Variable importance and explainable AI

Art B. Owen

Stanford University

Based on joint work with:

Masayoshi Mase, Hitachi, Ltd.

Ben Seiler, Stanford Statistics

Opinions are my own, and not those of Stanford, the NSF, or Hitachi, Ltd.

Black box algorithms

- deep neural networks
- random forests
- etc.

State of the art accuracy, but

- hard to interpret / explain
- concerns over fairness

Additionally

Variable importance has direct interest

Variable importance

A step towards explanation

Criteria from Jiang & O (2003)

For $\boldsymbol{x} = (x_1, \dots, x_d)$, x_j is important if

1) x_j affects Y ***causally***
2) x_j affects fits $f(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) = \widehat{\mathbb{E}}(Y \mid \boldsymbol{x})$; call it ***mechanically***
3) omitting x_j deteriorates the fit, e.g., R^2

Explaining a prediction is about case 2

Which x_j are important for $f(\boldsymbol{x})$?

Variable importance literatures

Statistics and uncertainty quantification

P. Wei, Z. Lu, and J. Song. (2015)

Survey of 197 papers

Including 24 survey papers

Global sensitivity analysis

Razavi et al. (2021)

All star team of 26 GSA authors

100s of references

Explainable AI

C. Molnar (2018)

online book

Other areas

law / insurance / fairness / economics (e.g., Shapley value)

Easy!

We can compute any counter-factual $f(\boldsymbol{x}) - f(\boldsymbol{x}')$

Actually no

It is still hard.

Harder than causal inference

We want ***causes of effects***

not ***effects of causes***

Holland (1988) makes this point;

refers to philosopher Mill (1843)

rules out experiments for 'causes of effects'

The difference

Dawid & Musio (2021)

Does taking Lipitor increase the chance of type II diabetes?

Did Juanita get type II diabetes because of Lipitor?

Two very different questions

Example

Accident caused by many variables all going wrong at once (e.g. Tenerife)

maybe no accident **but for**

fog, crowding, extra fuel, distractions $\cdots$ communications

which is ***most*** causal?

https://en.wikipedia.org/wiki/Tenerife_airport_disaster

Why was $f(\boldsymbol{x}) > 0$?

We cannot use

- holdout samples
- bakeoffs on future data

Because $f(\boldsymbol{x})$ is completely known for all $\boldsymbol{x}$ we might want to try

Variable importance

A is an ***important variable*** if changing A changes B

where B is important

Why is B important?

It just is

so we avoid infinite regress

or a circular argument

Upshot

For us, importance is ***transferred*** not created

Quantifying importance

We have

$$f(\boldsymbol{x}), \quad \boldsymbol{x} = (x_1, x_2, \dots, x_d) \qquad x_j \in \mathcal{X}_j$$

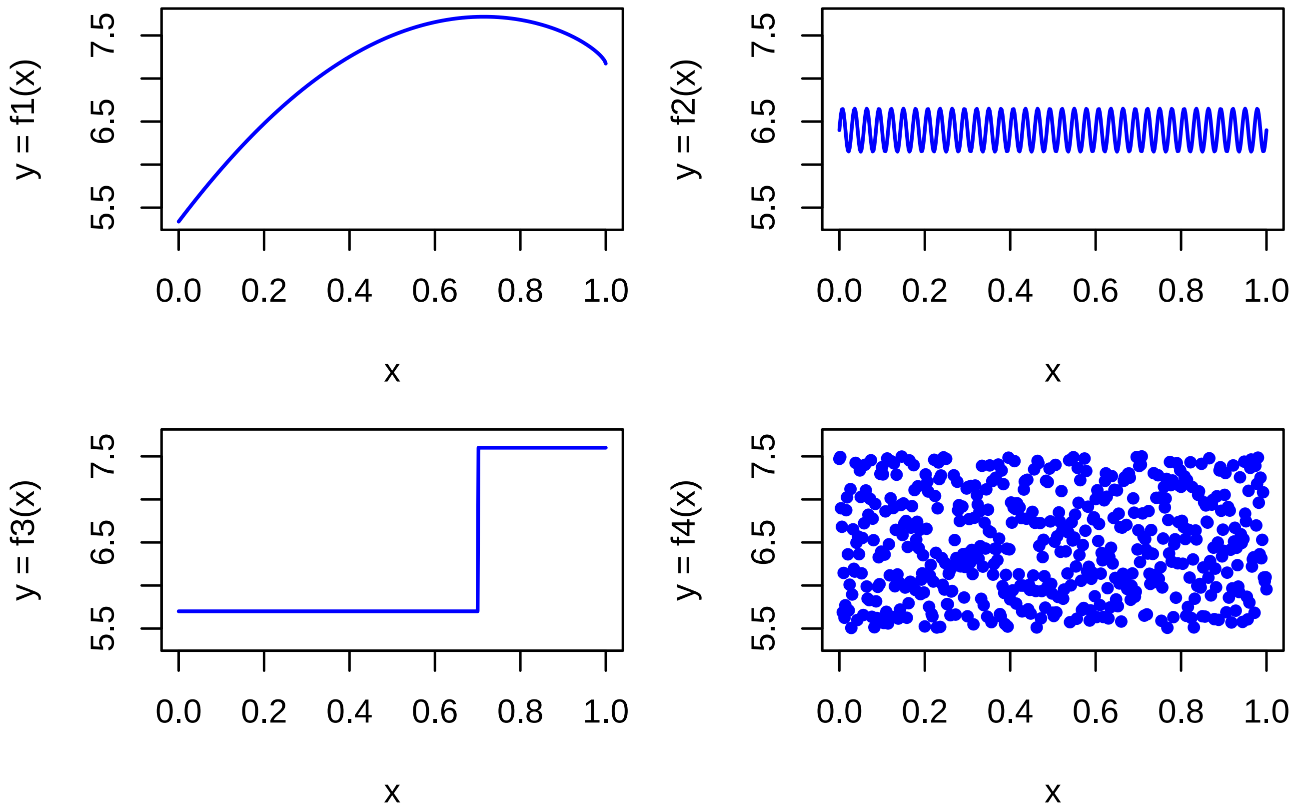
Importance of x_j on $\hat{y} = f(\boldsymbol{x})$

