Studying Social Inequality with Data Science

INFO 3370 / 5371
Spring 2024

Predicting life outcomes

Results of the PSID Income Prediction Challenge

Learning goals for today

By the end of class, you will be able to

- know who had the best predictions!
- reason about predictability of life outcomes

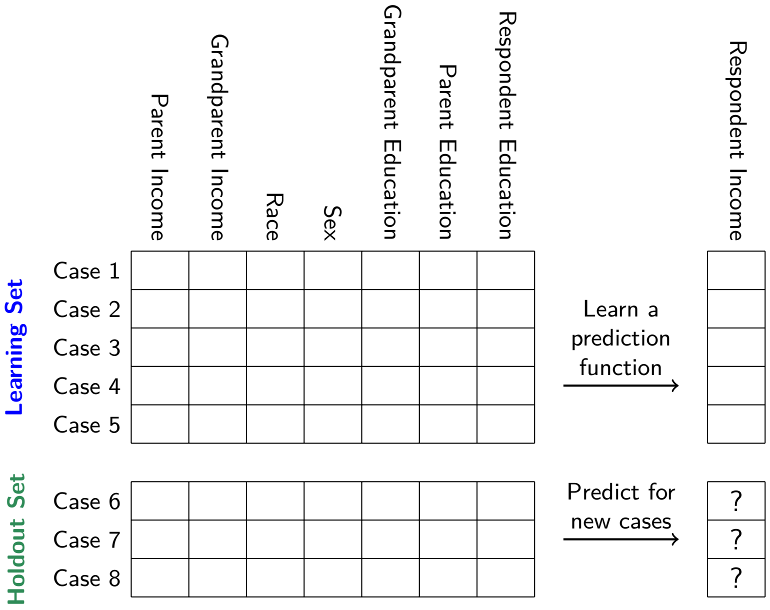
Equality Opportunity and Prediction

Possible claim

To the degree that we can predict life outcomes,

people do not have equal opportunity

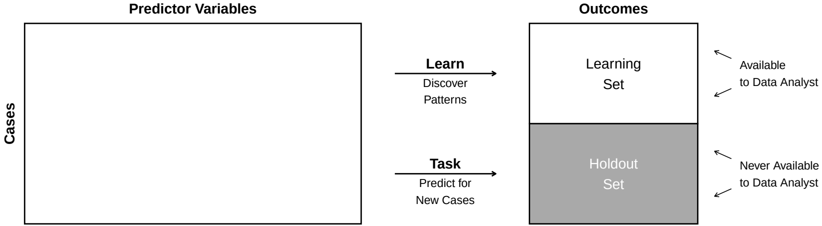
Equality Opportunity and Prediction



The model selection problem

In supervised machine learning, the goal is to

- learn patterns in the available data
- predict outcomes for previously unseen cases

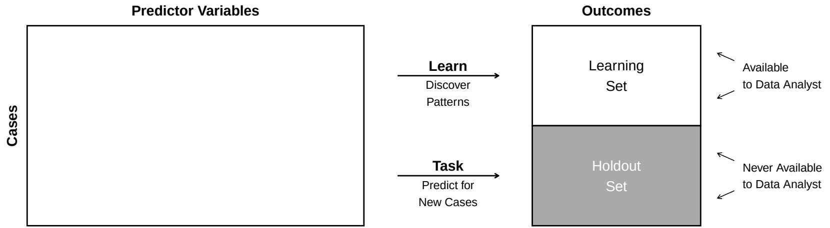


The model selection problem

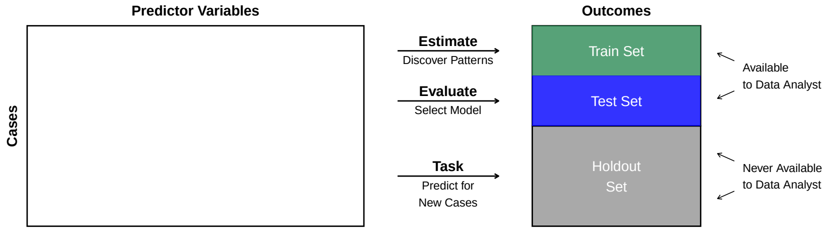
When a task involves unseen data,

mimic the task with data we have

The model selection problem



The model selection problem



Prepare environment

```r
library(tidyverse)
library(rsample)
set.seed(14850)
```

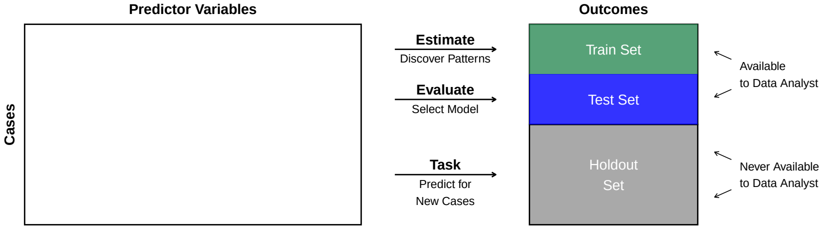
Load data

```r
learning <- read_csv("learning.csv")
holdout_public <- read_csv("holdout_public.csv")
```

Create a train-test split within `learning`

Using the `rsample` package,

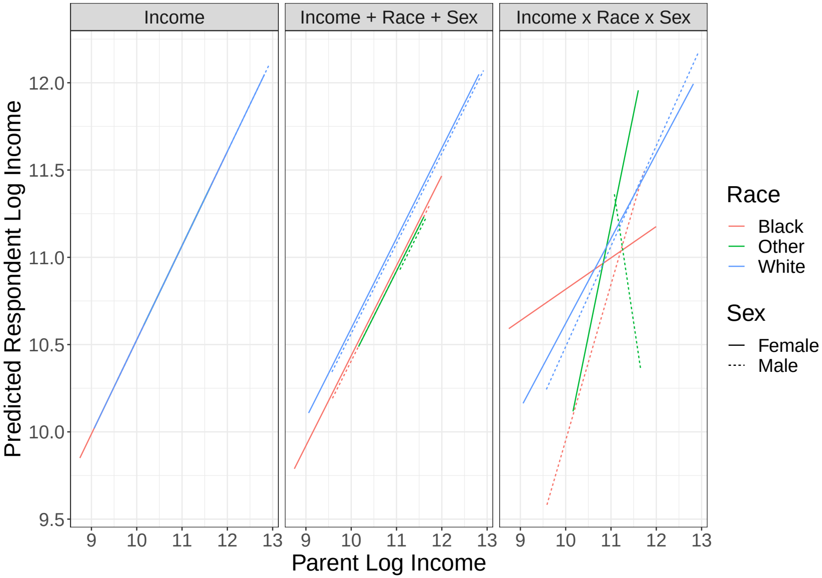
```
split <- learning |>
  initial_split(prop = 0.5)
```


