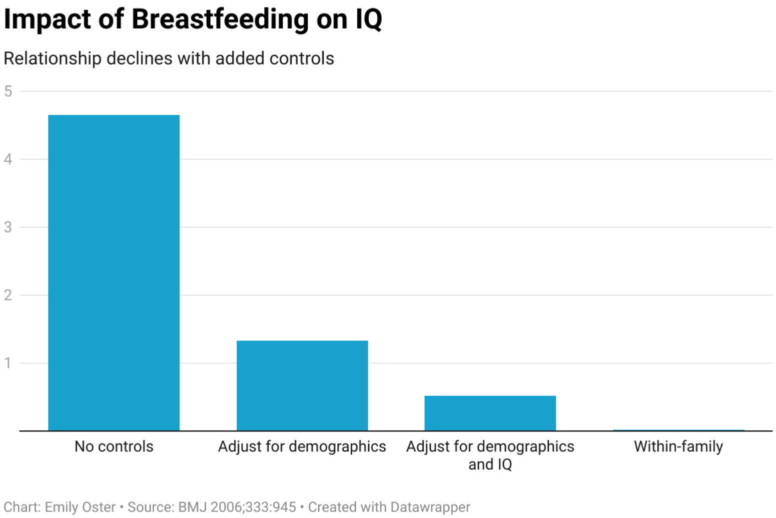
Building a Robot Judge: Data Science for Decision-Making

10. Algorithms and Decisions II

Correlation vs. Causation: Effect of Breastfeeding on Child IQ



https://toktopics.com/2023/03/24/why-i-look-at-data-differently-by-emily-oster/

When are recividism risk scores sufficient for bail/release decisions?

(1) Benefits of detention are mostly homogeneous.

- why not?

When are recividism risk scores sufficient for bail/release decisions?

(1) Benefits of detention are mostly homogeneous.

- why not?

(2) Potential offenders do not change behavior in response to the algorithm.

- why not?

When are recividism risk scores sufficient for bail/release decisions?

(1) Benefits of detention are mostly homogeneous.

- why not?

(2) Potential offenders do not change behavior in response to the algorithm.

- why not?

(3) Judge follows a threshold rule.

- why not?

When are recividism risk scores sufficient for bail/release decisions?

(1) Benefits of detention are mostly homogeneous.

- why not?

(2) Potential offenders do not change behavior in response to the algorithm.

- why not?

(3) Judge follows a threshold rule.

- why not?

(4) Judges get feedback on prediction accuracy to assess domain shift.

- why not?
- **important distinction: if decision is about inspecting, versus jailing/treating/etc**

Alternative: Doctor's testing decision

Mullainathan and Obermeyer (2019)

- Consider the problem of a doctor deciding whether to order a test for a heart blockage.
  - if blockage is detected, useful treatment can be given
  - if no blockage, then test was wasted (test is costly to administer)

Alternative: Doctor's testing decision

Mullainathan and Obermeyer (2019)

- Consider the problem of a doctor deciding whether to order a test for a heart blockage.
  - if blockage is detected, useful treatment can be given
  - if no blockage, then test was wasted (test is costly to administer)
- Optimal testing strategy:
  - form predicted prior probability of a positive test $\hat{Y}(X_i)$
  - test all i with predicted prior probability above some threshold $\bar{Y}$.

When are test result priors sufficient for testing decisions?

(1) Benefits of ________ are mostly homogeneous.

- why not?

When are test result priors sufficient for testing decisions?

(1) Benefits of ________ are mostly homogeneous.

- why not?

(2) ________ do not change behavior in response to the algorithm.

- why not?

When are test result priors sufficient for testing decisions?

(1) Benefits of ________ are mostly homogeneous.

- why not?

(2) _________ do not change behavior in response to the algorithm.

- why not?

(3) ______ follows a threshold rule.

- why not?

When are test result priors sufficient for testing decisions?

(1) Benefits of ________ are mostly homogeneous.

- why not?

(2) _________ do not change behavior in response to the algorithm.

- why not?

(3) ______ follows a threshold rule.

- why not?

Note: Given (1) through (3), the doctor testing decision is a **prediction problem**.

