Building a Robot Judge: Data Science for Decision-Making

10. Algorithms and Decisions I

Weekly Q&A

https://padlet.com/eash44/rdfswr8dzcxhka0z

Outline

Summary

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.
  - when we say "bias" or "endogeneity", that is talking about internal validity

Summary

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.
  - when we say "bias" or "endogeneity", that is talking about internal validity
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings.
  - this is usually much more speculative.

Internal Validity (from week 3)

Linear regression model:

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

- Exogeneity assumption: $\mathrm{Cov}[s_i, \epsilon_i] = 0$
  - no omitted variable bias (unobserved confounders), no joint causality.

Internal Validity (from week 3)

Linear regression model:

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

- Exogeneity assumption: $\text{Cov}[s_i, \epsilon_i] = 0$
  - no omitted variable bias (unobserved confounders), no joint causality.
  - $\rightarrow$ then OLS estimates for $\hat{\beta}$ converge to β in large samples.

Internal Validity (from week 3)

Linear regression model:

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

- Exogeneity assumption: $\mathrm{Cov}[s_i, \epsilon_i] = 0$
  - no omitted variable bias (unobserved confounders), no joint causality.
  - $\rightarrow$ then OLS estimates for $\hat{\beta}$ converge to β in large samples.
- Standard errors are correct.
  - accounting for heteroskedasticity (use the "robust" option)
  - accounting for serial correlation (clustering at level of treatment)

Internal Validity (from week 3)

Linear regression model:

$$Y_i = \alpha + \beta s_i + \epsilon_i$$

- Exogeneity assumption: $\text{Cov}[s_i, \epsilon_i] = 0$
  - no omitted variable bias (unobserved confounders), no joint causality.
  - $\rightarrow$ then OLS estimates for $\hat{\beta}$ converge to β in large samples.
- Standard errors are correct.
  - accounting for heteroskedasticity (use the "robust" option)
  - accounting for serial correlation (clustering at level of treatment)

Under these conditions, causal inferences (statistical estimates on treatment effects) are valid for the population studied.

Internal validity (machine learning)

- In machine learning, we would gauge "internal validity" by proper train/test splits, and avoidance of data leakage.
  - $\rightarrow$ then performance metrics are valid to that dataset, or othe samples from the same data generating process.

External Validity: Does it generalize?

- If internal validity is satisfied, then ML metrics and causal inferences are valid for the population studied.
- External validity is about whether the results generalize to other populations.
  - in machine learning, this is often called "domain shift"

External Validity: Does it generalize?

- If internal validity is satisfied, then ML metrics and causal inferences are valid for the population studied.
- External validity is about whether the results generalize to other populations.
  - in machine learning, this is often called "domain shift"
- Examples:
  - say we have a study of UZH students for the effect of coffee on productivity.
    - Does that generalize to ETH students?

External Validity: Does it generalize?

- If internal validity is satisfied, then ML metrics and causal inferences are valid for the population studied.
- External validity is about whether the results generalize to other populations.
  - in machine learning, this is often called "domain shift"
- Examples:
  - say we have a study of UZH students for the effect of coffee on productivity.
    - Does that generalize to ETH students?
  - medical trials are often run with men, but medicines are then used to treat both men and women.

External Validity: Does it generalize?

- If internal validity is satisfied, then ML metrics and causal inferences are valid for the population studied.
- External validity is about whether the results generalize to other populations.
  - in machine learning, this is often called “domain shift”
- Examples:
  - say we have a study of UZH students for the effect of coffee on productivity.
    - Does that generalize to ETH students?
  - medical trials are often run with men, but medicines are then used to treat both men and women.
  - recividism risk prediction model trained in 2020, is it valid for 2021?

External Validity: Does it generalize?

- If internal validity is satisfied, then ML metrics and causal inferences are valid for the population studied.
- External validity is about whether the results generalize to other populations.
  - in machine learning, this is often called "domain shift"
- Examples:
  - say we have a study of UZH students for the effect of coffee on productivity.
    - Does that generalize to ETH students?
  - medical trials are often run with men, but medicines are then used to treat both men and women.
  - recividism risk prediction model trained in 2020, is it valid for 2021?
- In general **estimates/metrics are not valid for other populations**.
  - other populations are different. so treatment effects and predictions might be different.

Practice Quiz, Weeks 2-9

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.

Prediction vs Judgment

- **Prediction** is about guessing the state of the world
  - parameters θ from $\hat{Y}(X; \theta)$.
- **Judgment** is about knowing the utility or benefit function
  - parameters β from $W(X, Y; \beta)$.

A simple formalization (Agrawal et al 2019)

- When a defendant is accused of a crime, should he/she be kept in jail while awaiting trial?

A simple formalization (Agrawal et al 2019)

- When a defendant is accused of a crime, should he/she be kept in jail while awaiting trial?