Change $x_j \to x_j'$ and watch $\hat{y}$ respond.

1) Which x_j do we ***start*** with?
2) What x_j' do we change it ***to***?
3) Where is x_k for $k \neq j$ while this is going on?
4) How do we aggregate all those changes?

Too many choices to list

When is x is most influential?



Depends on how you want to keep score,

$\cdots$ which depends on your goals.

Easy case

$\boldsymbol{x} = (x_1, \dots, x_d)$ for independent $x_j \in \mathcal{X}_j$ and

$$f(\boldsymbol{x}) = \sum_{j=1}^{d} f_j(x_j) \qquad \text{additive}$$

We can use single variable measures, e.g.,

$$\mathrm{Var}(f_j(x_j))$$
$$\mathbb{E}(|f_j(x_j) - f_j(x_j')|)$$
$$\int |f_j'(x)| \, \mathrm{d}x$$
$$\max_x |f_j'(x)|$$
$$\max_x f_j(x) - \min_x f_j(x)$$

Inputs

$f_j(x_j) - f_j(x_j')$ for $x_j, x_j' \in \mathcal{X}_j$

Multivariable complexities

- Interactions

  effect of changing x_1 depends on $x_2, x_3, \ldots, x_d$

- Correlation / dependency

  should changes to x_1 change x_2?

Most methods change ***some*** of the components of $\boldsymbol{x}$ but not all

Hybrid points

$\boldsymbol{x} = (x_1, x_2, \dots, x_9)$
$\boldsymbol{z} = (z_1, z_2, \dots, z_9)$

$u = \{1, 3, 7, 8\}$
$-u \equiv u^c = \{1, 2, \dots, 9\} \setminus u = \{2, 4, 5, 6, 9\}$

Combine two points: $\boldsymbol{x}$, $\boldsymbol{z}$

$$
\begin{array}{rcccccccccccc}
\boldsymbol{x} = & (& x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 &) \\
 & & & \downarrow & & \downarrow & \downarrow & \downarrow & & & \downarrow & \\
\boldsymbol{x}_{-u}{:}\boldsymbol{z}_u = & (& z_1 & x_2 & z_3 & x_4 & x_5 & x_6 & z_7 & z_8 & x_9 &) \\
 & & \uparrow & & \uparrow & & & & \uparrow & \uparrow & & \\
\boldsymbol{z} = & (& z_1 & z_2 & z_3 & z_4 & z_5 & z_6 & z_7 & z_8 & z_9 &)
\end{array}
$$