Generalizing these points

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?
  - $D^* = \max_D u(D, Y, X)$

Generalizing these points

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?
  - $D^* = \max_D u(D, Y, X)$

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
   - $u(D, Y, X) = u(D, Y)$, and hence $D^*(Y, X) = D^*(Y)$

Generalizing these points

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?
  - $D^* = \max_D u(D, Y, X)$

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
   - $u(D, Y, X) = u(D, Y)$, and hence $D^*(Y, X) = D^*(Y)$
2. Environment factors (i.e. decision subjects) do not respond to the algorithm.
   - X is not a function of $D^*(\cdot)$

Generalizing these points

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?
  - $D^* = \max_D u(D, Y, X)$

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
   - $u(D, Y, X) = u(D, Y)$, and hence $D^*(Y, X) = D^*(Y)$
2. Environment factors (i.e. decision subjects) do not respond to the algorithm.
   - X is not a function of $D^*(\cdot)$
3. Each decision-maker j follows the algorithm threshold rule.
   - $D(X, \hat{Y}, j) = D^*(\hat{Y})$

Practice Quiz, Weeks 2-8

Outline

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
2. **Environment factors (i.e. decision subjects) do not respond to the algorithm.**
3. **Decision-makers follow the algorithm threshold rule.**

Outline

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
2. **Environment factors (i.e. decision subjects) do not respond to the algorithm.**
3. Decision-makers follow th algorithm threshold rule

Incentive Responses to Decision Systems

A policy implemented today $D_t(\cdot)$ could change features tomorrow X_{t+1}.

Incentive Responses to Decision Systems

A policy implemented today $D_t(\cdot)$ could change features tomorrow X_{t+1}.

- Take the case of ML-based credit scoring:
  - Some strategic responses are benign/helpful – e.g., pay back existing debts to improve scores

Incentive Responses to Decision Systems

A policy implemented today $D_t(\cdot)$ could change features tomorrow X_{t+1}.

- Take the case of ML-based credit scoring:
  - Some strategic responses are benign/helpful – e.g., pay back existing debts to improve scores
  - Other responses could be costly manipulation – e.g., open more credit accounts to increase credit score, which increase default risk.

Incentive Responses to Decision Systems

A policy implemented today $D_t(\cdot)$ could change features tomorrow X_{t+1}.

- Take the case of ML-based credit scoring:
  - Some strategic responses are benign/helpful – e.g., pay back existing debts to improve scores
  - Other responses could be costly manipulation – e.g., open more credit accounts to increase credit score, which increase default risk.
- More generally:
  - ML subjects can pay some cost and manipulate their features to improve their predicted label.

Milli et al, “The Social Cost of Strategic Classification” (2019)

Model sequential decision of modeler (“institution”) and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0, 1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)
- Subject gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$:

$$u(X'; X) = \hat{y}(X') - c(X, X')$$

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)
- Subject gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$:

$$u(X'; X) = \hat{y}(X') - c(X, X')$$

- The subject reports

$$\Delta(X) = \arg\max_{X'} u(X'; X)$$

- and therefore the institution's objective is

$$\max_{\hat{y}(\cdot)} \Pr(\hat{y}(\Delta(X)) = Y).$$

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)
- Subject gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$:

$$u(X'; X) = \hat{y}(X') - c(X, X')$$

- The subject reports

$$\Delta(X) = \arg\max_{X'} u(X'; X)$$

- and therefore the institution's objective is

$$\max_{\hat{y}(\cdot)} \Pr(\hat{y}(\Delta(X)) = Y).$$

- Equilibrium:
  - features x_j that are costly to change (high $\frac{\partial c}{\partial x_j}$) will be used by the designer. features that are less costly to change will not be used.
  - in strategic context, designer chooses overall more conservative decision threshold.

Milli et al, "The Social Cost of Strategic Classification" (2019)

Model sequential decision of modeler ("institution") and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

- Each subject has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $\hat{y} : X \to Y$ equal to $V = \Pr(\hat{y}(X) = Y)$.
  - (implicitly treats classification as equal to decision: $D(\hat{Y}) = \hat{Y}$)
- Subject gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$:

$$u(X'; X) = \hat{y}(X') - c(X, X')$$

- The subject reports

$$\Delta(X) = \arg\max_{X'} u(X'; X)$$

- and therefore the institution's objective is

$$\max_{\hat{y}(\cdot)} \Pr(\hat{y}(\Delta(X)) = Y).$$

- Equilibrium:
  - features x_j that are costly to change (high $\frac{\partial c}{\partial x_j}$) will be used by the designer. features that are less costly to change will not be used.
  - in strategic context, designer chooses overall more conservative decision threshold.
- The costs $c(\cdot)$ are socially wasteful, but responses to manipulation increase them.
  - $c(\cdot)$ could be different across groups, causing inequity

