Random Forest, Causal Trees and Causal Forest

Alexander Quispe

November 5, 2021

Citation

Images are taken from Sujan Dutta notes on Random Forest. Notes on Causal Trees and Causal Forest are taken from Susan Athey Lecture Notes on Machine Learning and Causal Inference, 2021.

Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

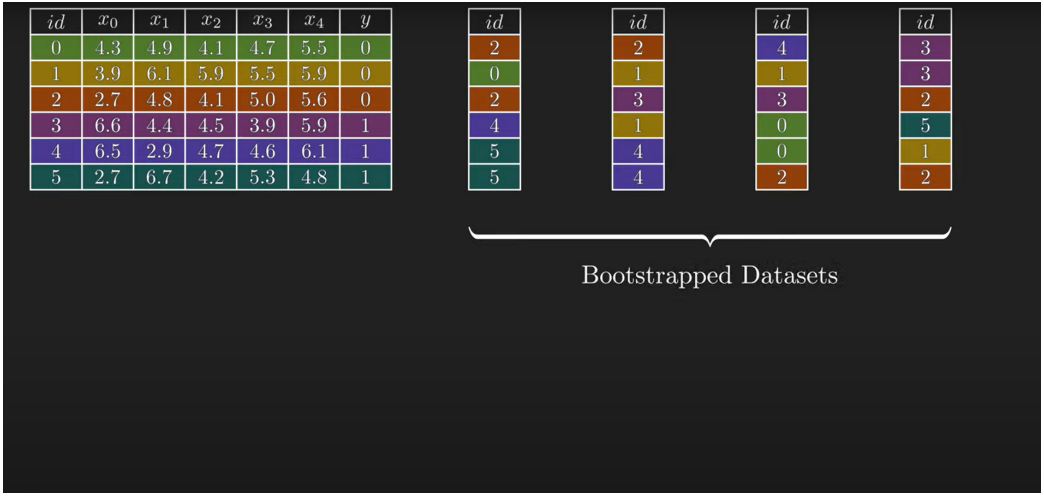
id
2
0
2
4
5
5

id
2
1
3
1
4
4

id
4
1
3
0
0
2

id
3
3
2
5
1
2

Random Forest



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

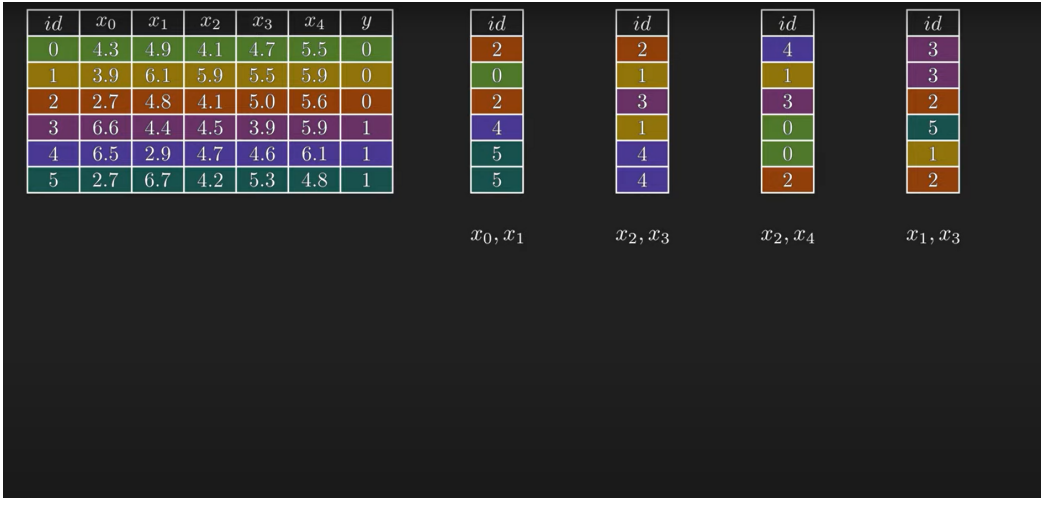
id
2
1
3
1
4
4

id
4
1
3
0
0
2

id
3
3
2
5
1
2

x_0, x_1

Random Forest



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

x_0, x_1

id
2
1
3
1
4
4

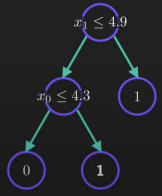
x_2, x_3

id
4
1
3
0
0
2

x_2, x_4

id
3
3
2
5
1
2

x_1, x_3



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

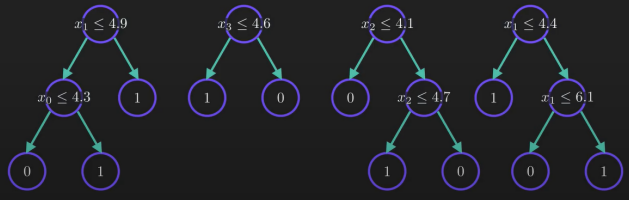
id
2
0
2
4
5
5

id
2
1
3
1
4
4

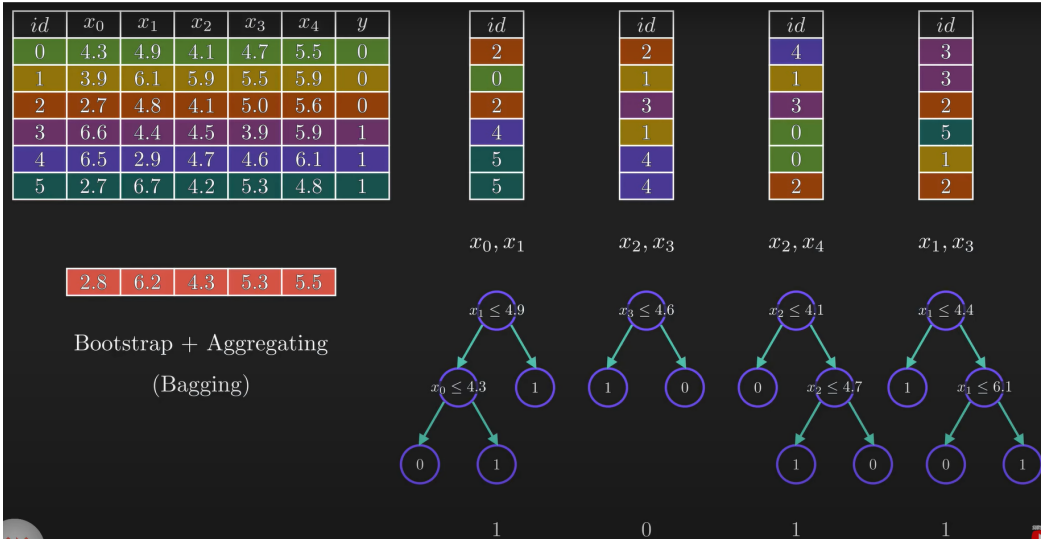
id
4
1
3
0
0
2

id
3
3
2
5
1
2

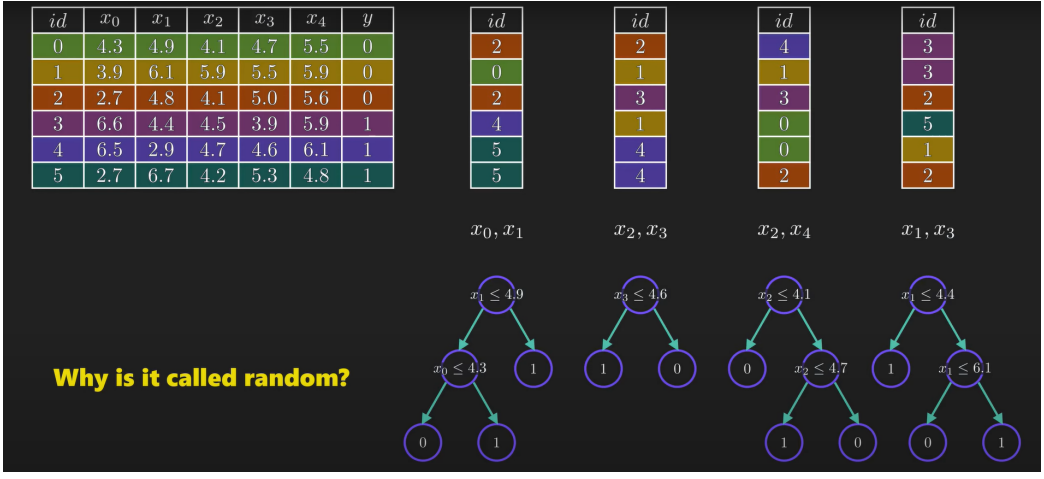
x_0, x_1 x_2, x_3 x_2, x_4 x_1, x_3



Random Forest



Random Forest



Random Forest

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

x_0, x_1

id
2
1
3
1
4
4

x_2, x_3

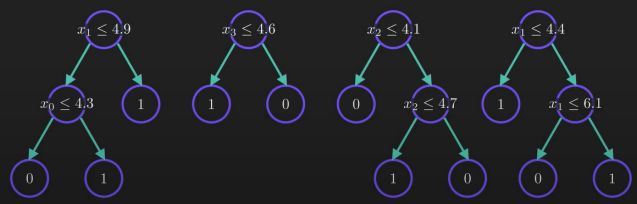
id
4
1
3
0
0
2