Predictor Variables
Cases
Outcomes
Estimate
Discover Patterns
Evaluate
Select Model
Task
Predict for
New Cases
Train Set
Test Set
Holdout
Set
Available
to Data Analyst
Never Available
to Data Analyst

Learn candidates in the train set

```r
candidate_1 <- lm(
  g3_log_income ~ g2_log_income,
  data = training(split)
)
candidate_2 <- lm(
  g3_log_income ~ g2_log_income + race + sex,
  data = training(split)
)
candidate_3 <- lm(
  g3_log_income ~ g2_log_income * race * sex,
  data = training(split)
)
```

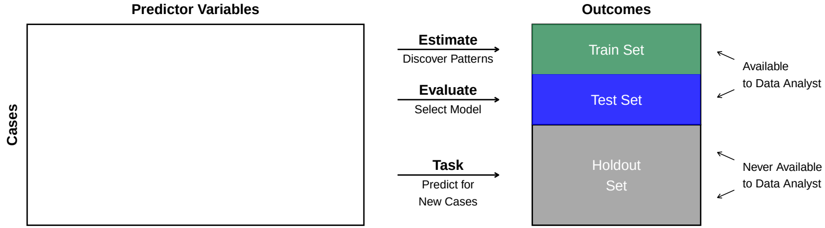
Learn candidates in the train set



Evaluate performance on the test set. Choose a model

```r
fitted |>
  group_by(model) |>
  mutate(error = g3_log_income - yhat) |>
  mutate(squared_error = error ^ 2) |>
  summarize(mse = mean(squared_error))
```

```
## # A tibble: 3 x 2
##   model         mse
##   <chr>       <dbl>
## 1 candidate_1 0.439
## 2 candidate_2 0.437
## 3 candidate_3 0.477
```


Predictor Variables
Cases
Outcomes
Estimate
Discover Patterns
Evaluate
Select Model
Task
Predict for
New Cases
Train Set
Test Set
Holdout
Set
Available
to Data Analyst
Never Available
to Data Analyst

Apply your chosen model

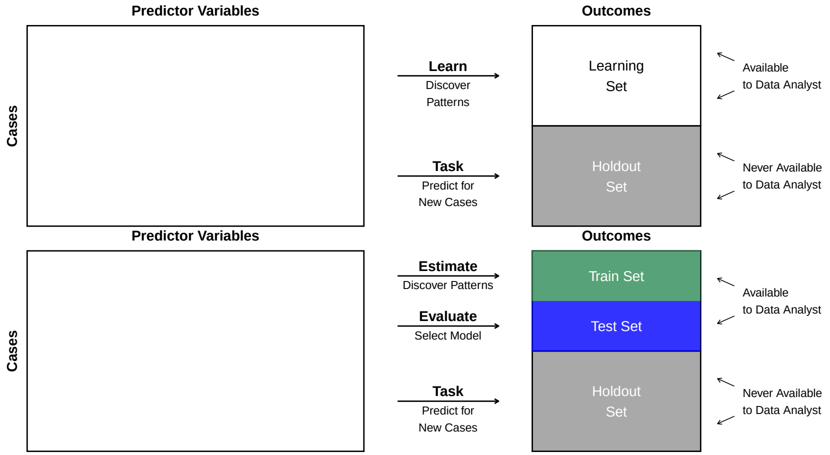
Learn in the full learning set

```
chosen <- lm(
  g3_log_income ~ g2_log_income +
    race + sex,
  data = learning
)
```

Predict for the holdout set

```
predicted <- holdout_public %>%
  mutate(
    predicted = predict(
      chosen,
      newdata = holdout_public
    )
  )
```

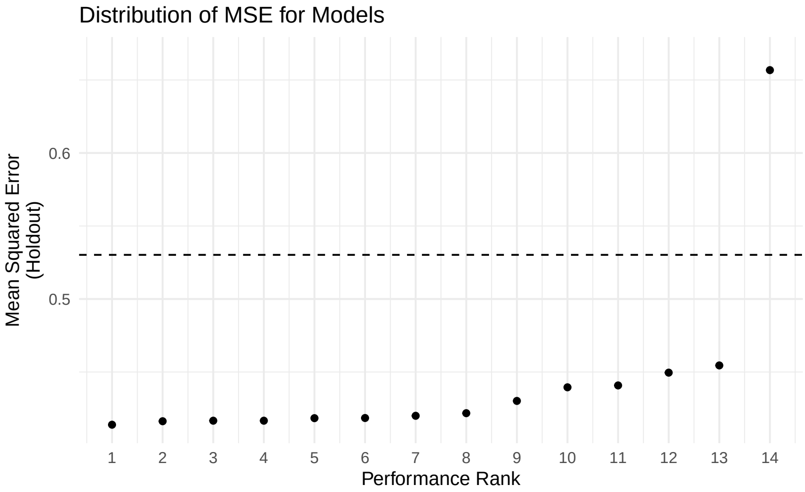
Summary



Predictor Variables
Outcomes
Cases
Learn
Discover
Patterns
Learning
Set
Available
to Data Analyst
Task
Predict for
New Cases
Holdout
Set
Never Available
to Data Analyst
Predictor Variables
Outcomes
Cases
Estimate
Discover Patterns
Train Set
Available
to Data Analyst
Evaluate
Select Model
Test Set
Task
Predict for
New Cases
Holdout
Set
Never Available
to Data Analyst