Compare

$f(\boldsymbol{x}_{-u}{:}\boldsymbol{z}_u) - f(\boldsymbol{x})$

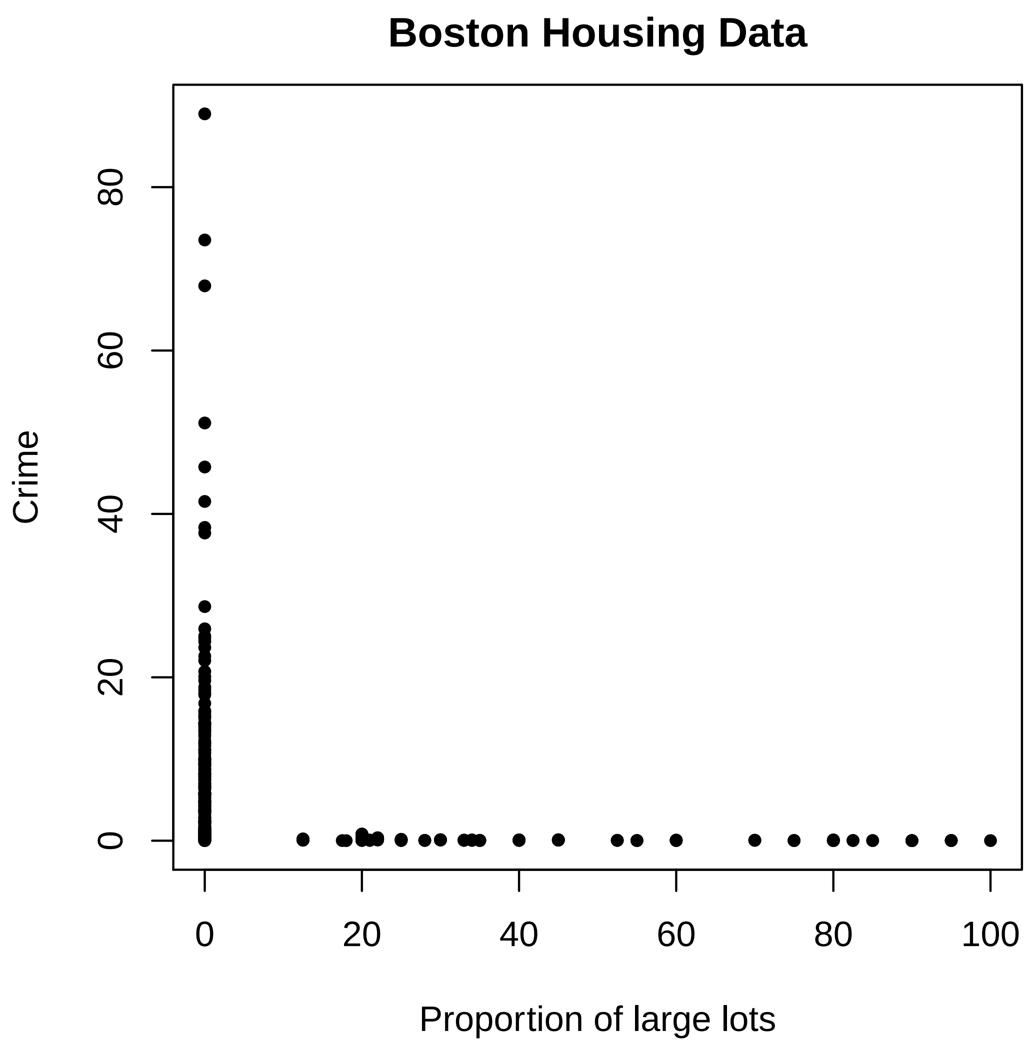
carries clues to importance of variables $j \in u$

Awkward combinations

If x_1 and x_2 are highly correlated (or structured)

$\Longrightarrow$ x_1:z_2 could be quite unlikely

Y = median housing value: 506 regions and 13 predictors Harrison & Rubinfeld (1978)