There are two actions:

- jail defendant (safe)
  - always generates payoff S.

A simple formalization (Agrawal et al 2019)

- When a defendant is accused of a crime, should he/she be kept in jail while awaiting trial?

There are two actions:

- jail defendant (safe)
  - always generates payoff S.
- release defendant (risky)
  - generates payoff R for a good defendant, r for a bad defendant.

$$r < S < R$$

A simple formalization (Agrawal et al 2019)

- When a defendant is accused of a crime, should he/she be kept in jail while awaiting trial?

There are two actions:

- jail defendant (safe)
  - always generates payoff S.
- release defendant (risky)
  - generates payoff R for a good defendant, r for a bad defendant.

$$r < S < R$$

- Assume probability of bad defendant is $\bar{Y}$.
  - Then choice is "always jail" if

$$\bar{Y}R + (1 - \bar{Y})r < S$$

A simple formalization (Agrawal et al 2019)

- When a defendant is accused of a crime, should he/she be kept in jail while awaiting trial?

There are two actions:

- jail defendant (safe)
  - always generates payoff S.
- release defendant (risky)
  - generates payoff R for a good defendant, r for a bad defendant.

$$r < S < R$$

- Assume probability of bad defendant is $\bar{Y}$.
  - Then choice is "always jail" if

$$\bar{Y}R + (1 - \bar{Y})r < S$$

- Suppose there is a prediction technology where the decision-maker observes $\hat{Y}(X) \in [0,1]$.
  - choice function becomes

$$\hat{Y}R + (1 - \hat{Y})r < S$$

Example: Allocating fire/health 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 fire/health 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%).

Under what conditions are predictions sufficient for optimal allocation of inspectors?

Example: Allocating fire/health 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%).

Under what conditions are predictions sufficient for optimal allocation of inspectors?

1. **Benefits of fixing problems are mostly homogeneous.**
2. **Establishments do not change behavior in response to the algorithm.**
3. **Inspectors respond predictably to the algorithm.**

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(1) Benefits 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 fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(1) Benefits 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.
- restaurants with high health risk might not have many customers $\rightarrow$ could be better to inspect the more popular restaurants.

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(2) Establishments do not change behavior in response to the algorithm.

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(2) Establishments do not change behavior in response to the algorithm.

- inspections are an incentive mechanism – after implementing an algorithm, firms will respond to that.

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(2) Establishments do not change behavior in response to the algorithm.

- inspections are an incentive mechanism – after implementing an algorithm, firms will respond to that.
- This could be a good thing, if firms reduce health code violations.

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(2) Establishments do not change behavior in response to the algorithm.

- inspections are an incentive mechanism – after implementing an algorithm, firms will respond to that.
- This could be a good thing, if firms reduce health code violations.
    - but it is also a **domain shift** $\rightarrow$ predictions using pre-reform data are no longer externally valid.

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(2) Establishments do not change behavior in response to the algorithm.

- inspections are an incentive mechanism – after implementing an algorithm, firms will respond to that.
- This could be a good thing, if firms reduce health code violations.
  - but it is also a **domain shift** $\rightarrow$ predictions using pre-reform data are no longer externally valid.
- Responses could be heterogeneous:
  - some firms may be more sensitive to penalties than others,
  - it 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 fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(3) Inspectors respond predictably to the algorithm.

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(3) Inspectors respond predictably to the algorithm.

- What if inspectors ignore the algorithm?
  - e.g., they see a few errors and then go back to following their own judgment.

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(3) Inspectors respond predictably to the algorithm.

- What if inspectors ignore the algorithm?
  - e.g., they see a few errors and then go back to following their own judgment.
- What if inspectors rely too heavily on the algorithm?
  - e.g., they ignore some obvious special circumstances or variables that aren't in the dataset (e.g. a building being next door to a fire house; a restaurant serving only pre-packaged food).

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(1) Benefits of fixing problems are mostly homogeneous.
(2) Establishments do not change behavior in response to the algorithm.
(3) Inspectors respond predictably to the algorithm.

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(1) Benefits of fixing problems are mostly homogeneous.
(2) Establishments do not change behavior in response to the algorithm.
(3) Inspectors respond predictably to the algorithm.
(4) Others?

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(1) Benefits of fixing problems are mostly homogeneous.
(2) Establishments do not change behavior in response to the algorithm.
(3) Inspectors respond predictably to the algorithm.
(4) Others?

- We will come back to all of these in more detail.
- Main Lesson: inspection policy is not just a machine prediction problem
  - it is also a causal inference problem.

Example: Allocating fire/health inspectors

Under what conditions are predictions sufficient for optimal allocation of inspectors?
(1) Benefits of fixing problems are mostly homogeneous.
(2) Establishments do not change behavior in response to the algorithm.
(3) Inspectors respond predictably to the algorithm.
(4) Others?