Your submissions

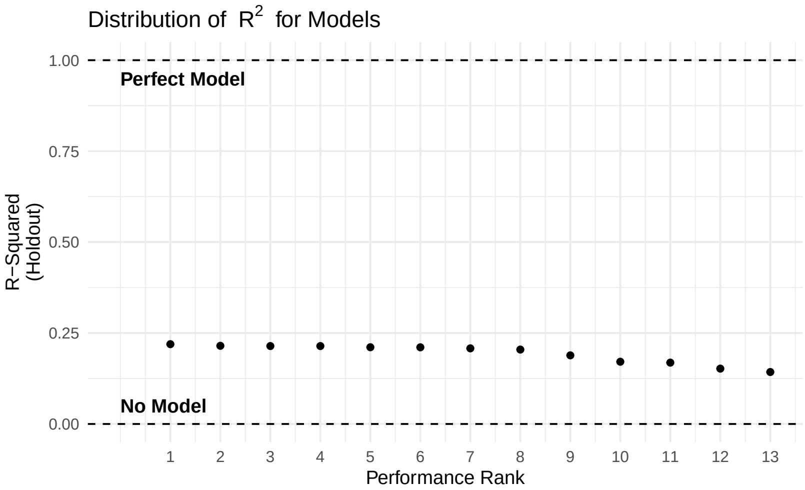
- 21 submissions
- 20 submissions predicting for all holdout cases
- 17 submissions with non-missing predictions
- 14 submissions by unique teams



Distribution of MSE for Models
Mean Squared Error
(Holdout)
0.6
0.5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Performance Rank

$$R^2 = 1 - \frac{\mathrm{MSE}_{\mathrm{Model}}}{\mathrm{MSE}_{\mathrm{No\ Model}}}$$

- score of $1 =$ perfect! $\mathrm{MSE}_{\mathrm{Model}} = 0$
- score of $0 =$ no better than no model at all



How would you make sense of this?

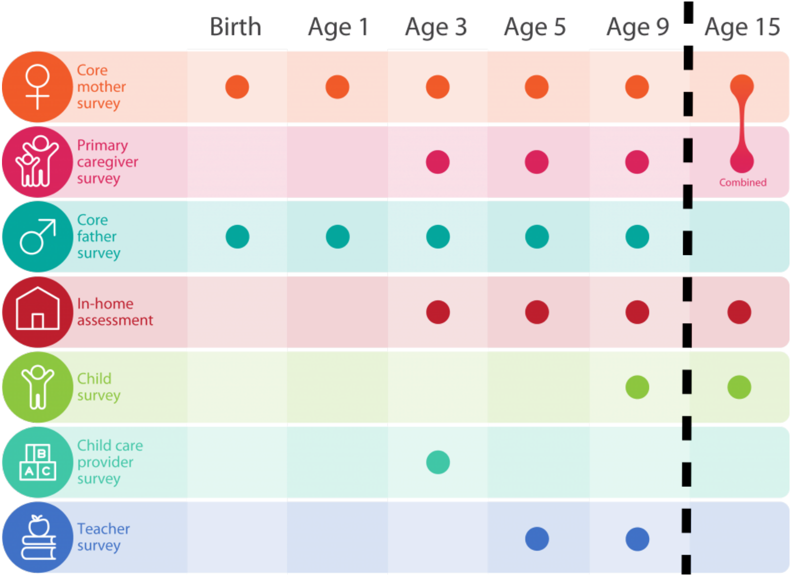
our exercise was a particular case

of a broader research project

Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew J. Salganik[a,1], Ian Lundberg[a], Alexander T. Kindel[a], Caitlin E. Ahearn[b], Khaled Al-Ghoneim[c], Abdullah Almaatouq[d,e], Drew M. Altschul[f], Jennie E. Brand[b,g], Nicole Bohme Carnegie[h], Ryan James Compton[i], Debanjan Datta[j], Thomas Davidson[k], Anna Filippova[l], Connor Gilroy[m], Brian J. Goode[n], Eaman Jahani[o], Ridhi Kashyap[p,q,r], Antje Kirchner[s], Stephen McKay[t], Allison C. Morgan[u], Alex Pentland[e], Kivan Polimis[v], Louis Raes[w], Daniel E. Rigobon[x], Claudia V. Roberts[y], Diana M. Stanescu[z], Yoshihiko Suhara[e], Adaner Usmani[aa], Erik H. Wang[z], Muna Adem[bb], Abdulla Alhajri[cc], Bedoor AlShebli[dd], Redwane Amin[ee], Ryan B. Amos[y], Lisa P. Argyle[ff], Livia Baer-Bositis[gg], Moritz Büchi[hh], Bo-Ryehn Chung[ii], William Eggert[jj], Gregory Faletto[kk], Zhilin Fan[ll], Jeremy Freese[gg], Tejomay Gadgil[mm], Josh Gagné[gg], Yue Gao[nn], Andrew Halpern-Manners[bb], Sonia P. Hashim[y], Sonia Hausen[gg], Guanhua He[oo], Kimberly Higuera[gg], Bernie Hogan[pp], Ilana M. Horwitz[qq], Lisa M. Hummel[gg], Naman Jain[x], Kun Jin[rr], David Jurgens[ss], Patrick Kaminski[bb,tt], Areg Karapetyan[uu,vv], E. H. Kim[gg], Ben Leizman[y], Naijia Liu[z], Malte Möser[y], Andrew E. Mack[z], Mayank Mahajan[y], Noah Mandell[ww], Helge Marahrens[bb], Diana Mercado-Garcia[qq], Viola Mocz[xx], Katariina Mueller-Gastell[gg], Ahmed Musse[yy], Qiankun Niu[ee], William Nowak[zz], Hamidreza Omidvar[aaa], Andrew Or[y], Karen Ouyang[y], Katy M. Pinto[bbb], Ethan Porter[ccc], Kristin E. Porter[ddd], Crystal Qian[y], Tamkinat Rauf[gg], Anahit Sargsyan[eee], Thomas Schaffner[y], Landon Schnabel[gg], Bryan Schonfeld[z], Ben Sender[fff], Jonathan D. Tang[y], Emma Tsurkov[gg], Austin van Loon[gg], Onur Varol[ggg,hhh], Xiafei Wang[iii], Zhi Wang[hhh,jjj], Julia Wang[y], Flora Wang[fff], Samantha Weissman[y], Kirstie Whitaker[kkk,lll], Maria K. Wolters[mmm], Wei Lee Woon[nnn], James Wu[ooo], Catherine Wu[y], Kengran Yang[aaa], Jingwen Yin[ll], Bingyu Zhao[ppp], Chenyun Zhu[ll], Jeanne Brooks-Gunn[qqq,rrr], Barbara E. Engelhardt[y,ii], Moritz Hardt[sss], Dean Knox[z], Karen Levy[ttt], Arvind Narayanan[y], Brandon M. Stewart[a], Duncan J. Watts[uuu,vvv,www], and Sara McLanahan[a,1]