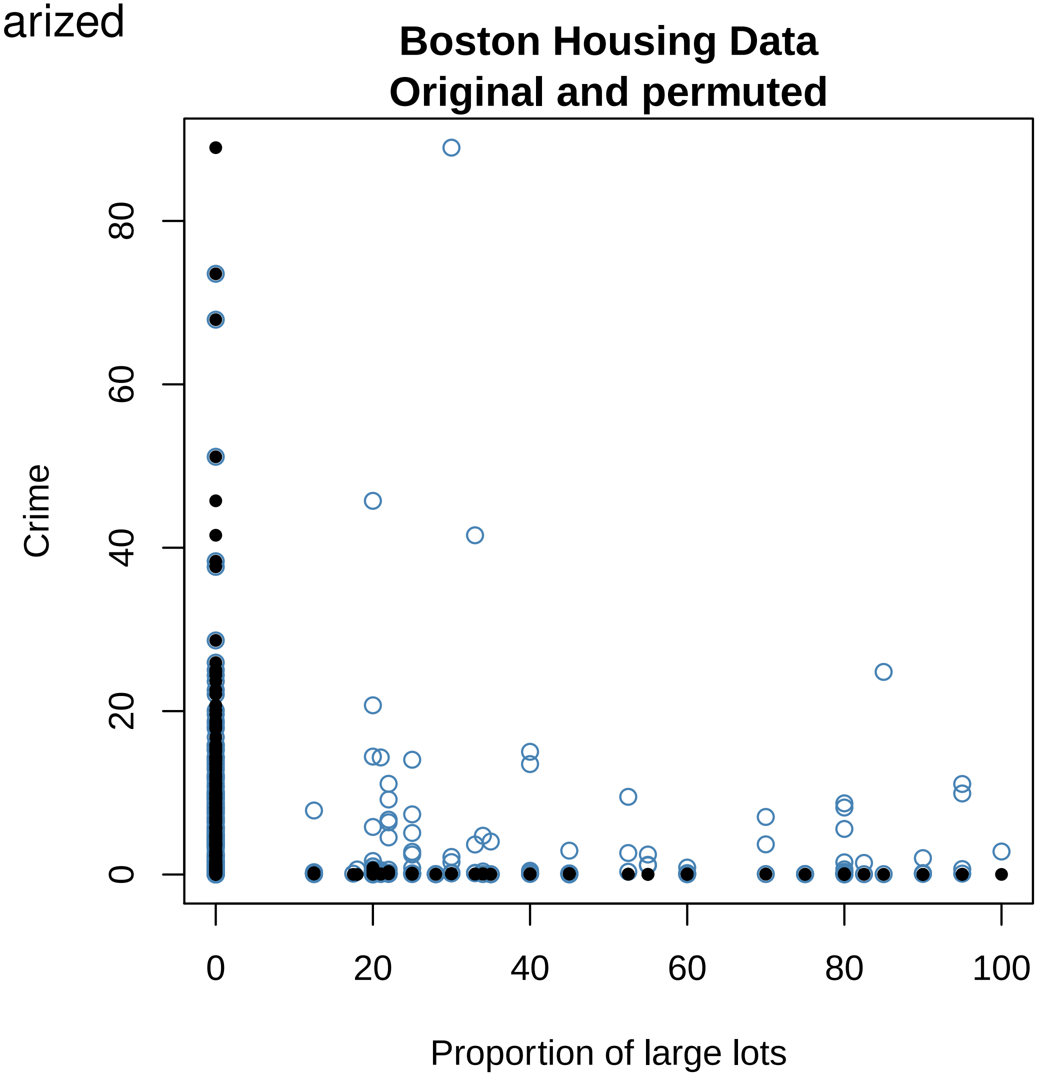


Awkward combinations

Random pairings do not describe 1970s Boston

Any predictions at such points are problematic

Not well regularized



Breiman's permutation

Random forests: Breiman (2001) $f(\boldsymbol{x}) = \widehat{\mathbb{E}}(Y \mid \boldsymbol{x})$

To judge x_j, permute $x_{1j}, x_{2j}, \dots, x_{nj}$

Old $\boldsymbol{x}$'s	New $\boldsymbol{x}$'s
(x_{11}, x_{12})	(x_{11}, x_{32})
(x_{21}, x_{22})	(x_{21}, x_{22})
(x_{31}, x_{32})	(x_{31}, x_{52})
(x_{41}, x_{42})	(x_{41}, x_{42})
(x_{51}, x_{52})	(x_{51}, x_{12})

Recompute $\sum_i (y_i - f(x_i))^2$ on permuted values

Like a Sobol' index.

Uses problematic inputs.

Physically impossible

- Birth date $>$ graduation date
- Systolic blood pressure $<$ diastolic
- Longitude / lattitude combination $\Longrightarrow$ dwelling in ocean
- County $=$ Los Angeles & State $=$ Colorado

Problems

- We cannot trust any explanation that used these combinations
- Hard to avoid them computationally

Logically impossible

- $x_{\text{Annual}} = x_{\text{Jan}} + x_{\text{Feb}} + \cdots + x_{\text{Dec}} \neq z_{\text{Annual}}$
- Patient's Min. blood O_2 $>$ Avg. blood O_2
- Min $O_2 \neq$ Max O_2 while # measurements $= 1$ (or 0)

Sobol' and Shapley

Sobol' indices handles interactions among independent variables

Shapley handles interactions and dependence

Global sensitivity analysis

This is a large literature since the early 1990s

See SIAM / ASA Journal of Uncertainty Quantification

Global sensitivity analysis books

Fang, Li & Sudijanto (2010),

Saltelli, Chan & Scott (2009),

Saltelli, Ratto & Andres (2008),

Cacuci, Ionescu-Bujor & Navon (2005),

Saltelli, Tarantola & Campolongo (2004),

Santner, Williams & Notz (2003)

and there are many more articles.

Many references on Sobol' indices:

driven by variance explained

Shapley value

Baseline Shapley plus survey

Najmi & Sundararajan (2020)

Uncertainty quantification

O (2014), Song, Nelson Staum (2016), O & Prieur (2017)

Shapley for interactions

Rabitti & Borgonovo (2019)

Computations

Plischke, Rabitti & Borgonovo (2019)

Black box explanations

Strumbelj & Kononenko (2010)

SHapley Additive exPlanations (SHAP)

Lundberg & Lee (2017)

Data Shapley

Gorbani & Zou (2019,2020)

Qualms

Kumar et al. (2020)

From economics

How to attribute a reward among multiple causes or team members.

Solved by Shapley (1953)

\$15 million

Shapley's (1953) value measures contributions of team members.

We need to know what each subset of the team would have accomplished.

Example from Bank of International Settlement

Team	Output value
$\varnothing$	0
A	4,000,000
B	4,000,000
C	4,000,000
A,B	9,000,000
A,C	10,000,000
B,C	11,000,000
A,B,C	15,000,000

Q: How should we split the \$15,000,000 earned by A, B, C among them?

$15 million

Example from Bank of International Settlement

Team	Output value
∅	0
A	4,000,000
B	4,000,000
C	4,000,000
A,B	9,000,000
A,C	10,000,000
B,C	11,000,000
A,B,C	15,000,000

Q: How should we split the $15,000,000 earned by A, B, C among them?

A: Shapley (1953) says: A gets $4,500,000, B gets $5,000,000, C gets $5,500,000

Shapley setup

Team $u \subseteq \mathcal{D} \equiv \{1, 2, \dots, d\}$ creates value $\textbf{val}(u)$.

Total value is $\textbf{val}(\mathcal{D})$.

Player j should get ϕ_j.