Outline

Behavioral Responses to Algorithms

Responses by Subjects

Responses by Decision-Makers

Selective Labeling

Further Discussion: Using Machine Learning to Guide Audit Policy

- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
2. Environment factors (i.e. decision subjects) do not respond to the algorithm.
3. **Decision-makers respond predictably to the algorithm.**

Decision-makers are usually separate from the algorithm

- So far we have treated the decision D as a deterministic function of $\hat{Y}$: $D = 1$ if $\hat{Y} > \bar{Y}$, $D = 0$ otherwise.
  - means that $\frac{\partial D}{\partial x_j} = 0, \forall j$: decisions are not sensitive to case characteristics, after conditioning on $\hat{Y}$.

Decision-makers are usually separate from the algorithm

- So far we have treated the decision D as a deterministic function of $\hat{Y}$: $D = 1$ if $\hat{Y} > \bar{Y}$, $D = 0$ otherwise.
  - means that $\frac{\partial D}{\partial x_j} = 0, \forall j$: decisions are not sensitive to case characteristics, after conditioning on $\hat{Y}$.
- But there could be many reasons that this assumption does not hold, e.g.:
  - judges caring about whether a defendant has children or not.
  - tax/fraud auditors not wanting to audit their friends / family members
  - doctor wanting to save people with more years of life left / not terminally ill

Decision-makers are usually separate from the algorithm

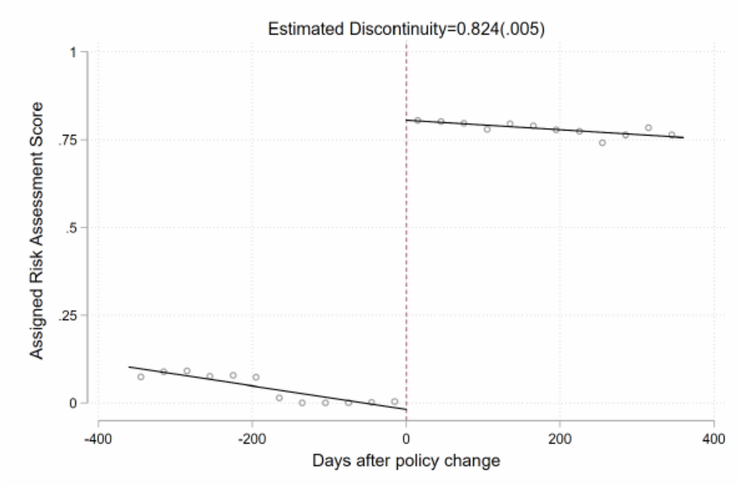
- So far we have treated the decision D as a deterministic function of $\hat{Y}$: $D = 1$ if $\hat{Y} > \bar{Y}$, $D = 0$ otherwise.
  - means that $\frac{\partial D}{\partial x_j} = 0, \forall j$: decisions are not sensitive to case characteristics, after conditioning on $\hat{Y}$.
- But there could be many reasons that this assumption does not hold, e.g.:
  - judges caring about whether a defendant has children or not.
  - tax/fraud auditors not wanting to audit their friends / family members
  - doctor wanting to save people with more years of life left / not terminally ill

$\rightarrow$ empirical evidence is needed on how decision-makers respond to algorithms.

First Stage: Discrete Reform introducing risk scoring

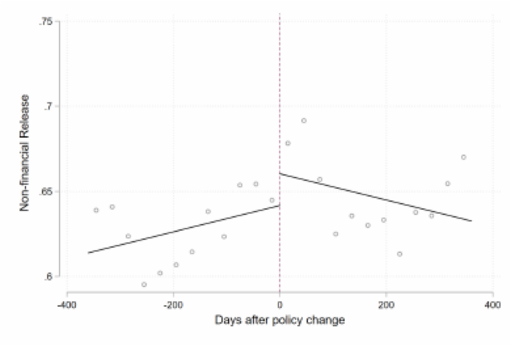
Sloan et al 2018

Figure 4: Regression Discontinuity Results for the Probability of Receiving a Risk Assessment Score