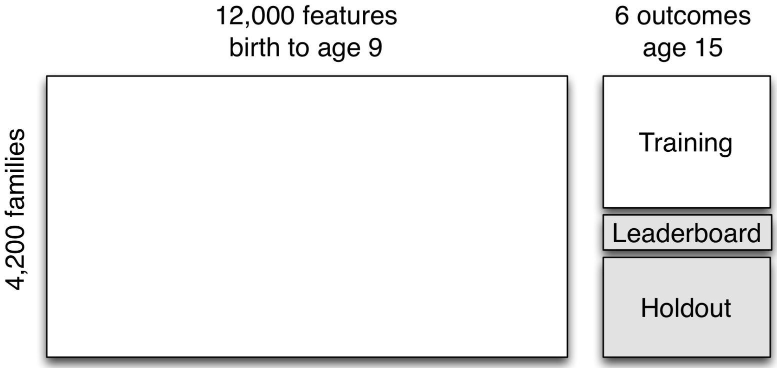
	Birth	Age 1	Age 3	Age 5	Age 9
Core mother survey	●	●	●	●	●
Primary caregiver survey			●	●	●
Core father survey	●	●	●	●	●
In-home assessment			●	●	●
Child survey					●
Child care provider survey			●		
Teacher survey				●	●



Birth
Age 1
Age 3
Age 5
Age 9
Age 15
Core mother survey
Primary caregiver survey
Combined
Core father survey
In-home assessment
Child survey
Child care provider survey
Teacher survey

Six age 15 outcomes:

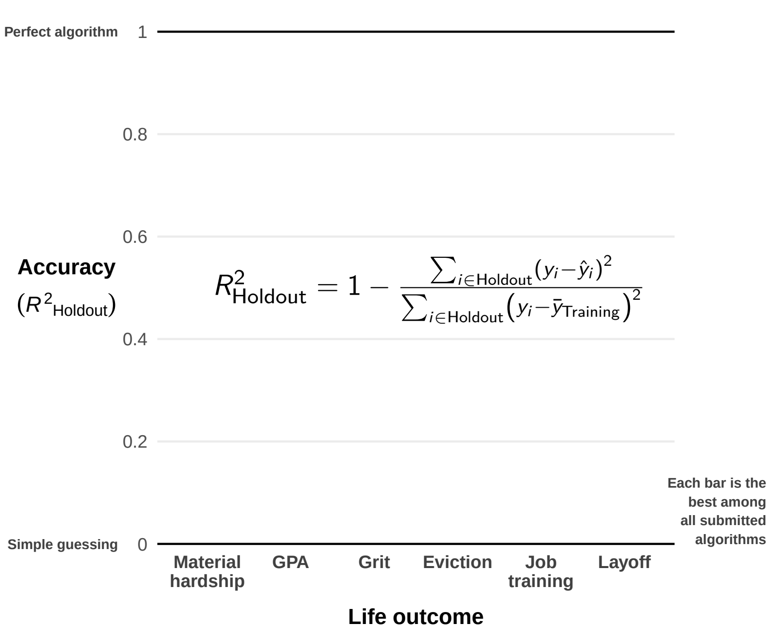
- GPA
- Material Hardship
- Grit
- Evicted
- Job training
- Job loss

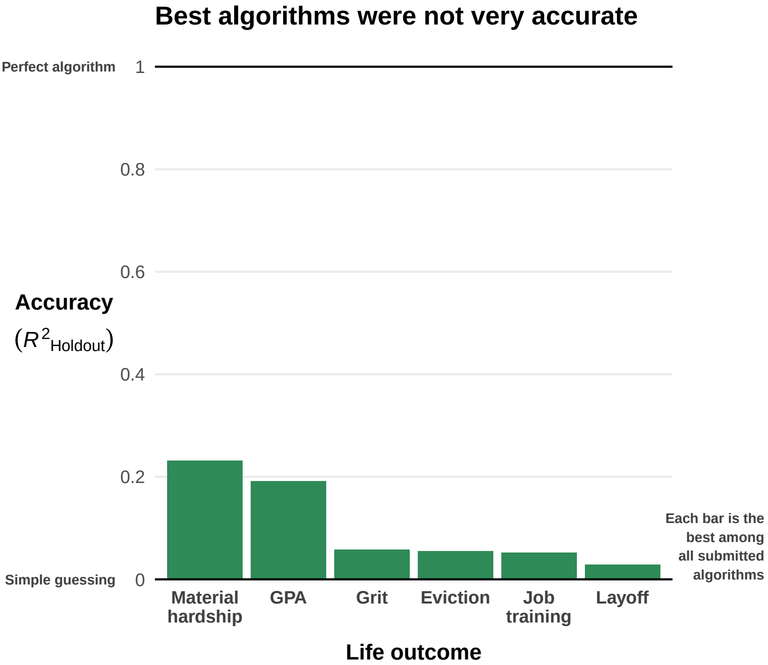

12,000 features
birth to age 9
6 outcomes
age 15
4,200 families
Training
Leaderboard
Holdout