Incremental value of j given u

$$\textbf{val}(j \mid u) = \textbf{val}(u \cup \{j\}) - \textbf{val}(u)$$

Shapley axioms

Efficiency	$\sum_{j=1}^{d} \phi_j = \textbf{val}(\mathcal{D})$
Dummy	If $\textbf{val}(j \mid u) = 0$, all u then $\phi_j = 0$
Symmetry	If $\textbf{val}(i \mid u) = \textbf{val}(j \mid u)$, when $u \cap \{i, j\} = \varnothing$ then $\phi_i = \phi_j$
Additivity	If games **val**, $\textbf{val}'$ have values ϕ, ϕ' then $\textbf{val} + \textbf{val}'$ has value $\phi_j + \phi'_j$

Unique solution

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \textbf{val}(j \mid u)$$

For variable importance

Variables $x_1, x_2, \dots, x_d$ team up to explain f.

Variance explained:

$$\textbf{val}(u) = \mathrm{Var}(\mathbb{E}(f(\boldsymbol{x}) \mid \boldsymbol{x}_u))$$

Variance explained under dependence

Song, Nelson & Staum (2016),

O & Prieur (2017)

Local importance

Variance explained is ***global***, i.e., all data or a distribution

Local questions

why was target person turned down for a loan?

why did the algo recommend intensive care unit?

Target subject t

For some $t \in 1{:}n$ we want to "explain" $f(\boldsymbol{x}_t)$

Baseline Shapley

Najmi & Sundararajan (2020)

n subjects $i = 1, \dots, n$

Target subject $t \in 1{:}n$ has $f(\boldsymbol{x}_t)$

Baseline point $\boldsymbol{x}_b = (x_{b1}, x_{b2}, \dots, x_{bd})$
Your choice. Could be $\boldsymbol{x}_b = \bar{\boldsymbol{x}} \equiv (1/n)\sum_{i=1}^n \boldsymbol{x}_i$

To explain $f(\boldsymbol{x}_t) - f(\boldsymbol{x}_b)$

$$\textbf{val}(u) = f(\boldsymbol{x}_{t,u}{:}\boldsymbol{x}_{b,-u})$$ "Baseline Shapley"

$$\textbf{val}(u) = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_{t,u}{:}\boldsymbol{x}_{i,-u})$$ "random Baseline Shapley"

$$\textbf{val}(u) = \mathbb{E}(f(\boldsymbol{x}) \mid \boldsymbol{x}_u)$$ "cond expectation Shapley"

Given the value function, Shapley does the rest

Cost is exponential in d

Use Monte Carlo for large d

Our contributions

Three papers on arxiv by Mase, Seiler, O

- arXiv:1911.00467

  introduces cohort Shapley

- arXiv:2105.07168

  uses it for fairness

- arXiv:2105.08013

  uses it to quantify what variable(s) identify you

Cohort Shapley

Motivation:

avoid impossible combinations

by only using actually observed combinations

counters some adversarial attacks described in Slack et al (2020)

close to conditional expectation Shapley with empirical distribution

Mase, Seiler, O (2019) arXiv:1911.00467

Similarity

Target has $\boldsymbol{x}_t = (x_{t1}, \dots, x_{td})$.

Define

$$z_{ij} = z_{ij}(t) = \begin{cases} 1, & x_{ij} \text{ `similar' to } x_{tj} \\ 0, & \text{else.} \end{cases}$$

E.g.: $x_{ij} = x_{tj}$, or $|x_{ij} - x_{tj}| \leqslant \delta_j$

Toy example

	Subj	Color	Breakfast	$Z_{i1}(5)$	$Z_{i2}(5)$	$Z_{i,\{1,2\}}(5)$
	1	red	eggs	0	1	0
	2	red	cereal	0	0	0
	3	blue	cereal	1	0	0
	4	red	eggs	0	1	0
Target	5	blue	eggs	1	1	1

Cohorts

$$\{1,4,5\} \rightarrow \{5\}$$

similar food $\uparrow$ $\uparrow$

$$\{1,2,3,4,5\} \rightarrow \{3,5\}$$

similar color

Toy continued

$$\begin{array}{lccc} & \{1,4,5\} & \rightarrow & \{5\} \\ \text{similar food} & \uparrow & & \uparrow \\ & \{1,2,3,4,5\} & \rightarrow & \{3,5\} \\ & & \text{similar color} & \end{array}$$

Similarity constraints

$$\begin{array}{lccc} & \{2\} & \rightarrow & \{1,2\} \\ \text{similar food} & \uparrow & & \uparrow \\ & \varnothing & \rightarrow & \{1\} \\ & & \text{similar color} & \end{array}$$

Value function

Cohorts

$$C_{t,u} = \{i \in 1{:}n \mid z_{ij}(t) = 1,\ \text{all } j \in u\}$$

Cohort means

$$\textbf{val}(u) = \textbf{val}(u;t) \equiv \bar{y}_{t,u} = \frac{1}{|C_{t,u}|} \sum_{i \in C_{t,u}} f(\boldsymbol{x}_i)$$

Cohort refinement

Start with

$$C_{t,\varnothing} = \{1, 2, \dots, n\}$$

Each j added to u refines the cohort by removing dissimilar subjects.

