PLSC 476: Empirical Legal Studies

Christopher Zorn

April 13, 2021

CAREERS IN THE LAW 2021

Join us for Virtual Careers in the Law on Friday, April 16, 2021

Careers in the Law is designed to allow students to learn more about a variety of practice areas by speaking with alumni and friends of Penn State who have experience in various practice areas.

Benefits for Alumni and Friends

- Meet current Penn State Law students
- Share your experiences with the next generation of attorneys
- Network with other attorneys
- Reconnect with your classmates and professors

Benefits for Students

- Hone your networking skills
- Learn more about the wide variety of practice areas in the law
- Connect with practicing attorneys
- Learn about potential job openings

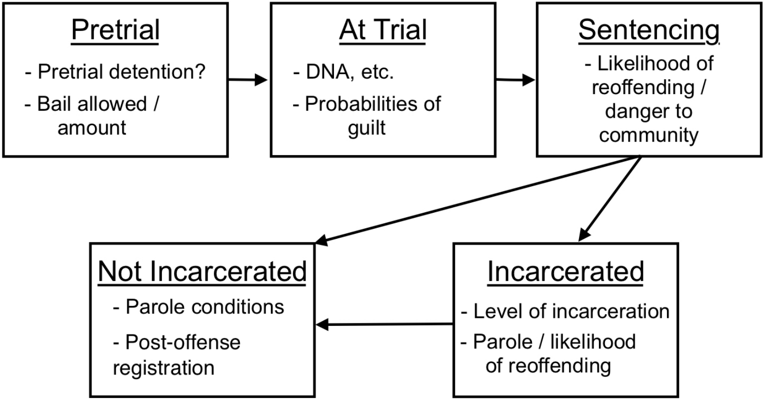
What Students Can Expect

- Network with alumni, faculty, and friends of Penn State that represent 15-20 practice areas
- 5-8 OCI-participating law firms from a variety of locales
- Business attire is strongly preferred

Register

Date/Time:
Friday, April 16, 2021 - 5:00pm to 7:30pm
Location:
Online (Remo)

Empirics in Criminal Justice



“Risk Assessment Instruments” (RAIs)

- Tools for predicting risk / negative outcomes
- Inputs: Characteristics of the offender / offense(s) / context (“risk factors”)
- Output(s): Estimates of the risk of bad outcomes (fleeing/failing to appear, reoffense, etc.)
- Uses:
  - Inform judge / parole board / etc. decision-making
  - Determine offender’s required actions
  - Inform communities (e.g., via registries)
- Methods: Vary widely...

Example RAI: New York's Sex Offender Registration RAI

SEX OFFENDER REGISTRATION ACT
RISK ASSESSMENT INSTRUMENT

RISK FACTOR	VALUE	SCORE
I. CURRENT OFFENSE(S)		
1. Use of Violence		
Used forcible compulsion	+10	
Inflicted physical injury	+15	
Armed with a dangerous instrument	+30	
2. Sexual Contact with Victim		
Contact over clothing	+5	
Contact under clothing	+10	
Sexual intercourse, deviate sexual intercourse or aggravated sexual abuse	+25	
3. Number of Victims		
Two	+20	
Three or more	+30	
4. Duration of offense conduct with victim		
Continuing course of sexual misconduct	+20	
5. Age of victim		
11 through 16	+20	
10 or less, 63 or more	+30	
6. Other victim characteristics		
Victim suffered from mental disability or incapacity or from physical helplessness	+20	
7. Relationship with victim		
Stranger or established for purpose of victimizing or professional relationship	+20	
II. CRIMINAL HISTORY		
8. Age at first act of sexual misconduct		
20 or less	+10	
9. Number and nature of prior crimes		
Prior history/no sex crimes or felonies	+5	
Prior history/non-violent felony	+15	
Prior violent felony, or misdemeanor sex crime or endangering welfare of a child	+30	
10. Recency of prior felony or sex crime		
Less than 3 years	+10	
11. Drug or Alcohol abuse		
History of abuse	+15	
COLUMNS 1- 11 SUBTOTAL		