441 registered participants

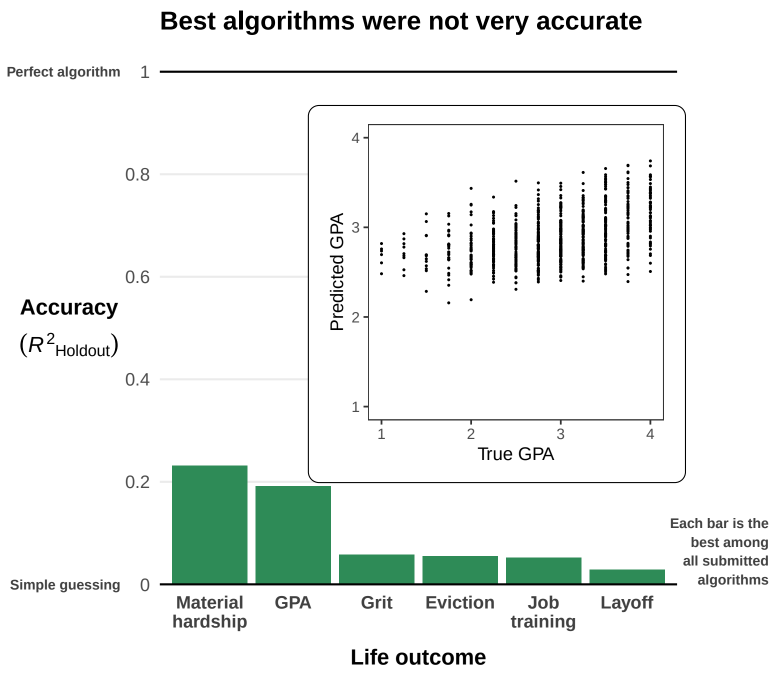
- social scientists and data scientists
- undergraduates, grad students, and professionals
- many working in teams

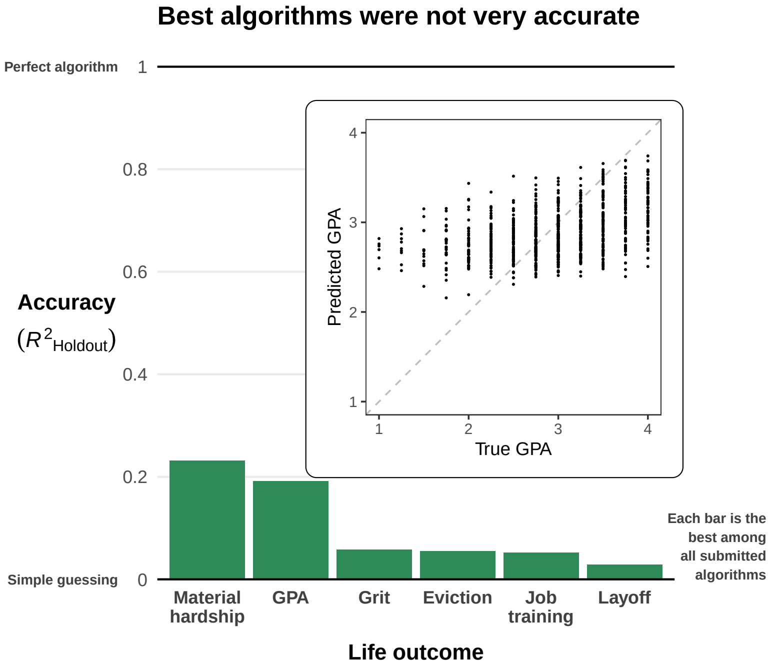
How did they do?

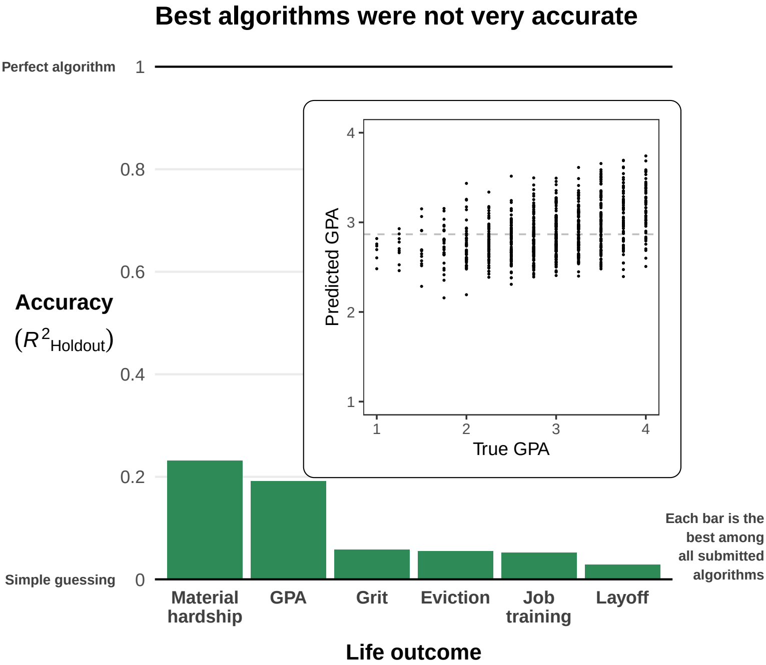

Perfect algorithm
1
0.8
0.6
Accuracy
(R^2_{Holdout})
$$R^2_{\text{Holdout}} = 1 - \frac{\sum_{i \in \text{Holdout}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{Holdout}} \left(y_i - \bar{y}_{\text{Training}}\right)^2}$$
0.4
0.2
Simple guessing
0
Each bar is the best among all submitted algorithms
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Life outcome