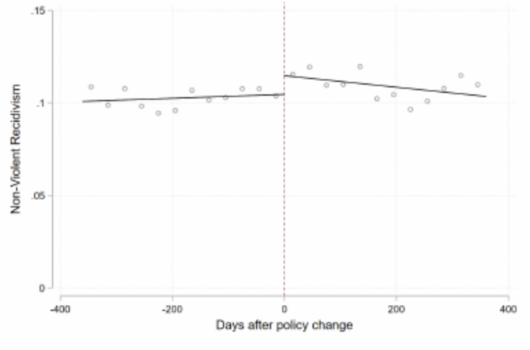


Risk scoring increases release rates and recidivism

Sloan et al 2018



(a) Non-financial Bond



(a) Probability of Non-Violent Recidivism

- In response to risk scoring, judges release more poor defendants.

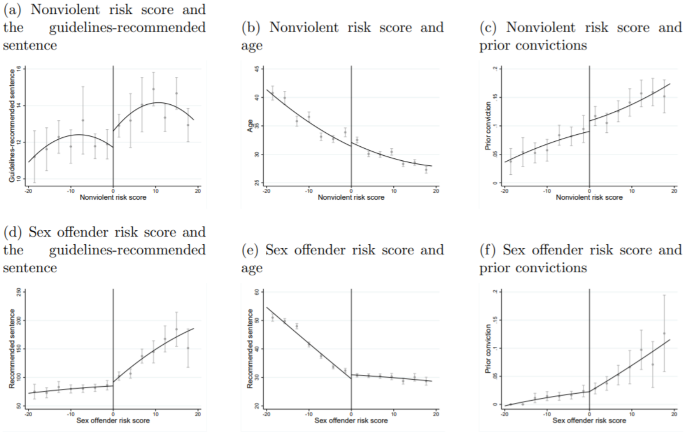
Stevenson and Doleac: Method

- RD using a continuous risk score – above a discrete cutoff, defendant is labeled "risky".

Stevenson and Doleac: Method

- RD using a continuous risk score – above a discrete cutoff, defendant is labeled "risky".
- Identification check: Other predetermined characteristics are flat around the cutoff (covariate balance):

Figure 2: Covariate balance across risk score cutoffs



(a) Nonviolent risk score and the guidelines-recommended sentence

(b) Nonviolent risk score and age

(c) Nonviolent risk score and prior convictions

(d) Sex offender risk score and the guidelines-recommended sentence

(e) Sex offender risk score and age

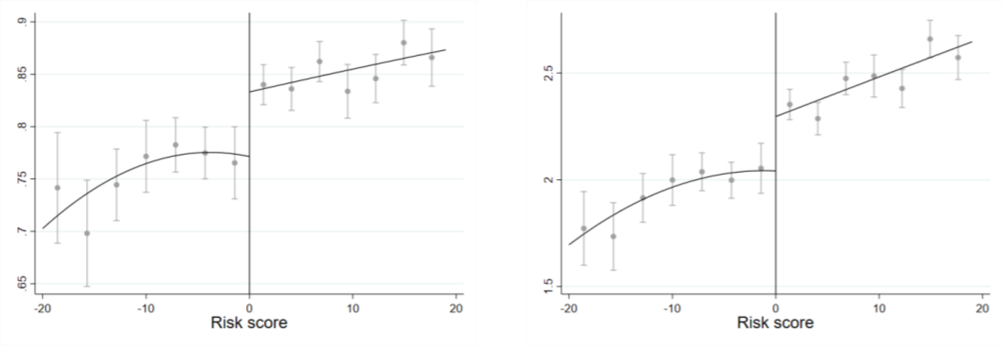
(f) Sex offender risk score and prior convictions

Stevenson and Doleac: Result (RDD)

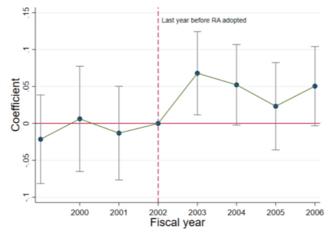
Figure 3: Does the risk classification affect defendants' sentences at the margin?

