Building a Robot Judge:
Data Science for Decision-Making

9. Algorithms and Decisions I

Weekly Q&A

http://bit.ly/BRJ_Padlet9

Outline

Learning Objectives

1. Implement and evaluate machine learning pipelines.
2. Implement and evaluate causal inference designs.

Learning Objectives

1. Implement and evaluate machine learning pipelines.
2. Implement and evaluate causal inference designs.
3. **Understand how (not) to use data science tools (ML and CI) to support expert decision-making.**

Learning Objectives

1. Implement and evaluate machine learning pipelines.
2. Implement and evaluate causal inference designs.
3. **Understand how (not) to use data science tools (ML and CI) to support expert decision-making.**
   - Appreciate the connections/distinctions between **prediction**, **inference**, and **decisions**.
   - Evaluate proposed policies/systems that use algorithms for decision support – along accuracy, bias, gaming, and other dimensions.

Learning Objectives

1. Implement and evaluate machine learning pipelines.
2. Implement and evaluate causal inference designs.
3. **Understand how (not) to use data science tools (ML and CI) to support expert decision-making.**
   - Appreciate the connections/distinctions between **prediction**, **inference**, and **decisions**.
   - Evaluate proposed policies/systems that use algorithms for decision support – along accuracy, bias, gaming, and other dimensions.
   - Read and critique research papers reporting on these policies/systems.
   - If you are signed up for the project: Implement/analyze such a system and write a paper about it.

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Yes, if:

- Costs of fixing problems are mostly homogeneous.
  - it could be that buildings with high fire risk also have old wiring that is costly to replace $\rightarrow$ better to inspect buildings with cheaper fixes.

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Yes, if:

- Costs of fixing problems are mostly homogeneous.
  - it could be that buildings with high fire risk also have old wiring that is costly to replace $\rightarrow$ better to inspect buildings with cheaper fixes.
- Establishments subject to inspection do not respond to the algorithm.
  - some firms may be more sensitive to penalties than others, or may be easier for some firms to game the predictors.

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Yes, if:

- Costs of fixing problems are mostly homogeneous.
  - it could be that buildings with high fire risk also have old wiring that is costly to replace $\rightarrow$ better to inspect buildings with cheaper fixes.
- Establishments subject to inspection do not respond to the algorithm.
  - some firms may be more sensitive to penalties than others, or may be easier for some firms to game the predictors.
  - some firms might know they have a low inspection due to a low violation probability (because of their neighborhood, for example), and reduce safety measures.

Example: Allocating Inspectors

Athey 2017; Glaeser et al, AER P&P 2016

- Governments can conserve resources by inspecting establishments that are likely to have violations, e.g.:
  - NYC's Firecast algorithm predicts fire risk and code violation
  - Glaeser et al.'s (2016) algorithm predicts health code violations in Boston restaurants (improved violation detection rates by 30%).
- Are these predictions sufficient for optimal allocation of inspectors?

Yes, if:

- Costs of fixing problems are mostly homogeneous.
  - it could be that buildings with high fire risk also have old wiring that is costly to replace $\rightarrow$ better to inspect buildings with cheaper fixes.
- Establishments subject to inspection do not respond to the algorithm.
  - some firms may be more sensitive to penalties than others, or may be easier for some firms to game the predictors.
  - some firms might know they have a low inspection due to a low violation probability (because of their neighborhood, for example), and reduce safety measures.
- Overall, the inspection policy problem is a causal inference problem:
  - What is the expected improvement in overall quality of units (e.g., food poisoning rates) in the city under a new inspector allocation regime?

Example: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the return on investment for advertising clicks was estimated at 1400%.

Example: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the return on investment for advertising clicks was estimated at 1400%.
  - eBay then purchased a ton of search engine advertising.

Example: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the return on investment for advertising clicks was estimated at 1400%.
  - eBay then purchased a ton of search engine advertising.
- Blake et al. (2015) used city-level longitudinal data, with a difference-in-difference approach, to estimate causal effect of search engine advertising on sales.