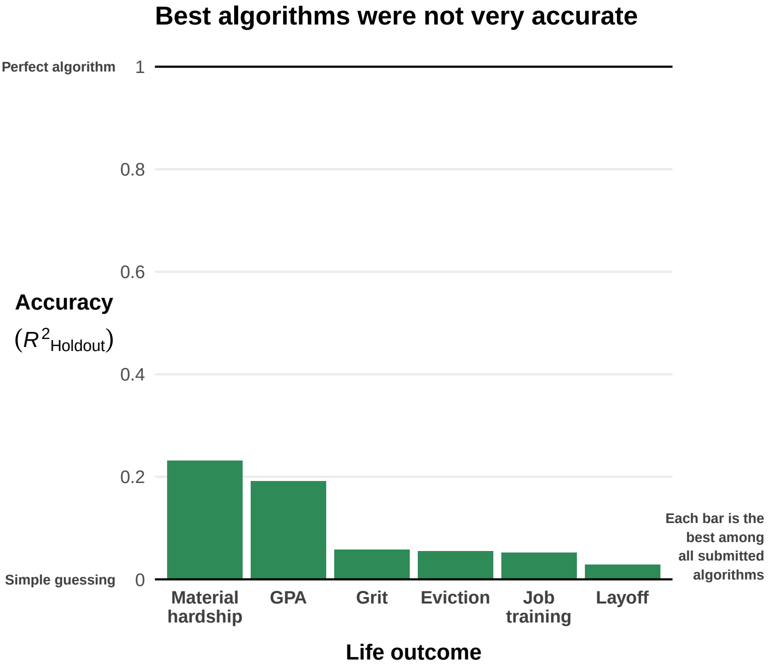


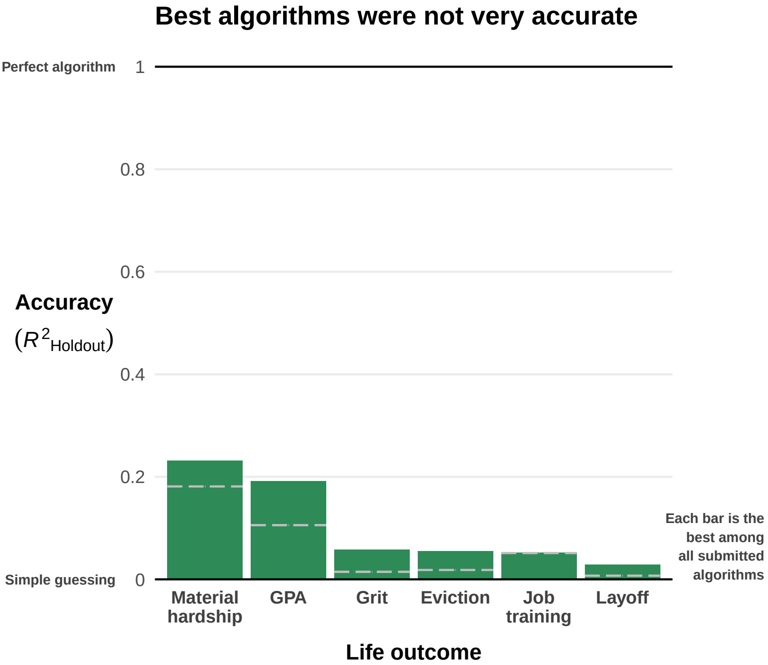
Best algorithms were not very accurate
Perfect algorithm
1
0.8
0.6
Accuracy
(R^2_{Holdout})
0.4
0.2
Simple guessing
0
Each bar is the best among all submitted algorithms
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Life outcome


Best algorithms were not very accurate
Perfect algorithm
1
0.8
0.6
Accuracy
(R^2_{Holdout})
0.4
0.2
Simple guessing
0
Predicted GPA
4
3
2
1
True GPA
1
2
3
4
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Each bar is the best among all submitted algorithms
Life outcome


Best algorithms were not very accurate
Perfect algorithm
1
0.8
0.6
Accuracy
(R^2_{Holdout})
0.4
0.2
Simple guessing
0
Predicted GPA
4
3
2
1
1
2
3
4
True GPA
Each bar is the best among all submitted algorithms
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Life outcome

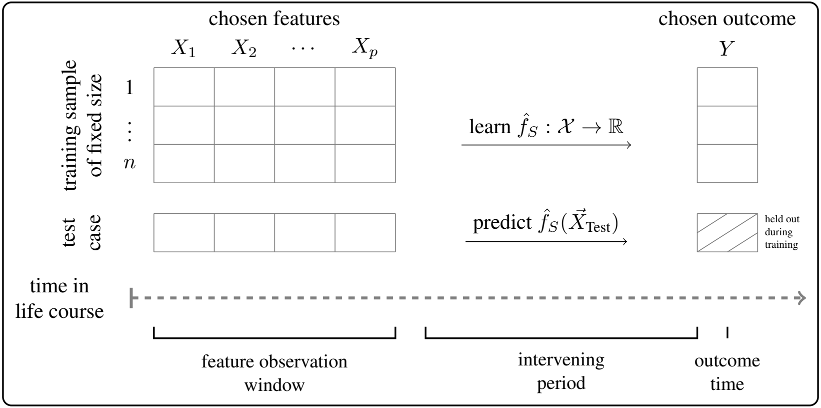

Best algorithms were not very accurate
Perfect algorithm
1
0.8
0.6
0.4
0.2
0
Simple guessing
Accuracy
(R^2_{Holdout})
Predicted GPA
4
3
2
1
True GPA
1
2
3
4
Each bar is the best among all submitted algorithms
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Life outcome


Best algorithms were not very accurate
Perfect algorithm
1
0.8
0.6
Accuracy
(R^2_{Holdout})
0.4
0.2
Simple guessing
0
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Each bar is the best among all submitted algorithms
Life outcome


Best algorithms were not very accurate
Perfect algorithm
1
0.8
0.6
Accuracy
(R^2_{Holdout})
0.4
0.2
Simple guessing
0
Material hardship
GPA
Grit
Eviction
Job training
Layoff
Each bar is the best among all submitted algorithms
Life outcome

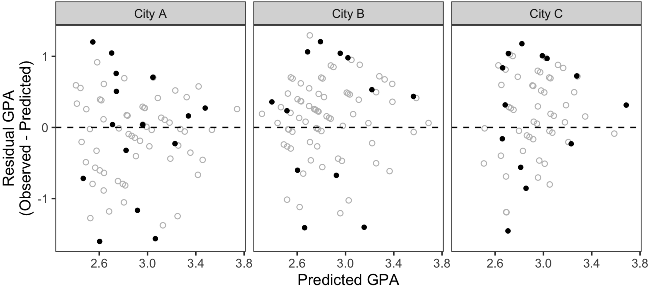
Lundberg et al. 2024.