x_2, x_4

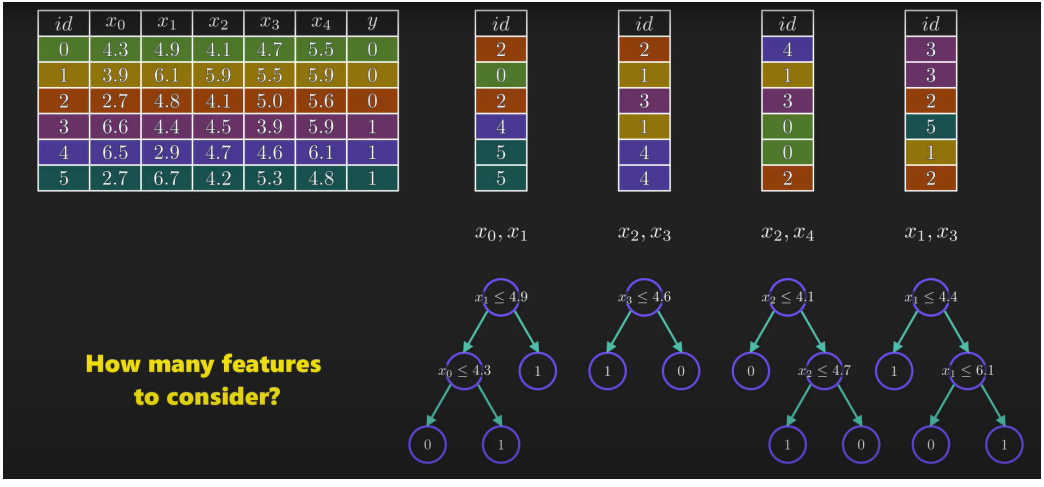
id
3
3
2
5
1
2

x_1, x_3

Why Bootstrapping and Feature Selection?



Random Forest



Random Forest - Concluding Remarks

From ISL2 we have :

- Random forests provide an improvement over bagged trees by way of a small tweak that decorrelates the trees. This reduces the variance when we average the trees.
- But when building these decision trees, each time a split in a tree is considered, **a random selection of m predictors is chosen** as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors.
- A fresh selection of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$, that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors

Causal trees and Causal Regression

Similarities

- Divide data en subgroups in order to maximize the discriminatory power of the split
- Both uses recursive binary splitting (greedy aproach) into covariate space
- Solve over-fitting problem through Croos- Validation (CV) to determine the depth of the tree

Causal trees and Causal Regression

Differences

- Goal: estimate treatment effect heterogeneity
- Optimize sub-groups to estimate treatment effect heterogeneity
- Divide population into subgroups to minimize MSE in treatment effects **(objective function)** instead in outcomes