Important j move the cohort means the most

Value function

$$\textbf{val}_{\mathrm{CS}}(u) = \bar{y}_{t,u} \qquad \text{or} \quad \bar{y}_{t,u} - \bar{y}_{t,\varnothing}$$

Centering doesn't change ϕ_j

Fourth importance

Start from blank slate

- reveal x_{tj} in any order
- revealing an important variable tells more about y_t

I.e., ***knowledge*** about x_{tj} is informative about $f(\boldsymbol{x}_t)$

Variables not in the model

Cnsider $f(\boldsymbol{x}) = g(x_1, x_3, x_4)$ with $x_2 \approx x_1$

Is x_2 important?

Baseline Shapley attributes it all to x_1

Cohort Shapley shares importance

similar $x_1 \iff$ similar x_2

Any choice we make is ***a feature and a bug***

Catch-22 according to Kumar et al. (2020)

Cohort Shapley can detect redlining

It could also find false positives

COMPAS recidivism risk score

Correctional Offender Management Profiling for Alternative Sanctions

See e.g., Chouldechova (2017)

Sources

Proprietary algorithm from NorthPointe Inc.

Broward County data 2013, 2014 available via ProPublica

Variables

We used $n = 5278$ obs (Black and White) of 6172

$p = 5$ predictors:

Age, Race, Gender, # Priors, Crime (felony vs misdemeanor)

discretized as in Chouldechova (2017)

Responses

$Y =$ reoffended

$\hat{Y} =$ predicted to reoffend

Properties

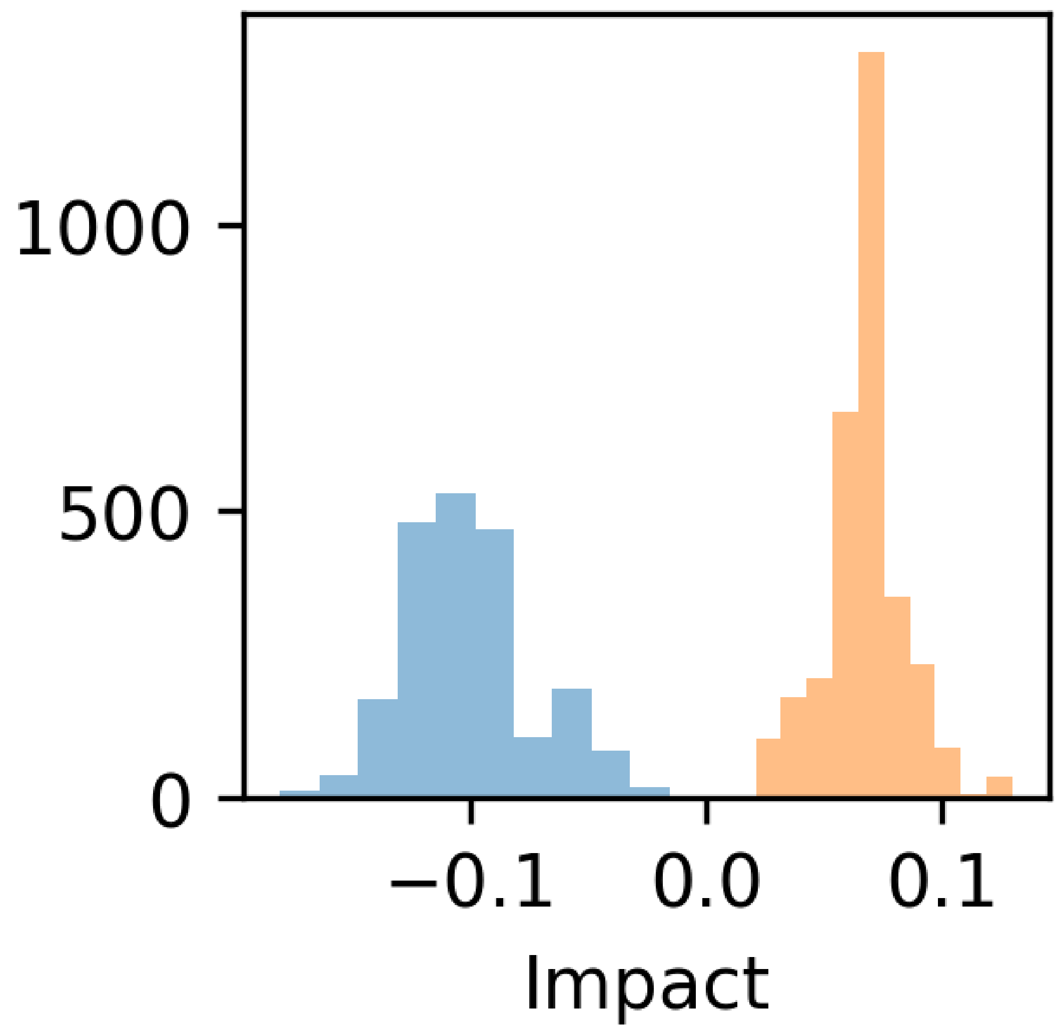
1) COMPAS did not use race
2) Proprietary algorithm: we don't have $f(\cdot)$
3) Algo was not trained on Broward County

We can still apply cohort Shapley

We get variable importance for each person's race / gender etc.