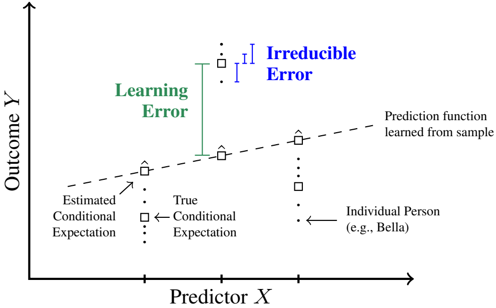
The origins of unpredictability in life outcome prediction tasks


chosen features
X_1 X_2 $\cdots$ X_p
chosen outcome
Y
training sample of fixed size
1
$\vdots$
n
learn $\hat{f}_S : \mathcal{X} \rightarrow \mathbb{R}$
test case
predict $\hat{f}_S(\vec{X}_{\text{Test}})$
held out during training
time in life course
feature observation window
intervening period
outcome time

In-depth, qualitative interviews

- 73 respondents in 40 families
- Separate interviews with the youth and primary caregiver
- Life history of the youth from birth to the interview ($\approx$ age 18)

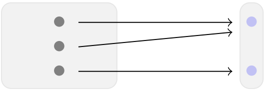



Irreducible
Error
Learning
Error
Prediction function
learned from sample
Estimated
Conditional
Expectation
True
Conditional
Expectation
Individual Person
(e.g., Bella)
Outcome Y
Predictor X

Irreducible error

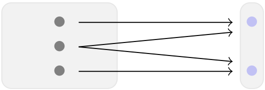
Zero Irreducible Error	Non-Zero Irreducible Error
Irreducible error is zero if each feature value maps to **one** outcome value	Irreducible error is non-zero if at least one feature value maps to **multiple** outcome values

Feature Value | Outcome Value



feature observation window | outcome observation time

Feature Value | Outcome Value



feature observation window | outcome observation time

Irreducible error: Unmeasurable features

Unmeasurable features occur after the feature observation window

Irreducible error: Unmeasurable features

Unmeasurable features occur after the feature observation window

- Bella: A lasting event

Irreducible error: Unmeasurable features

Unmeasurable features occur after the feature observation window

- Bella: A lasting event
  - after age 9, her father died

Irreducible error: Unmeasurable features

Unmeasurable features occur after the feature observation window

- Bella: A lasting event
  - after age 9, her father died
  - high school went off course

Irreducible error: Unmeasurable features

Unmeasurable features occur after the feature observation window

- Bella: A lasting event
  - after age 9, her father died
  - high school went off course
- Charles: A fleeting event

Irreducible error: Unmeasurable features

Unmeasurable features occur after the feature observation window

- Bella: A lasting event
  - after age 9, her father died
  - high school went off course
- Charles: A fleeting event
  - online high school

Irreducible error: Unmeasurable features

Unmeasurable features occur after the feature observation window

- Bella: A lasting event
  - after age 9, her father died
  - high school went off course
- Charles: A fleeting event
  - online high school
  - worked in the basement for one semester

Irreducible error: Unmeasurable features

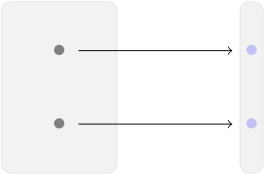
Unmeasurable features occur after the feature observation window

- Bella: A lasting event
  - after age 9, her father died
  - high school went off course
- Charles: A fleeting event
  - online high school
  - worked in the basement for one semester
  - video games = bad grades that semester

Irreducible error: Unmeasurable features

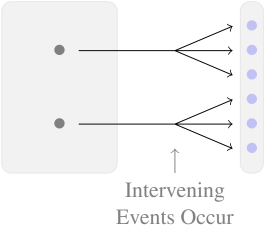
Zero Irreducible Error

Without intervening events,



Non-Zero Irreducible Error

With intervening events,



Irreducible error: Unmeasured features

Irreducible error: Unmeasured features

Lola's social network

Irreducible error: Unmeasured features

Lola's social network

- elderly neighbor got Lola ready for school each day

Irreducible error: Unmeasured features

Lola's social network

- elderly neighbor got Lola ready for school each day
- grandparents remodeled the basement to house Lola

Irreducible error: Unmeasured features

Lola's social network

- elderly neighbor got Lola ready for school each day
- grandparents remodeled the basement to house Lola
- aunt employed Lola's mother in a family business

Irreducible error: Unmeasured features

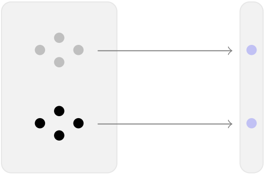
Lola's social network

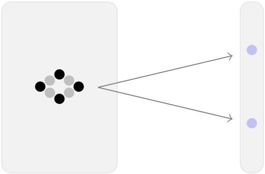
- elderly neighbor got Lola ready for school each day
- grandparents remodeled the basement to house Lola
- aunt employed Lola's mother in a family business

Predicted GPA: 3.04 Actual GPA: 3.75

Irreducible error: Unmeasured features

Zero Irreducible Error	Non-Zero Irreducible Error
Feature is measured,	Feature is unmeasured,





Irreducible error: Imperfectly measured features

Irreducible error: Imperfectly measured features

How close do you feel to your mom? Would you say…

Extremely close, ... 1
Quite close,.. 2
Fairly close, or, ... 3
Not very close?.. 4
REFUSED .... -1
DON’T KNOW .... -2

Irreducible error: Imperfectly measured features

How close do you feel to your mom? Would you say…

Extremely close,	1
Quite close,	2
Fairly close, or,	3
Not very close?	4
REFUSED	-1
DON'T KNOW	-2

A daughter told us about her "not very close" mother

Irreducible error: Imperfectly measured features

How close do you feel to your mom? Would you say…

Extremely close, ... 1
Quite close,.................... 2
Fairly close, or, ... 3
Not very close?.................... 4
REFUSED .... -1
DON'T KNOW .... -2

A daughter told us about her "not very close" mother

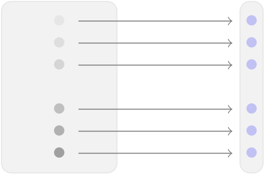
- kicked her out of the house and called police
- mother: "you better start treating me better, because I might not live that long.''
- daughter: "I couldn't even focus in class. . . I was shaking.''