SEX OFFENDER REGISTRATION ACT
RISK ASSESSMENT INSTRUMENT

RISK FACTOR	VALUE	SCORE
III. POST-OFFENSE BEHAVIOR		
12. Acceptance of Responsibility		
Not accepted responsibility	+10	
Not accepted responsibility and refused or expelled from treatment	+15	
13. Conduct while confined / supervised		
Unsatisfactory	+10	
Unsatisfactory with sexual misconduct	+20	
IV. RELEASE ENVIRONMENT		
14. Supervision		
Release with specialized supervision	0	
Release with supervision	+5	
Release without supervision	+15	
15. Living / employment situation		
Living or employment inappropriate	+10	
COLUMNS 12-15 SUBTOTAL		
COLUMNS 1-11 SUBTOTAL		
TOTAL RISK FACTOR SCORE (add 2 subtotals)		
1 2 3		

Level 1 (low)	=	0 to +70
Level 2 (moderate)	=	+75 to +105
Level 3 (high)	=	+110 to +300

Note: The Sex Offender Registration Act requires the court or Board of Examiners of Sex Offenders to consider any victim impact statement in determining a sex offender's level of risk.

Offender Name: ________

NYSID #: ________

Docket #: ________

Risk Level: ________

Assessor's Signature ________

Date: ________

A. Overrides (If any override is circled, offender is presumptively a Level 3)

1. Offender has a prior felony conviction for a sex crime
2. Offender inflicted serious physical injury or caused death
3. The offender has made a recent threat that he will reoffend by committing a sexual or violent crime
4. There has been a clinical assessment that the offender has a psychological, physical, or organic abnormality that decreases ability to control impulsive sexual behavior

B. Departure

1. A departure from the risk level is warranted

   ☐ Yes ☐ No

2. If yes, circle the appropriate risk level 1 2 3
3. If yes, explain the basis for departure (See Summary)

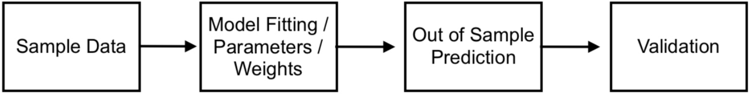
(Source)

Goal: *Out of sample* ("OOS") prediction of some event, using existing data on (past) events / non-events.

Challenges:

- Counterfactuals ("What if...?")
- Overfitting
- Validation / assessing performance

The usual process:



Predictive Models: An Example

- Sample (or "Training") Data:
  - N observations
  - *Outcome* Y_i (say, $\in \{0, 1\}$)
  - *K predictors* $\mathbf{X}_i$

- Model Fitting:

$$Y_i = f[\beta_0 + \beta_1 X_{1i} + ... + \beta_1 X_{Ki}]$$

- $\rightarrow$ Parameter estimates $\{\hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_K\}$

- Out-of-Sample (or "Test") Data: $j \in \{1, 2, ...M\}$ observations with information on $\mathbf{X}_j$

- $\rightarrow$ Predictions: For test-data observation j with $[X_{1j}, X_{2j}, ...X_{Kj}]$,

$$\hat{Y}_j = f[\hat{\beta}_0 + \hat{\beta}_1 X_{1j} + ... + \hat{\beta}_K X_{Kj}]$$

- Validation: Compare $\hat{Y}_j$ to Y_j.... (how?)

Validation Methods

- "Holdout"
  - *Hold out* (a random) part of the sample data (say, 10 percent)
  - Fit the models to the non-held-out data
  - Generate predictions & validate on the "held out" fraction

- *K*-fold Cross-Validation
  - Split the data into k random subgroups
  - Conduct "holdout" validation k times, using each of the successive subgroups as the "test" data and the remaining $k-1$ subgroups as the "training" data
  - Average over the K performance measures

- Others (e.g., "Leave-One-Out" Cross-Validation)

Assessing Predictive Performance

The "Confusion Matrix":

	Predicted Values $\hat{Y}_j$		
Actual Values Y_j	$\hat{Y}_j = 0$	$\hat{Y}_j = 1$	Total
$Y_j = 0$	"True Negatives"	"False Positives"	All Actual Negatives
$Y_j = 1$	"False Negatives"	"True Positives"	All Actual Positives
Total	All Predicted Negatives	All Predicted Positives	All Test Observations

Concepts:

- True Positive Rate ("sensitivity"): $TPR = \frac{\text{True Positives}}{\text{All Actual Positives}}$
- True Negative Rate ("specificity"): $TNR = \frac{\text{True Negatives}}{\text{All Actual Negatives}}$
- False Positive Rate: $FPR = \frac{\text{False Positives}}{\text{All Actual Negatives}} = 1 - \textit{Specificity}$