- We will come back to all of these in more detail.
- Main Lesson: inspection policy is not just a machine prediction problem
  - it is also a causal inference problem.
- Framed differently: What is the expected improvement in overall quality of units (e.g., fire damage, food poisoning rates) in the city under a new AI-powered inspector allocation regime?

Example 2: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the net return on search-engine ad spending based on clicks = estimated at 1400%.

Example 2: eBay advertising

Athey 2017; Blake et al 2015

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

Example 2: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the net return on search-engine ad spending based on clicks = 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 2: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the net return on search-engine ad spending based on clicks = 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 net return on search-engine ad spending = -63%!
  - eBay stopped buying so much advertising after that.

Example 2: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the net return on search-engine ad spending based on clicks = 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 net return on search-engine ad spending = -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.

Example 2: eBay advertising

Athey 2017; Blake et al 2015

- Historically, eBay measured advertising effectiveness with correlational model:
  - clicks were used to predict sales
  - the net return on search-engine ad spending based on clicks = 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 net return on search-engine ad spending = -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.
  - again: AI-suppported decision-making is both a machine learning and causal inference problem.

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 Release

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

Bail Decision: Detain or Release

- 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 Release

- 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 next 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 next 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 next 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

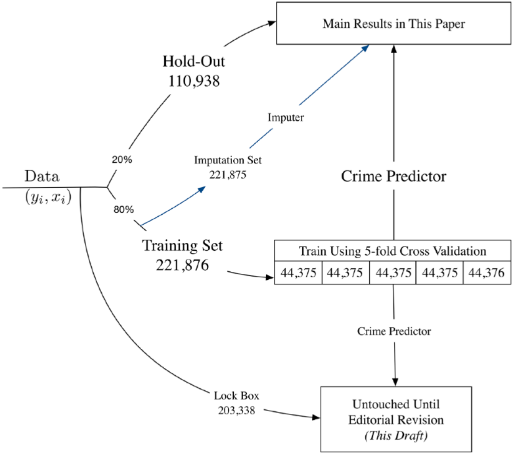


FIGURE I

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

Data: Defendant Features

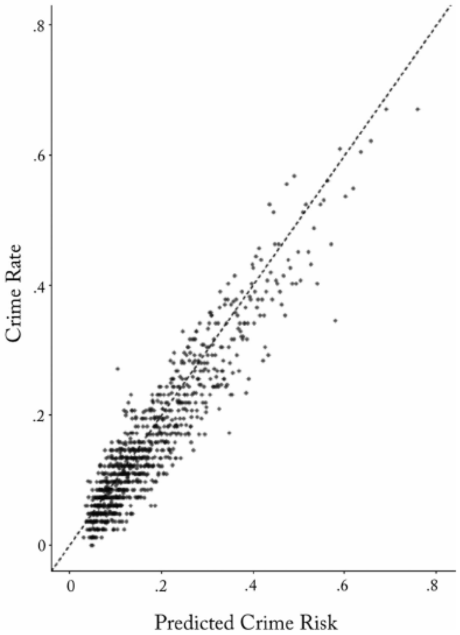
Kleinberg et al (2019)

`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`

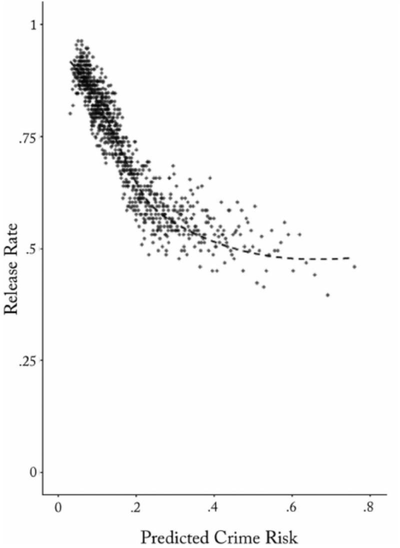
- excludes race, gender, and religion
  - not legal to include – will come back to this issue