Example: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the return on investment for advertising clicks was estimated at 1400%.
  - eBay then purchased a ton of search engine advertising.
- Blake et al. (2015) used city-level longitudinal data, with a difference-in-difference approach, to estimate causal effect of search engine advertising on sales.
  - the true return on investment: -63%!
  - eBay stopped buying so much advertising after that.

Example: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the return on investment for advertising clicks was estimated at 1400%.
  - eBay then purchased a ton of search engine advertising.
- Blake et al. (2015) used city-level longitudinal data, with a difference-in-difference approach, to estimate causal effect of search engine advertising on sales.
  - the true return on investment: -63%!
  - eBay stopped buying so much advertising after that.
- The problem with the previous approach: confounding.
  - many people who clicked on search advertisements would have purchased items from eBay anyway.

Algorithms are opaque and complex

- Most decision-makers / stakeholders want to understand the reason that a decision has been made.
  - or decision-makers may need to commit a decision rule to memory (e.g., doctors).

Algorithms are opaque and complex

- Most decision-makers / stakeholders want to understand the reason that a decision has been made.
  - or decision-makers may need to commit a decision rule to memory (e.g., doctors).
- Transparency and simplicity considerations might lead analysts to sacrifice predictive power.
  - more on this in week 12

Algorithms are opaque and complex

- Most decision-makers / stakeholders want to understand the reason that a decision has been made.
  - or decision-makers may need to commit a decision rule to memory (e.g., doctors).
- Transparency and simplicity considerations might lead analysts to sacrifice predictive power.
  - more on this in week 12
- Finally, there are considerations of fairness or discrimination.
  - more on this in week 11

Activity: Review, Four Types of Confounders

https://twitter.com/ASlavitt/status/1319870491063177216

https://padlet.com/eash44/p4h4614l70q4ixzd

Outline

Humans vs. Machines

- Given the same data/features X, machines tend to beat humans:
  - Games: Chess, AlphaGo, Poker
  - Image classification
  - Question answering (IBM Watson)

Humans vs. Machines

- Given the same data/features X, machines tend to beat humans:
  - Games: Chess, AlphaGo, Poker
  - Image classification
  - Question answering (IBM Watson)
- But humans see more than machines do.
  - Humans make decisions based on $X_H \supset X$.
  - could include common sense, knowledge about the future, etc.

Humans vs. Machines

- Given the same data/features X, machines tend to beat humans:
  - Games: Chess, AlphaGo, Poker
  - Image classification
  - Question answering (IBM Watson)
- But humans see more than machines do.
  - Humans make decisions based on $X_H \supset X$.
  - could include common sense, knowledge about the future, etc.
- So when should machines make decisions?

Bail Decision: Detain or Relase

- Costs of detention (avg. 2-3 months):
  - Consequential for jobs, families
  - Costs to taxpayers of jails

Bail Decision: Detain or Relase

- Costs of detention (avg. 2-3 months):
  - Consequential for jobs, families
  - Costs to taxpayers of jails
- Costs of release:
  - failure to appear at trial
  - commit more crimes

Bail Decision: Detain or Relase

- Costs of detention (avg. 2-3 months):
  - Consequential for jobs, families
  - Costs to taxpayers of jails
- Costs of release:
  - failure to appear at trial
  - commit more crimes
- Judge is implicitly making an assessment/prediction about these outcomes, and then making a decision based on that.

COMPAS

- COMPAS is a risk-scoring algorithm used by many U.S. courts (more than 1 million cases).
  - "Correctional Offender Management Profiling for Alternative Sanctions"
  - judges see a risk assessment of how likely a defendant is to commit more crimes if released on bail.

COMPAS

- COMPAS is a risk-scoring algorithm used by many U.S. courts (more than 1 million cases).
  - "Correctional Offender Management Profiling for Alternative Sanctions"
  - judges see a risk assessment of how likely a defendant is to commit more crimes if released on bail.
  - model is proprietary / closed-source.
  - apparently uses 137 predictors, which include defendant characteristics and criminal history.