- Use honesty tree, additionally, to address over-fitting.

How to address missing counterfactual

- Objective function is unfeasible $\sum_i[(\tau_i - \hat{\tau}(X_i))^2]$ (true treatment effect unobserved)

- Transform Outcome (Athey and Imbens, 2016)
- Applying off-the-shelf prediction to estimate CATE
- p is the assignment probability.

$$Y_i^* = Y_i/p, \ \ if \ \ W_i = 1, \ \ and \ \ let \ \ Y_i^* = -Y_i/(1-p) \ \ if \ \ W_i = 0 \tag{1}$$

- $\mathbb{E}[Y_i^*|X_i = x] = \tau(x)$ is noisy for individual treatment effect, but unbiased estimate of CATE
- Use estimated propensity score $\hat{p}$ or AIPW score as outcome in observational studies.

Honest Causal Tree

- Half sample to estimates tree and another half to estimate treatment effect.
- **Tradeoff**
- COST: sample splitting means build shallower tree, less personalized predictions and lower MSE of treatment effects
- BENEFIT: valid confidence intervals with coverage rates that do not deteriorate as data generating process get more complex or more covariates are added.
- **Inference**
- Can separate tree construction from treatment effect estimation
- tree constructed on training is independent of test sample
- Holding tree form training sample fixed, can use standard methods to conduct inference within leaf of the tree on test sample