(a) Probability of incarceration

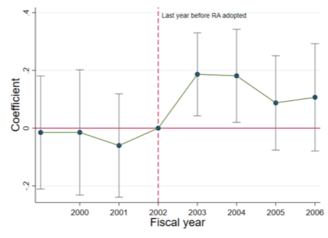
(b) The sentence length



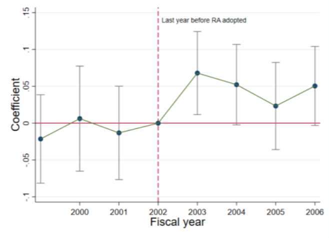
(c) Predicted risk score event study (outcome = pr(incarceration))



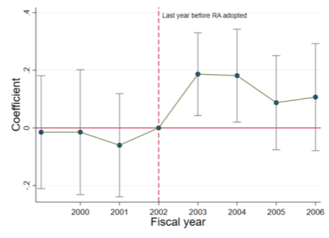
(d) Predicted risk score event-study (outcome = sentence length)



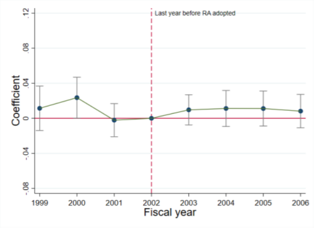
(c) Predicted risk score event study (outcome = pr(incarceration))



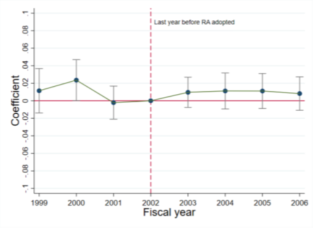
(d) Predicted risk score event-study (outcome = sentence length)



(a) Risk assessment's impact on pr(incarceration)



(b) Risk assessment's impact on sentence length (arcsinh)



"...despite explicit instructions that risk assessment was supposed to lower prison populations, there was no net reduction in incarceration. Nor do we detect any public safety benefits from its use..."

Outline

Checking for Domain Shift

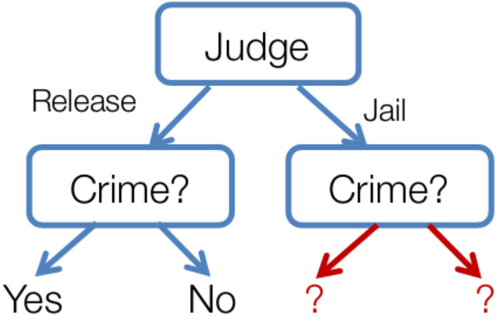
- Under what conditions are predictions $\hat{Y}(X)$ sufficient for making the optimal decision D^*?

1. Payoff of the decision does not depend on other factors besides $\hat{Y}$
2. Environment factors (i.e. decision subjects) do not respond to the algorithm.
3. Decision-makers respond predictably to the algorithm.
4. **Decision-maker gets continuous feedback on model accuracy.**

- What if decision-maker does not get this feedback?

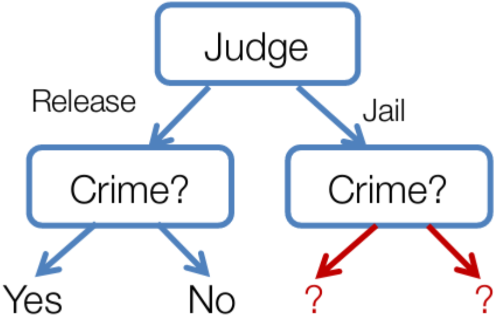
Bail decision: Judge is ***selectively labeling*** the dataset

Bail decision: Judge is ***selectively labeling*** the dataset



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

Bail decision: Judge is ***selectively labeling*** the dataset

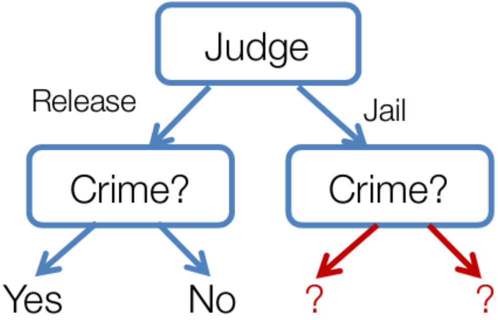


Selective labels introduce bias. Example:

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

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

Bail decision: 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 from Kleinberg et al: Contraction

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

Solution from Kleinberg et al: 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.
- **Why does this approach require random assignment of cases to judges to work?**

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

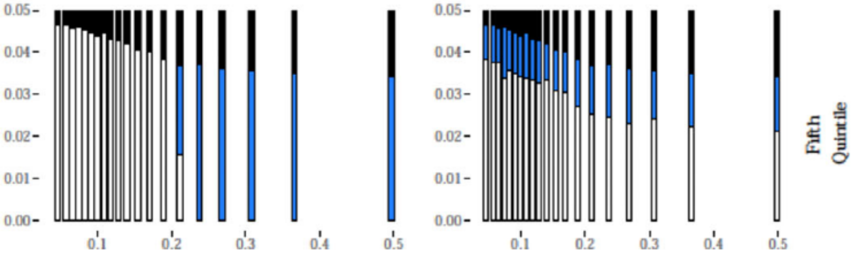


FIGURE VI

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

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.
  - Kleinberg et al could fix it because of random assignment of judges. But usually that is not possible either.