COMPAS

- COMPAS is a risk-scoring algorithm used by many U.S. courts (more than 1 million cases).
  - "Correctional Offender Management Profiling for Alternative Sanctions"
  - judges see a risk assessment of how likely a defendant is to commit more crimes if released on bail.
  - model is proprietary / closed-source.
  - apparently uses 137 predictors, which include defendant characteristics and criminal history.
- ProPublica built a dataset from 7000 criminal cases in Florida where COMPAS was used.
  - see this week's homework assignment.

COMPAS

- COMPAS is a risk-scoring algorithm used by many U.S. courts (more than 1 million cases).
  - "Correctional Offender Management Profiling for Alternative Sanctions"
  - judges see a risk assessment of how likely a defendant is to commit more crimes if released on bail.
  - model is proprietary / closed-source.
  - apparently uses 137 predictors, which include defendant characteristics and criminal history.
- ProPublica built a dataset from 7000 criminal cases in Florida where COMPAS was used.
  - see this week's homework assignment.
- Dress and Farid (Science Advances 2018):
  - a logistic regression model with two features is just as accurate as COMPAS

COMPAS

- COMPAS is a risk-scoring algorithm used by many U.S. courts (more than 1 million cases).
  - "Correctional Offender Management Profiling for Alternative Sanctions"
  - judges see a risk assessment of how likely a defendant is to commit more crimes if released on bail.
  - model is proprietary / closed-source.
  - apparently uses 137 predictors, which include defendant characteristics and criminal history.
- ProPublica built a dataset from 7000 criminal cases in Florida where COMPAS was used.
  - see this week's homework assignment.
- Dress and Farid (Science Advances 2018):
  - a logistic regression model with two features is just as accurate as COMPAS
  - majority vote by 20 non-specialist human participants (Amazon Mechanical Turk) predicts recidivism as accurately as COMPAS.

Kleinberg et al (2018) Data

- 750,000 individuals arrested in New York City between 2008-2013
- Same data on prior history that is available to judge (rap sheet, current offense, etc.)
  - Data on subsequent crimes to develop and evaluate performance of algorithm
  - Define "crime" as failing to show up at trial; objective is to jail those with highest risk of committing this crime
    - Other definitions of crime (e.g., repeat offenses) yield similar results

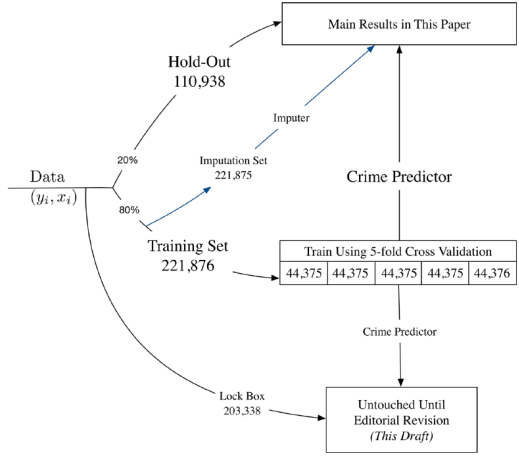


FIGURE I

Partition of New York City Data (2008–13) into Data Sets Used for Prediction and Evaluation

Data: Defendant Features

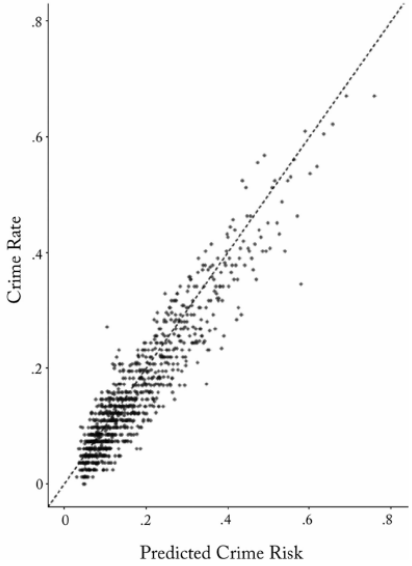
Kleinberg et al (2019)