Partition and Leaf Effects Estimates

- Three samples: model tree construction S^{tr} , estimation sample for leaf effects S^{est} and a test sample S^{te}.
- Sample average treatment effect in sample S^{est} for the leaf $l(X_i, \Pi)$ associated with covariates X_i

$$\hat{\tau}(X_i, S^{est}, \Pi) = \frac{1}{\sum_{j \in S^{est} \cap l(X_i, \Pi)} W_i} \sum_{i \in S^{est} \cap l(X_i, \Pi)} W_i Y_i \; -$$

$$\frac{1}{\sum_{j \in S^{est} \cap l(X_i, \Pi)} (1 - W_i)} \sum_{i \in S^{est} \cap l(X_i, \Pi)} (1 - W_i) Y_i$$

Estimating the MSE of treatment effects (EMSE)

- criterion for evaluating a partition Π anticipating re-estimating leaf effects using sample splitting:

$$MSE(S^{est}, S^{te}) = \frac{1}{n^{te}} \sum_{i \in S^{te}} (\tau_i - \hat{\tau}(X_i, S^{est}, \Pi))^2 \tag{2}$$

$$MSE(S^{est}, S^{te}) = \frac{1}{n^{te}} \sum_{i \in S^{te}} (\tau_i^2 - 2\tau_i * \hat{\tau}(X_i, S^{est}, \Pi) + \hat{\tau}^2(X_i, S^{est}, \Pi)) \tag{3}$$

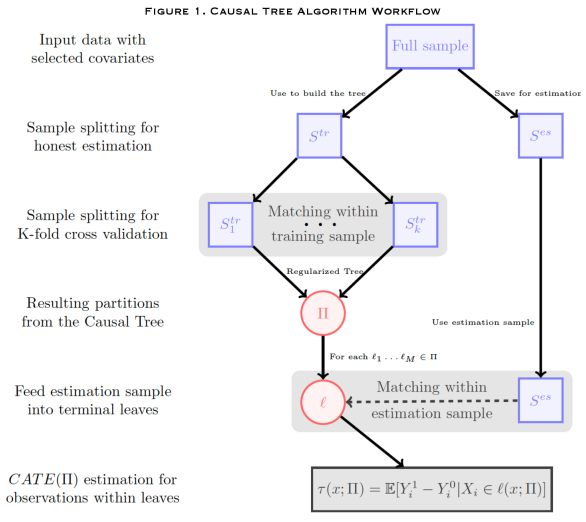
$$EMSE = E_{S^{est}, S^{te}}[MSE(S^{est}, S^{te})] \tag{4}$$

$$EMSE = \mathbf{V}_{S^{est}, S^{te}}[\hat{\tau}(X_i; \Pi, S^{est}] - E_{X_i}[\tau^2(X_i; \Pi)] + E[\tau_i^2] \tag{5}$$

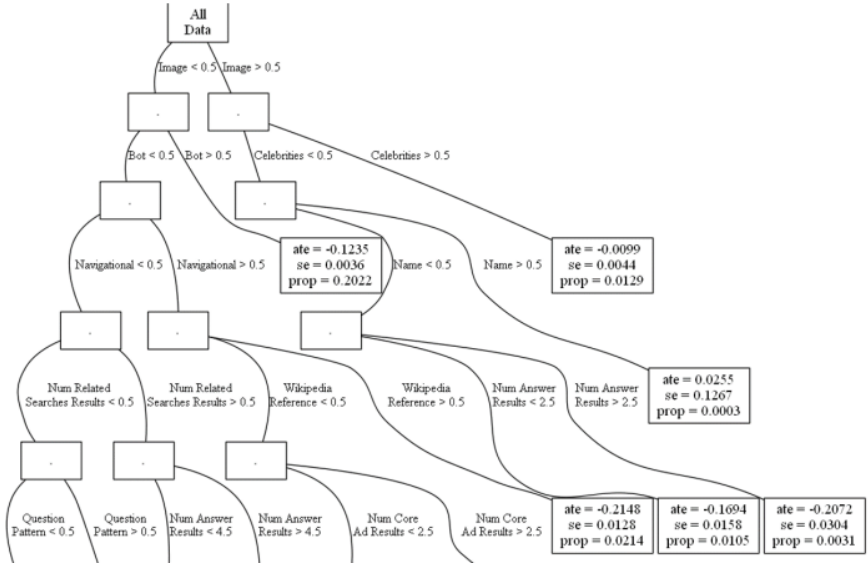
Causal Tree Algorithm

- Divide data into tree-building S^{tr} and estimation S^{est} samples
- Use recursively partition in covariate space X to deep partition Π
  - Each split is selected as the one that minimizes mean - square error estimation of treatment effect over all possible binary splits
  - **Preserve minimun number of treated and control units in each child leaf**
- Use cross - validation to select the depth d^* of the partition
- Select partition Π^* by pruning Π to depth d^*
- Estimate the treatment effects in each leaf of Π^* using estimation sample