Our one analysis is not necessarily definitive

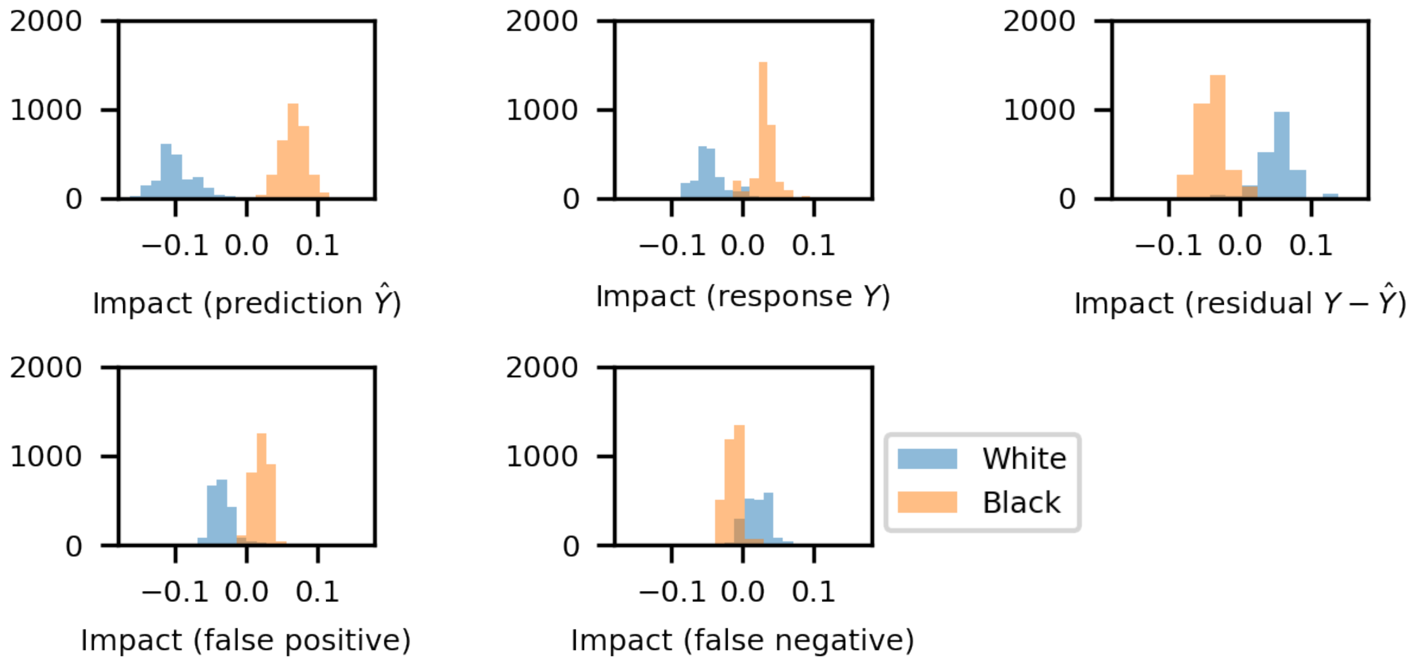
Cohort Shapley effects for race



Response is ‘predicted to re-offend’