`Age at first arrest, Times sentenced residential correction, Level of charge, Number of active warrants, Number of misdemeanor cases, Number of past revocations, Current charge domestic violence, Is first arrest, Prior jail sentence, Prior prison sentence, Employed at first arrest, Currently on supervision, Had previous revocation, Arrest for new offense while on supervision or bond, Has active warrant, Has active misdemeanor warrant, Has other pending charge, Had previous adult conviction, Had previous adult misdemeanor conviction, Had previous adult felony conviction, Had previous Failure to Appear, Prior supervision within 10 years`

- excludes race, gender, and religion
  - not legal to include – will come back to this issue

- Use labeled dataset (released defendants), to predict whether they fail to appear or commit more crimes.
  - preferred model: gradient boosting (GB): test-set AUC = .71

- Use labeled dataset (released defendants), to predict whether they fail to appear or commit more crimes.
  - preferred model: gradient boosting (GB): test-set AUC = .71



- Use labeled dataset (released defendants), to predict whether they fail to appear or commit more crimes.
  - preferred model: gradient boosting (GB): test-set AUC = .71

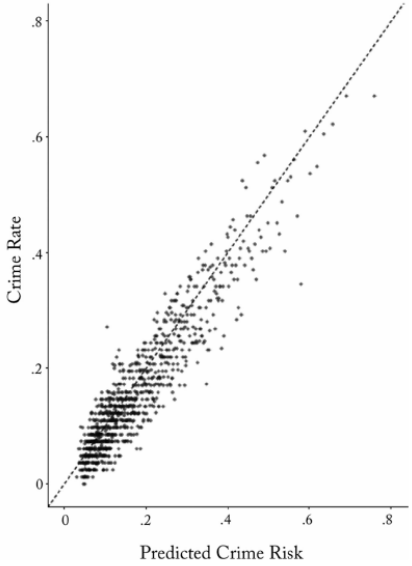
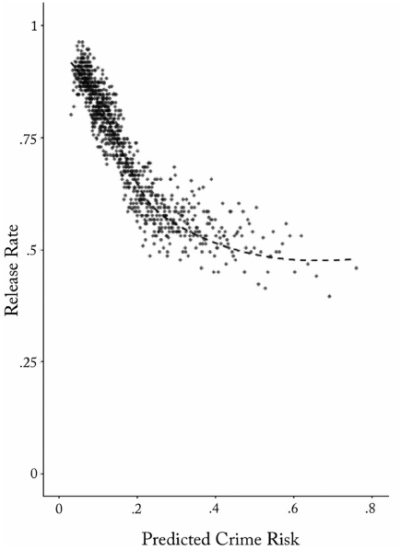


TABLE II
COMPARING LOGISTIC REGRESSION TO MACHINE-LEARNING PREDICTIONS OF CRIME RISK

Predicted risk percentile	ML/ logit overlap	Average observed crime rate for cases identified as high risk by: Both ML and logit	ML only	Logit only	All ML cases	All logit cases
1%	30.6%	0.6080 (0.0309)	0.5440 (0.0209)	0.3996 (0.0206)	0.5636 (0.0173)	0.4633 (0.0174)
5%	59.9%	0.4826 (0.0101)	0.4090 (0.0121)	0.3040 (0.0114)	0.4531 (0.0078)	0.4111 (0.0077)
10%	65.9%	0.4134 (0.0067)	0.3466 (0.0090)	0.2532 (0.0082)	0.3907 (0.0054)	0.3589 (0.0053)
25%	72.9%	0.3271 (0.0038)	0.2445 (0.0058)	0.1608 (0.0049)	0.3048 (0.0032)	0.2821 (0.0031)

Zoom Private Chat to me:
What do these exhibits tell us?

What human judges do



- Human judges tend to follow what algorithm suggests.
- But judge sees factors the machine does not
  - makes decisions based on $\Pr(Y|X_H)$
    - X_H includes other factors not seen by the machine – e.g., defendant demeanor.
  - Machine makes decisions based on $\Pr(Y|X)$, $X \subset X_H$.

