabs(~Recid+Recid.Preds,data=df)
> conf
     Recid.Preds
Recid    0    1
    0 6387  471
    1 2700  773

> ((conf[1,1]+conf[2,2])/(sum(conf))) * 100 # Accuracy
[1] 69.31
```

Predictive Accuracy: ROC and AUC-ROC

For each observation i, we generate a $\Pr(\widehat{\text{Recidivism}}_i = 1)$ between zero and one, as:

$$\Pr(\widehat{\text{Recidivism}}_i) = \frac{\exp(\mathbf{X}_i \hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{X}_i \hat{\boldsymbol{\beta}})}.$$

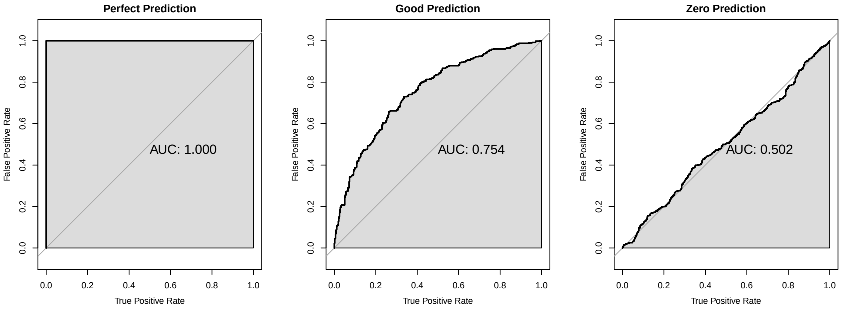
We then might imagine assigning individuals to either $\widehat{\text{Recidivism}}_i = 0$ or $\widehat{\text{Recidivism}}_i = 1$ for different values of a "threshold" τ of $\Pr(\widehat{\text{Recidivism}}_i = 1)$:

- If $\tau = 0$, then we'd assign every observation to be $\widehat{\text{Recidivism}}_i = 1$
- If $\tau = 1$, then we'd assign every observation to be $\widehat{\text{Recidivism}}_i = 0$
- In between – for, say, $\tau = \ell$, we assign observations with $\Pr(\widehat{\text{Recidivism}}_i = 1) > \ell$ to be $\widehat{\text{Recidivism}}_i = 1$, and those with $\Pr(\widehat{\text{Recidivism}}_i = 1) \leq \ell$ to be $\widehat{\text{Recidivism}}_i = 0$

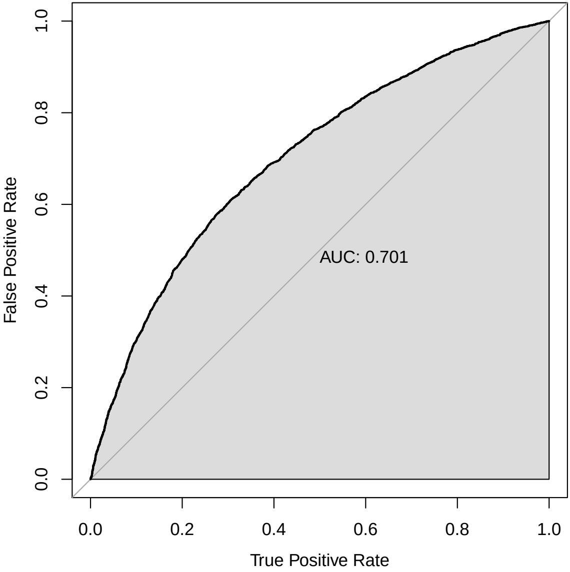
As we vary τ from zero to one, we'd get varying levels of "true positives" vs. "false positives":

- When $\tau = 0$ and every observation has $\widehat{\text{Recidivism}}_i = 1$, we'd have a 100 percent "true positive" rate but also a 100 percent "false positive" rate
- When $\tau = 1$ and every observation has $\widehat{\text{Recidivism}}_i = 0$, we'd have a 0 percent "true positive" rate but also a 0 percent "false positive" rate
- In between, we'd get varying levels of "true positives" vs. "false positives," depending on the predictive accuracy of the model

ROC Curves: Examples



ROC Curve: Our Model



Validation: "Hold-out"

Q: How well does our model do at predicting out of sample?

"Hold-out" validation:

- Randomly split the data into a "training" set (90%) and a "test" set (10%)
- Fit the predictive model to the "training" data
- Use the model parameters to generate predictions in the "test" (out-of-sample) data
- Examine the predictive accuracy of the model on the test data

Test Data Predictions

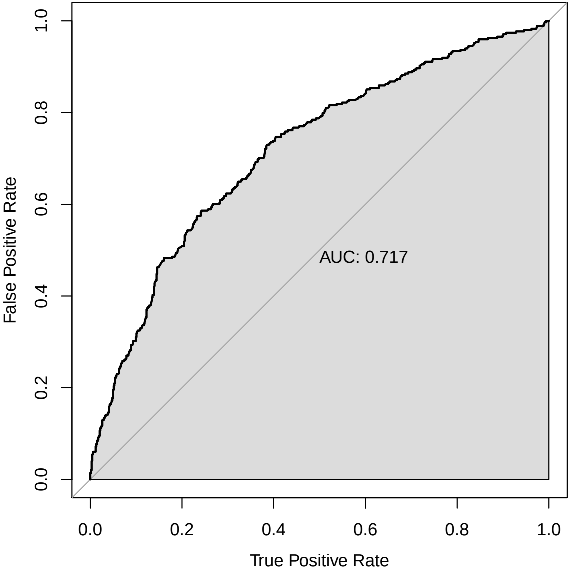
```
> Train.fit<-glm(Recid~Sex+Age+Race+JuvFelonies+JuvMisdem+
                 Priors+FelonyCharge,data=df.Train,
                 family="binomial")
> df.Test$Recid.Probs<-predict(Train.fit,newdata=df.Test,
                                type="response")
> df.Test$Recid.Preds<-ifelse(df.Test$Recid.Probs>0.5,1,0)