Orange is for Black subjects   Blue for White

Shapley effects for race, ctd



Responses

prediction $\hat{Y}$

response Y

false positive $Y = 0$ & $\hat{Y} = 1$

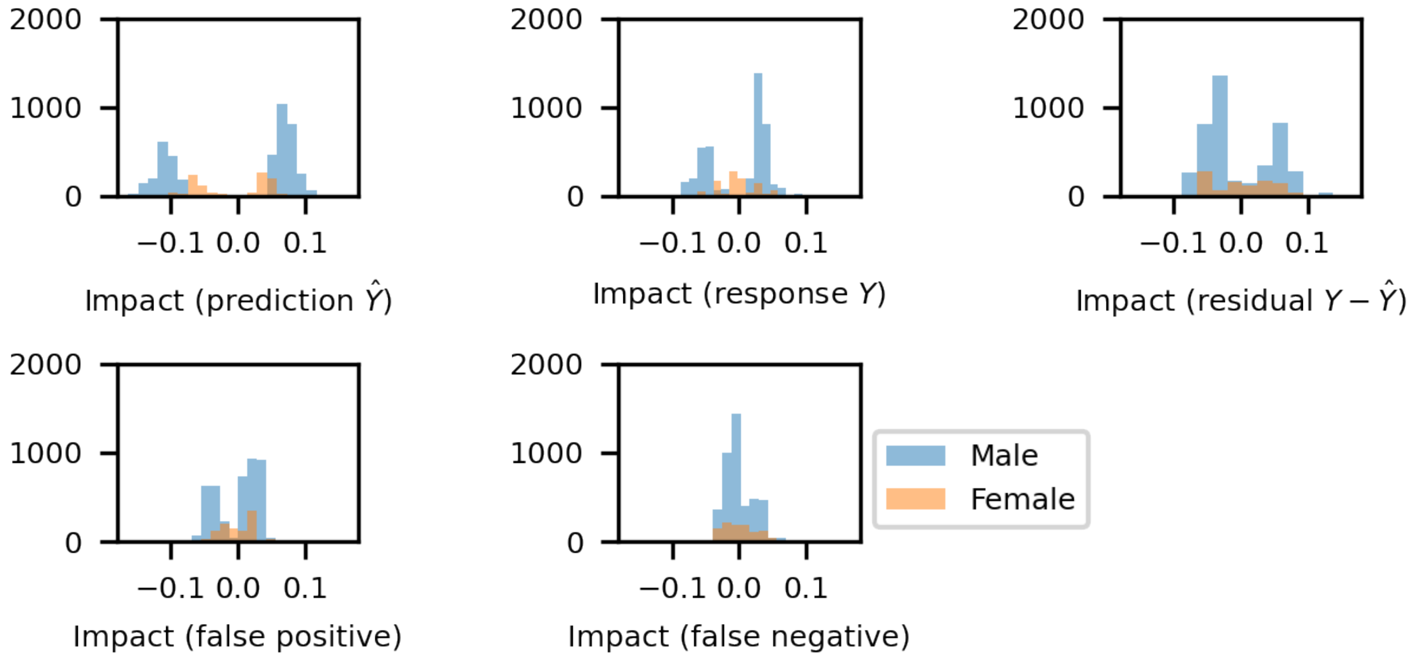
false negative $Y = 1$ & $\hat{Y} = 0$

There's a debate about $Y \mid \hat{Y}$ vs $\hat{Y} \mid Y$

Chouldechova (2017)

Gender split

Cohort Shapley for race



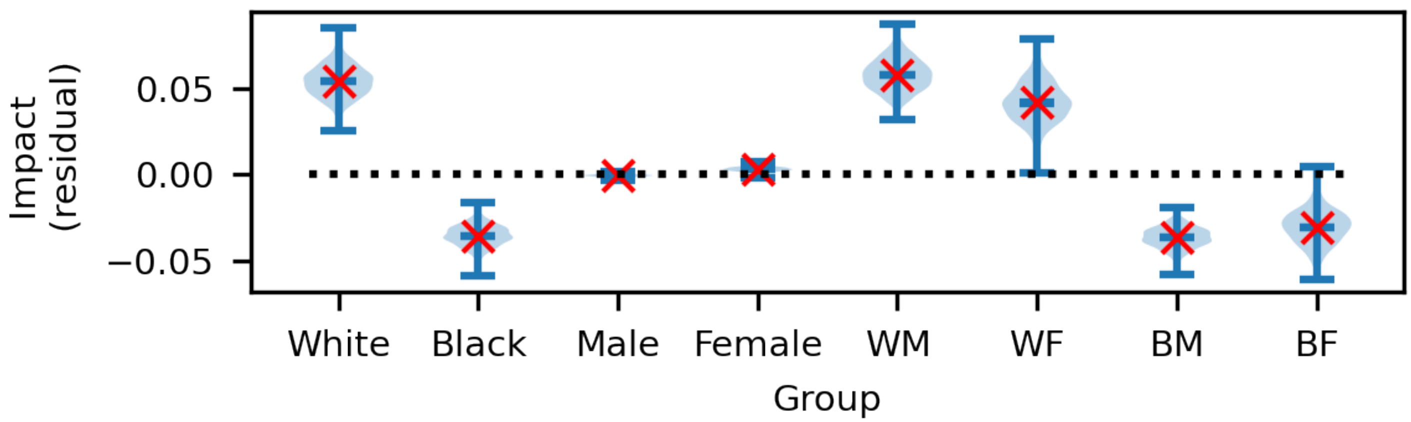
Responses

prediction $\hat{Y}$ response Y

false positive $Y = 0$ & $\hat{Y} = 1$ false negative $Y = 1$ & $\hat{Y} = 0$

Bootstrap

Aggregate cohort Shapley for $Y - \hat{Y}$



Violin plot from Bayesian bootstrap: Rubin (1981)

reweight observations by Exp(1) random variables

Uniqueness measure

Golle (2006)

In 1990 census data, 87% of the US population can be uniquely identified by gender, ZIP code and full date of birth

Uniqueness Shapley

val$(u) = -\log_2(\#C_{t,u})$ (log of cohort cardinality)

ϕ_j describes power to identify target t

North Carolina voter registration

$n = 7{,}538{,}125$

Huge speedup using all dimension trees of Moore & Lee (1998)

We can see how identifying: Zip Code, Race, Party, Gender, Age are

for individuals

for aggregates

Next steps

Think more about how to interpret Shapley impacts

E.g., what response is most appropriate?

What about missing variables?

Which variables to include/exclude

Which subsets of subjects?

Generalize to Shapley interactions

Thanks

- Masayoshi Mase, Benjamin Seiler, co-authors
- Chiara Sabatti for her literature course on fairness
- Hitachi, Ltd.
- NSF: IIS-1837931
- Priscilla Travis, Nick Cogan, Aseel Farhat, Sabrena Bouie, Billy Oates, Yousuff Hussaini
- William Becker, Samuele Lo Piano
- Giray Ökten