Outline

Behavioral Responses to Algorithms

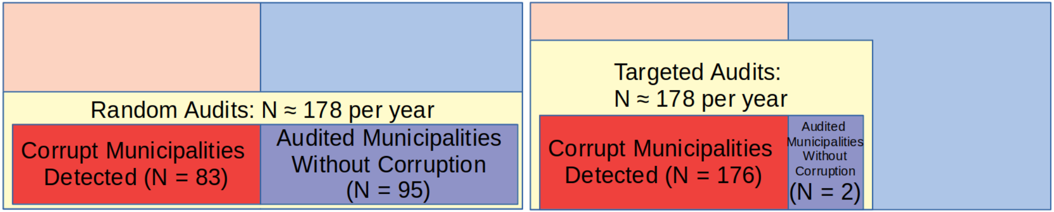
Responses by Subjects

Responses by Decision-Makers

Selective Labeling

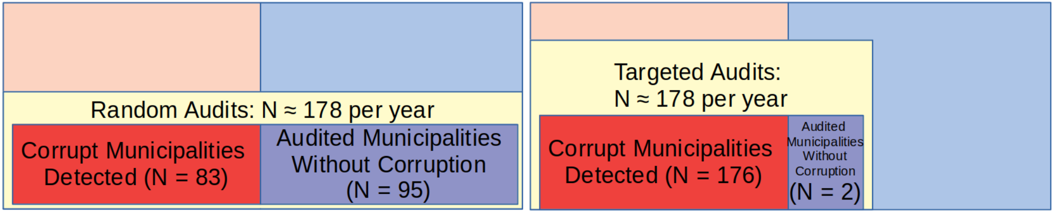
Further Discussion: Using Machine Learning to Guide Audit Policy

Comparison: Brazil Corruption Audits



- Holding number of audits constant, targeting increases detections by 120%.
- Detection probability per corrupt municipality more than doubles – from 2.9% to 6.7%.

Comparison: Brazil Corruption Audits



- Holding number of audits constant, targeting increases detections by 120%.
- Detection probability per corrupt municipality more than doubles – from 2.9% to 6.7%.
- To achieve same number of detections as status quo (83 municipalities), only 84 targeted audits are needed.
  - Decrease of 94 audits per year (53%), a major reduction in audit resources.
- ***Why don't we need to use the contraction method a la Kleinberg et al 2018?***

Incentive Effects of Targeted Audits

- Remember that one of our criteria for ML-powered decision-making is that decision subjects don't respond to the algorithm.
- But **in the case of detecting corruption, this is exactly what we want**:
  - corruption makes audits more likely $\rightarrow$ reduces incentives and probability of corruption!

Mechanism Design Issues

- With repeated audits, there could be behavioral responses by local officials.
  - could produce significant errors favoring savvy mayors.
  - Would still deter corrupt fiscal actions that are not easily substitutable.

How much information to publicize about audit targeting?

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is “the industry approach”, e.g., for how google/facebook detect violations.

How much information to publicize about audit targeting?

Option 1: Give **full information** about the policy and the associated model weights.

- Would increase deterrence against corruption actions captured by the model, that are not substitutable.
- But would make gaming the system easier.

Option 2: Give **no information** about how targeting is done.

- This is "the industry approach", e.g., for how google/facebook detect violations.
- mayors might learn how algorithm works over time.
- weights could be updated in response to behavioral responses

Mixing random and targeted audits

- Random audits could be maintained (along with targeted audits).
  - Preserves some deterrence incentive for all municipalities.
  - Results of random audits could be used to update algorithm parameters.

Required Reading for Next Week

"How a Discriminatory Algorithm Wrongly Accused Thousands of Families of Fraud"
(*Vice* article, linked on syllabus)

Video Presentation: Bjorkegren et al, Manipulation-Proof Machine Learning