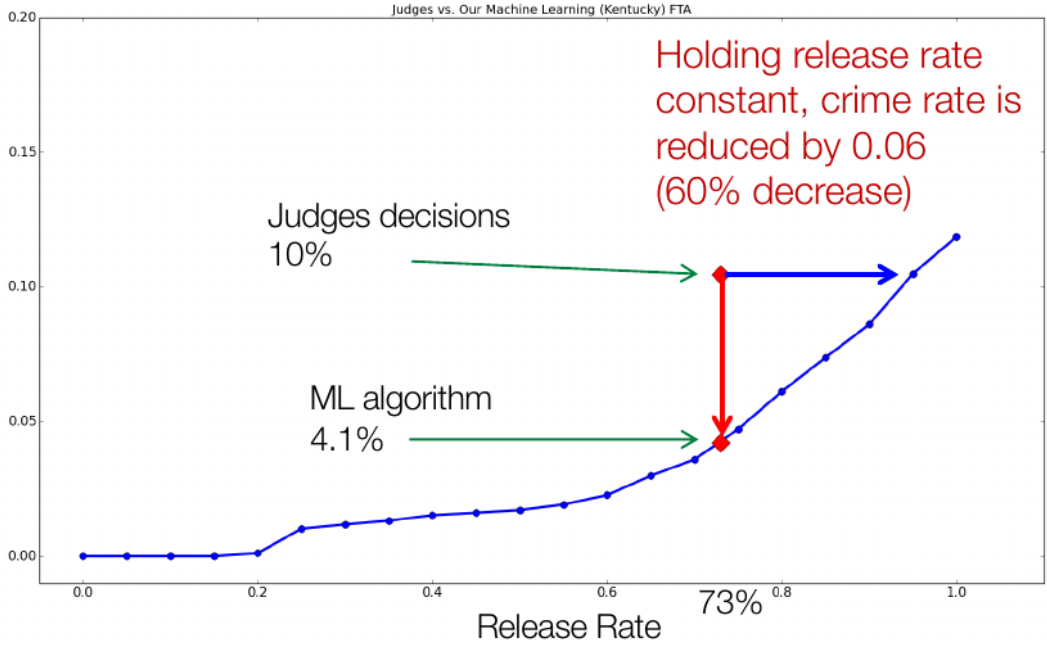
Prediction $\rightarrow$ Release Rule

- Kleinberg et al consder the following release rule based on recidivism predictions:
  - For every defendant predict $\hat{Y}(X_i)$
  - Sort by increasing $\hat{Y}(X_i)$
  - Release bottom N defendants, jail the rest.

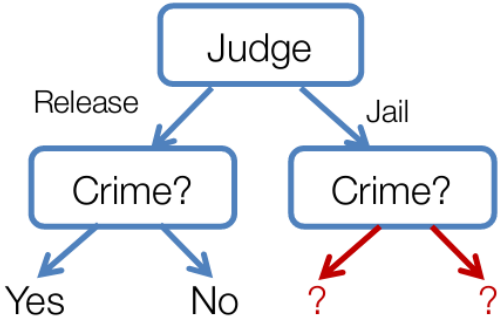
Prediction $\rightarrow$ Release Rule

- Kleinberg et al consder the following release rule based on recidivism predictions:
  - For every defendant predict $\hat{Y}(X_i)$
  - Sort by increasing $\hat{Y}(X_i)$
  - Release bottom N defendants, jail the rest.
- What is the fraction released vs. crime rate tradeoff?

Compare Judge to ML in predicted crime rate

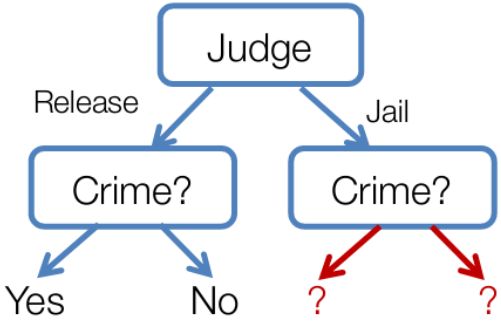


Problem: Judge is selectively labeling the dataset



- We can only train on released people:
  - By jailing, judge is selectively hiding labels!