Model Performance

- 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



What human judges do

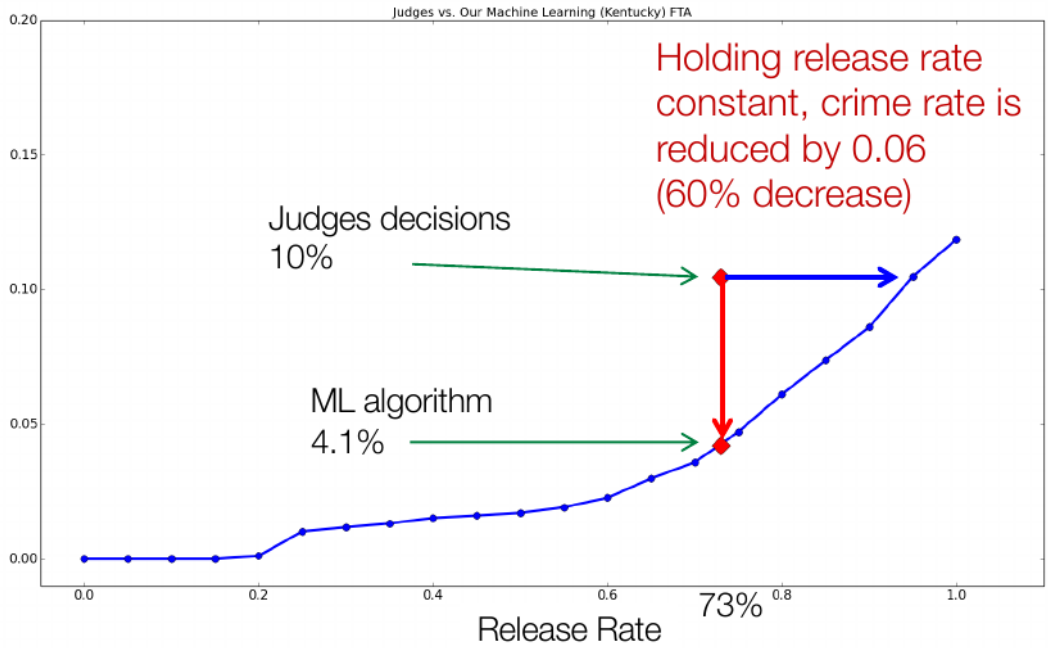


- Human judges tend to follow what algorithm suggests.
- But judge sees factors the machine does not
  - makes decisions based on $\Pr(Y|X_H)$
    - X_H includes other factors not seen by the machine – e.g., defendant demeanor.
  - Machine makes decisions based on $\Pr(Y|X)$, $X \subset X_H$.

Prediction $\rightarrow$ Release Rule

- Kleinberg et al consider the following release rule based on recidivism predictions:
  - For every defendant predict $\hat{Y}(X_i)$, probability of recidivism.
  - Sort by increasing $\hat{Y}(X_i)$
  - Release bottom N defendants, jail the rest.
- Kleinberg et al (2018) use this rule to analyze the tradeoff between fraction released and crime rate.

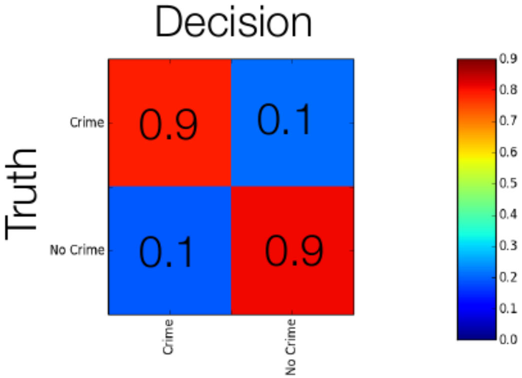
Compare Judge to ML in predicted crime rate

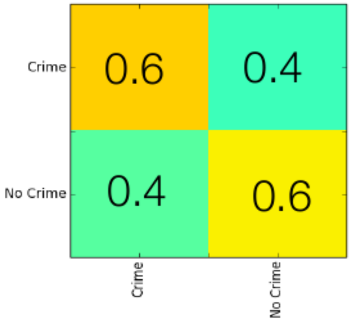


Analyzing judge mistakes

Analyzing judge mistakes

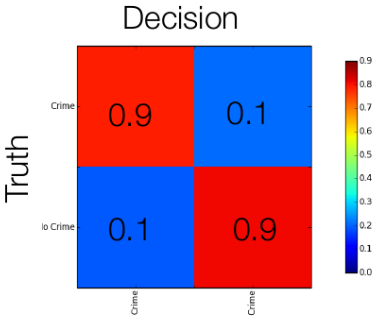
Experienced Judges	Inexperienced Judges

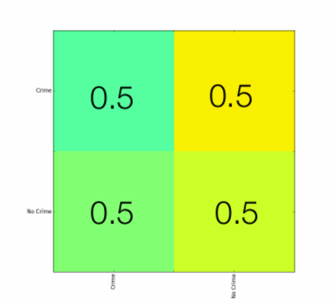




Source: Jure Leskovec slides.

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.

Activity

- **Rewrite the following statements about building inspectors, for the case of judges deciding on bail. For each requirement, give an example of when it won't hold.**
  Under what conditions are predictions sufficient for optimal allocation of inspectors?
  (1) Benefits of fixing problems are mostly homogeneous.
  (2) Establishments do not change behavior in response to the algorithm.
  (3) Inspectors respond predictably to the algorithm.
- **In 5 minutes, paste your answer into a padlet post here:** `https://padlet.com/eash44/u4rfjxc8mbgjd587`