Key: **TPR and FPR are positively related...**

“AUC-ROC”

Area Under the “Receiver Operating Characteristic” Curve

- Plot: TPR vs. FPR for various prediction thresholds

- “AUC-ROC”: Always > 0.5, but < 1.0

- Provides a summary measure of how well the model predicts outcomes...
  - AUC-ROC = 0.90-1.00 $\rightarrow$ Excellent (A)
  - AUC-ROC = 0.80-0.90 $\rightarrow$ Good (B)
  - AUC-ROC = 0.70-0.80 $\rightarrow$ Fair (C)
  - AUC-ROC = 0.60-0.70 $\rightarrow$ Poor (D)
  - AUC-ROC = 0.50-0.60 $\rightarrow$ Total Failure (F)

- Examples on Thursday...

RAIs: Empirical and Ethical Issues

Potential Pros:

- Data-driven $\rightarrow$ more accurate
- Reduce / eliminate decision-makers' (implicit and explicit) biases
- Fairness: Individuals with similar characteristics receive similar treatment

Potential Cons:

- "Black boxes" / lack of transparency
- Need for subjectivity / individualized treatment
- **Algorithmic bias** $\rightarrow$ lack of fairness

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
  - 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)

RAIs: Dressel & Farid (2018)

- Compare COMPAS predictions with nonexpert human guesses, based on vignettes with 5/6 inputs / predictors ("features")
- Based on $N = 1000$ defendant histories randomly selected from a *Pro Publica* FOIA-requested database of defendants in Broward County, FL during 2013-14.
- Findings:
  - Humans predicted as well as the COMPAS algorithm
  - Humans were no more biased than COMPAS vis-à-vis the race of the offender
  - Inclusion of information about offender's race in the predictors had no effect on the findings

Table 1. Human versus COMPAS algorithmic predictions from 1000 defendants. Overall accuracy is specified as percent correct, AUC-ROC, and criterion sensitivity (d') and bias (β). See also Fig. 1.

	(A) Human (no race)	(B) Human (race)	(C) COMPAS
Accuracy (overall)	67.0%	66.5%	65.2%
AUC-ROC (overall)	0.71	0.71	0.70
d'/β (overall)	0.86/1.02	0.83/1.03	0.77/1.08
Accuracy (black)	68.2%	66.2%	64.9%
Accuracy (white)	67.6%	67.6%	65.7%
False positive (black)	37.1%	40.0%	40.4%
False positive (white)	27.2%	26.2%	25.4%
False negative (black)	29.2%	30.1%	30.9%
False negative (white)	40.3%	42.1%	47.9%

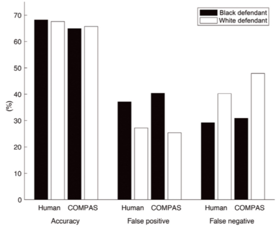


Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

RAIs: Lin et al. (2020)

- Replication & extension of Dressel & Farid (D&F), considering:
  - Effect of additional inputs / predictors
  - Effect of "kind" (w/feedback) vs. "wicked" (no feedback) environments
  - Different *base rates* for reoffending

- Findings:
  - D&F's findings were replicated, under similar circumstances
  - With additional ("enriched") inputs, algorithms outperformed humans
  - Humans did better with feedback than without[1]
  - Humans generally did worse when base rates of reoffending were unbalanced (here, low)

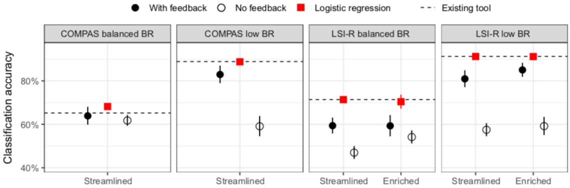


Fig. 2. Classification accuracy of human predictions, statistical models, and existing tools. Classification accuracy is shown for (i) human predictions, with and without immediate feedback; (ii) a logistic regression model that we trained using the same information provided to study participants; and (iii) the existing tools, COMPAS or LSI-R. For participants in the feedback condition, only the last 10 responses for each participant were used, to account for the effects of learning. Error bars represent 95% confidence intervals and are typically smaller than the height of red squares for the logistic regression models.

[1] Except in the case of "streamlined" features and balanced base rates.