Problem: Judge is selectively labeling the dataset

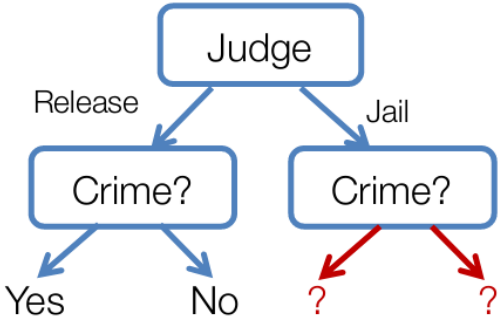


- We can only train on released people:
  - By jailing, judge is selectively hiding labels!

Selective labels introduce bias. Example:

- Say young people with no tattoos have no risk for crime. Judge releases them.
- Machine observes age, but does not observe tattoos.

Problem: Judge is selectively labeling the dataset



- We can only train on released people:
  - By jailing, judge is selectively hiding labels!

Selective labels introduce bias. Example:

- Say young people with no tattoos have no risk for crime. Judge releases them.
- Machine observes age, but does not observe tattoos.
- Machine would falsely conclude that all young people do no crime, and release all young people.

Solution: Contraction

- Selection problem is one-sided: We observe counterfactual (crime rate) for released defendants, but not jailed defendants.

Solution: Contraction

- Selection problem is one-sided: We observe counterfactual (crime rate) for released defendants, but not jailed defendants.

- **Contraction**:
  - Take released population of a lenient judge.
  - Then ask which additional defendant we would jail to minimize crime rate.
  - Compare change in crime rate to that observed for stricter judge.

Solution: Contraction

- Selection problem is one-sided: We observe counterfactual (crime rate) for released defendants, but not jailed defendants.

- **Contraction**:
  - Take released population of a lenient judge.
  - Then ask which additional defendant we would jail to minimize crime rate.
  - Compare change in crime rate to that observed for stricter judge.
- **This approach requires random assignment of cases to judges to work.**
  - **Zoom Private Chat: Explain why.**

Comparing Machine Judges (Left Panel) to Human Judges (Right Panel)

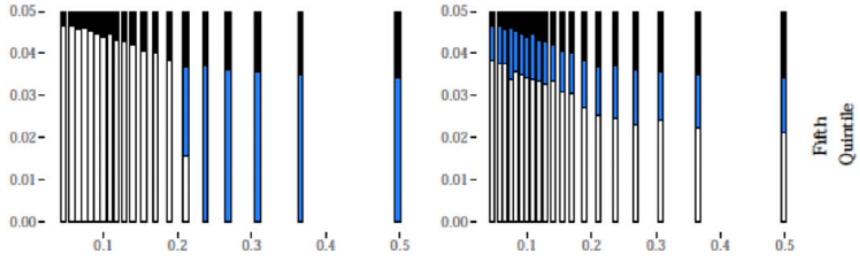


FIGURE VI

Who Do Stricter Judges Jail and Who Would the Algorithm Jail? Comparing Predicted Risk Distributions across Leniency Quintiles

- black = even most lenient judges (bottom quintile) would jail this defendant.
- blue = additional jailed by the strictest judges (top quintile). left panel = algorithm, right panel = human judges.
- white = who is released by all judges

Comparing Machine Judges (Left Panel) to Human Judges (Right Panel)