Outcome: Failed 8th grade. Low GPA. Dropped out.

Irreducible error: Imperfectly measured features

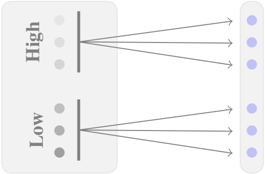
Zero Irreducible Error

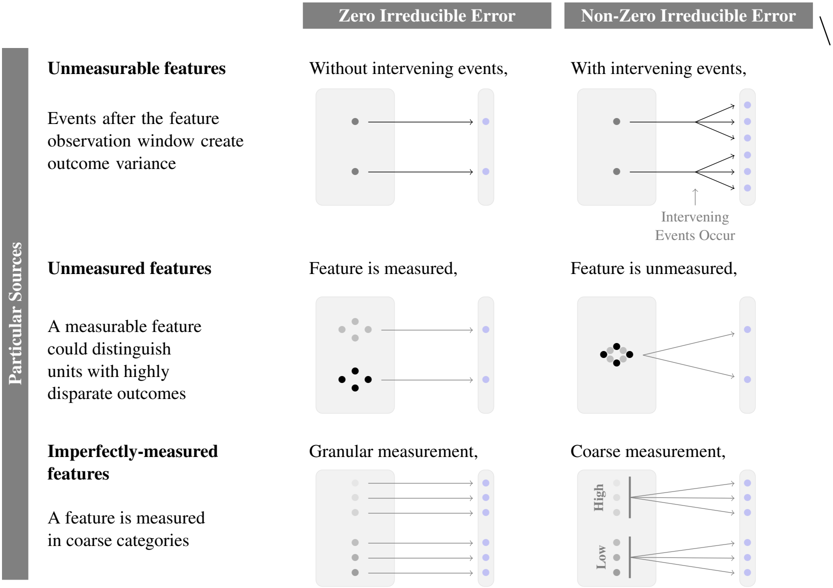
Granular measurement,



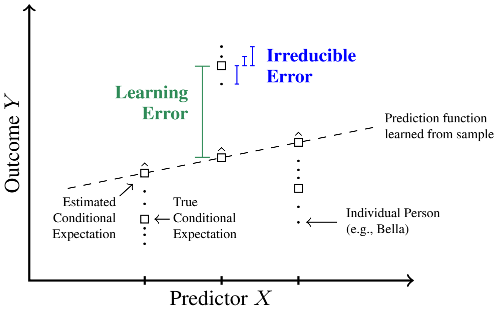
Non-Zero Irreducible Error

Coarse measurement,



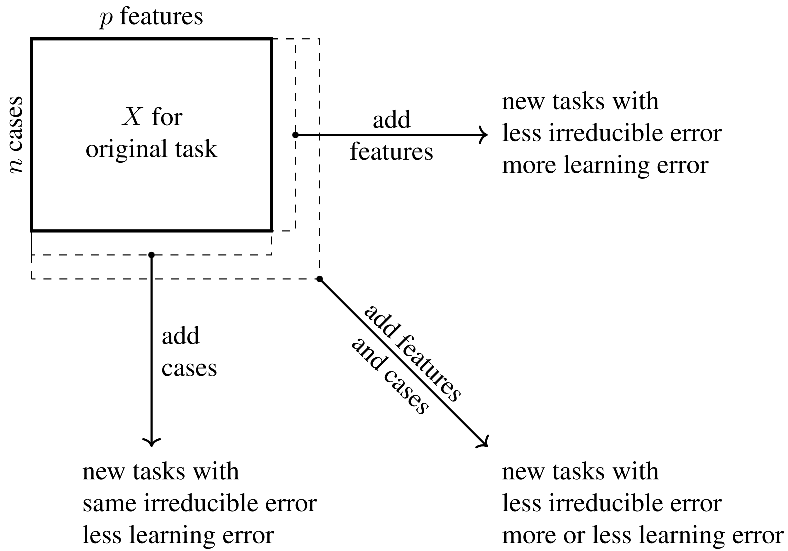


Zero Irreducible Error
Non-Zero Irreducible Error
Particular Sources
Unmeasurable features
Events after the feature observation window create outcome variance
Without intervening events,
With intervening events,
Intervening Events Occur
Unmeasured features
A measurable feature could distinguish units with highly disparate outcomes
Feature is measured,
Feature is unmeasured,
Imperfectly-measured features
A feature is measured in coarse categories
Granular measurement,
Coarse measurement,
High
Low


Irreducible
Error
Learning
Error
Prediction function
learned from sample
Outcome Y
Estimated
Conditional
Expectation
True
Conditional
Expectation
Individual Person
(e.g., Bella)
Predictor X

DISCUSSION

Generalizing to other life outcome prediction tasks



Implications for policy

Implications for policy

- life outcome predictions may be inaccurate

Implications for policy

- life outcome predictions may be inaccurate
  - if generated by algorithms
  - if generated by humans

Implications for policy

- life outcome predictions may be inaccurate
  - if generated by algorithms
  - if generated by humans
- from accuracy to impact evaluations

Implications for science

Implications for science

- old goal: between-group variability
  - how means vary across groups

Implications for science

- old goal: between-group variability
  - how means vary across groups
- new goal: within-group variability
  - how variances vary across groups

Implications for science

- old goal: between-group variability
  - how means vary across groups
- new goal: within-group variability
  - how variances vary across groups
- more work to better understand unpredictability
  - empirical estimates
  - formal models

Learning goals for today

By the end of class, you will be able to

- know who had the best predictions!
- reason about predictability of life outcomes