> conf2<-xtabs(~Recid+Recid.Preds,data=df.Test)

> conf2
     Recid.Preds
Recid   0   1
    0 649  36
    1 276  72

> ((conf2[1,1]+conf2[2,2])/(sum(conf2))) * 100 # Accuracy
[1] 69.8
```

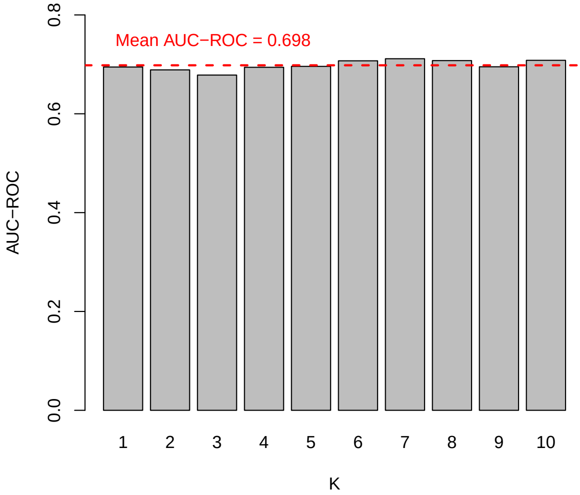
ROC Curve: Test Data



K-Fold Validation

- Randomly split the data into $k = \{1, 2, ...K\}$ equally-sized subsets
- For the first subset with $k = 1$:
  1. Use the data for $k = \{2, 3, ...K\}$ as the "training" data
  2. Fit the predictive model to the "training" data
  3. Use the resulting model parameters to generate predictions in the "test" (out-of-sample) data
  4. Examine the predictive accuracy of the model on the test data
- Repeat 1-4 for $k = 2$, $k = 3$, ... $k = K$
- Average over the K model fit measures to assess the predictive validity of the model

K-Fold Validation



Step Two: COMPAS Analysis

COMPAS provides two recidivism risk evaluations:

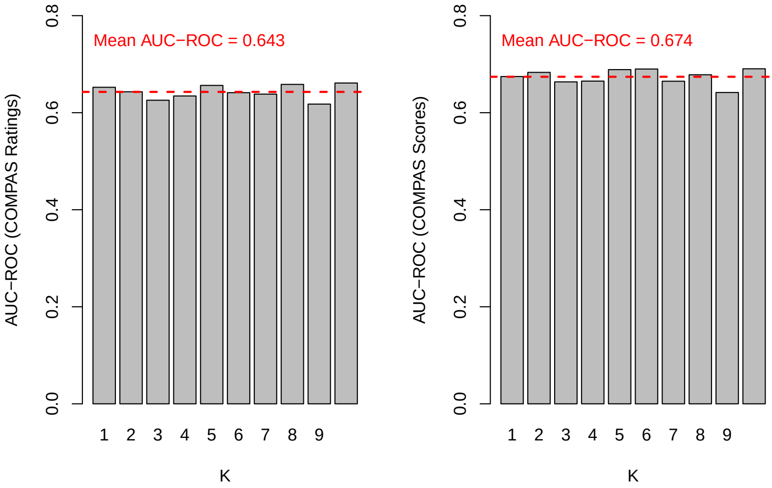
- A COMPAS Rating: "High Risk," "Medium Risk," or "Low Risk"
- A COMPAS Score: A numeric rating of recidivism risk ranging from 1 (lowest risk) to 10 (highest risk)
- From the ProPublica webpage: "(A)ccording to Northpointe's practitioners guide, COMPAS 'scores in the medium and high range garner more interest from supervision agencies than low scores, as a low score would suggest there is little risk of general recidivism' "

Goal: Assess the predictive value of COMPAS ratings and scores...

COMPAS Rating/Score Models

	Outcome: Recidivism	
	COMPAS Ratings	COMPAS Score
COMPAS Rating: Low Risk	-1.331^{***}	
	(0.055)	
COMPAS Rating: Medium Risk	-0.450^{***}	
	(0.060)	
COMPAS Rating: N/A	-1.634^{**}	
	(0.783)	
COMPAS Score		0.214^{***}
		(0.008)
Constant	0.130^{***}	-1.678^{***}
	(0.045)	(0.043)
Observations	10,331	10,320
Log Likelihood	−6,246	−6,172
Akaike Inf. Crit.	12,499	12,348
Note:		*p$<$0.1; **p$<$0.05; ***p$<$0.01

COMPAS Model: Cross-Validation



Wrap-Up

- In one data set, a relatively-simple (7-variable) model predicts recidivism slightly better than COMPAS scores/ratings
- One *could* do the same analysis for violent recidivism, using the COMPAS "Violent Risk" scores
- Some additional readings:
  - An excellent follow-up story on the ProPublica analysis, asking the question "What does it mean for an algorithm to be fair?"
  - The Age of Secrecy and Unfairness in Recidivism Prediction in the *Harvard Data Science Review* (read the commentaries, too)
  - A recent study on the impact of RAIs on judges' decisions...