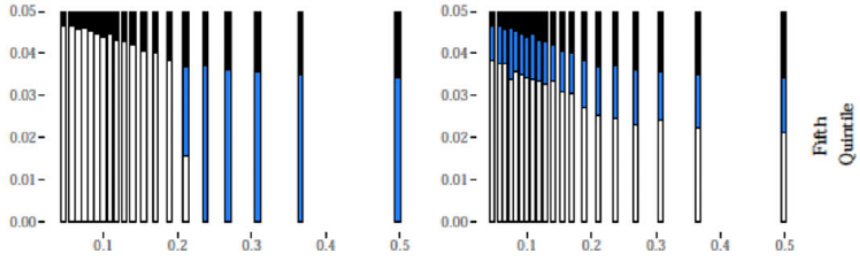


FIGURE VI

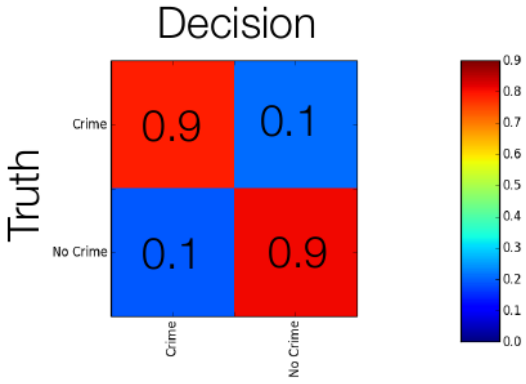
Who Do Stricter Judges Jail and Who Would the Algorithm Jail? Comparing Predicted Risk Distributions across Leniency Quintiles

- black = even most lenient judges (bottom quintile) would jail this defendant.
- blue = additional jailed by the strictest judges (top quintile). left panel = algorithm, right panel = human judges.
- white = who is released by all judges
- **Zoom Private Chat: What does this graph show?**

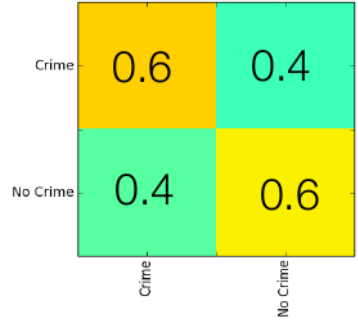
Analyzing judge mistakes

Analyzing judge mistakes

Experienced Judges

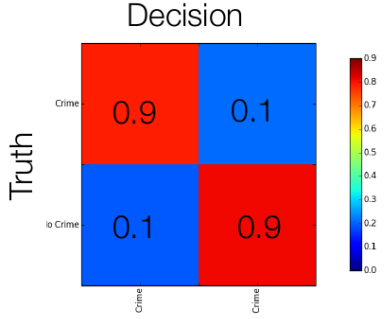


Inexperienced Judges

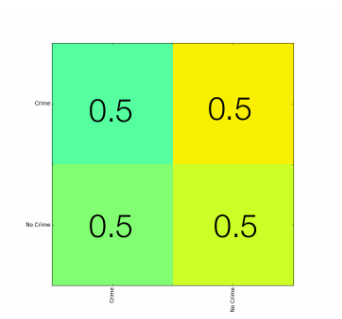


Source: Jure Leskovec slides.

Analyzing judge mistakes



Defendants who are single, did felonies, and moved a lot are accurately judged



Defendants who have kids are confusing to judges

- Or are judges balancing crime risk against kids' welfare?
  Source: Jure Leskovec slides.

Discussion

- Algorithms can help us understand if human judges make mistakes, and diagnose reasons for bias.

Discussion

- Algorithms can help us understand if human judges make mistakes, and diagnose reasons for bias.
- Not just about prediction. Key is starting with decision:
  - Performance benchmark: Current "human" decisions

Discussion

- Algorithms can help us understand if human judges make mistakes, and diagnose reasons for bias.
- Not just about prediction. Key is starting with decision:
  - Performance benchmark: Current "human" decisions
- Question: What are we really optimizing?

Labels are Driven by Decisions

- We don't see labels of people that are jailed
- This is a broader problem in policymaking systems:
  - Prediction $\rightarrow$ Decision $\rightarrow$ Outcome
- Which outcomes we see depends on our decisions

Focusing on re-arrest rates is limited

- Is minimizing the crime rate really the right goal?
- There are other important factors
  - Consequences of jailing on the family
  - Jobs and the workplace
  - Future behavior of the defendant
- How could we measure/model these?

Lightning Essay

For last minutes of class:

`https://bit.ly/BRJ-W9-Essay`