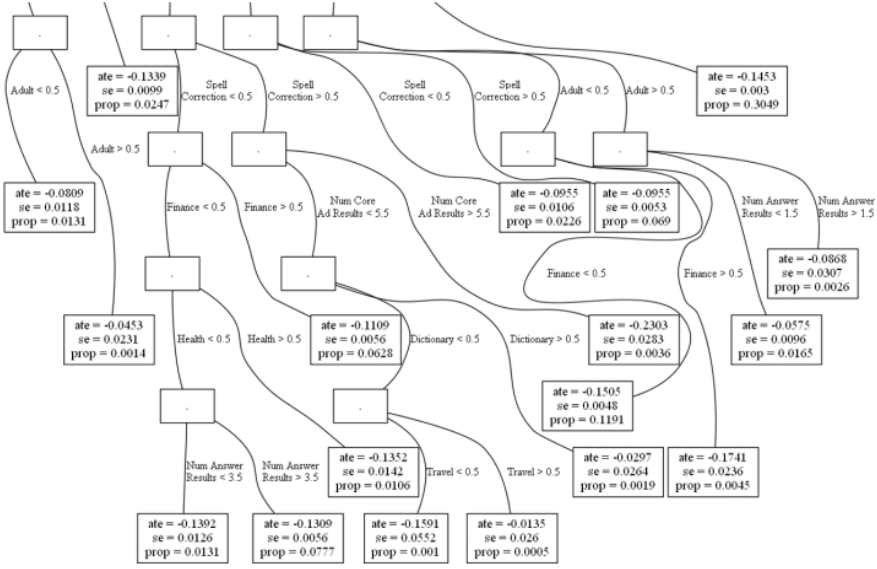
Causal Tree Algorithm



Causal Tree - Example

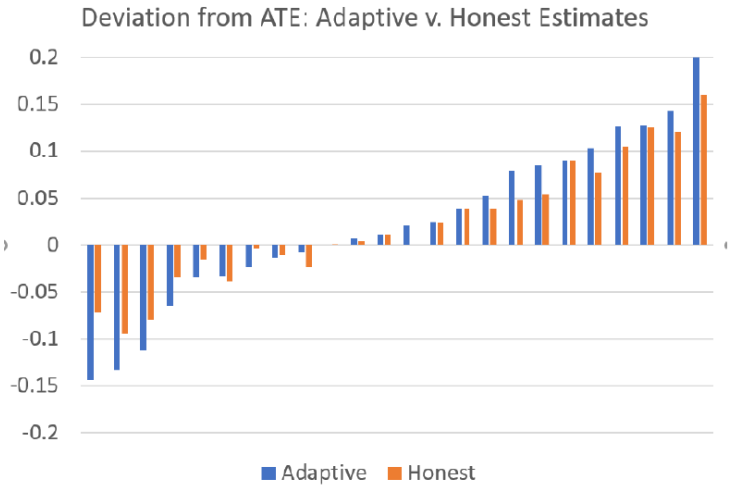


Causal Tree - Example



Estimating treatment effect: adaptive vs honest Causal Forest

- adaptive: same sample to build the tree (Π) and estimate treatment effect
- adaptive method creates deviation from ATE



Causal Tree - Example 2

The General Social Survey is an extensive survey, collected since 1972, that seeks to measure demographics, political views, social attitudes, etc. of the U.S. population.

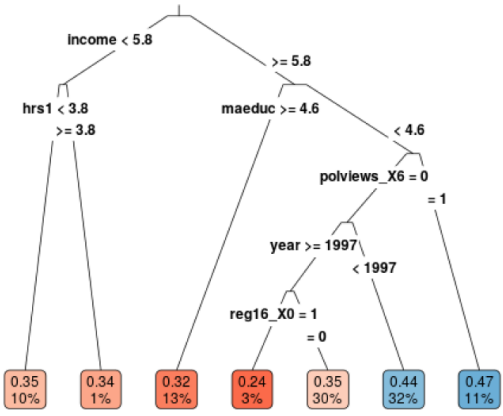
Of particular interest to us is a **randomized experiment**, for which we have data between 1986 and 2010.

- **Question A**: Are we spending too much, too little, or about the right amount on **welfare**?
- **Question B**: Are we spending too much, too little, or about the right amount on **assistance to the poor**?

Treatment effect: how much less likely are people to answer **too much** to question B than to question A.

- We want to understand how the treatment effect depends on **covariates**: political views, income, age, hours worked, ...

Causal Tree - Example 2




```
Linear hypothesis test

Hypothesis:
leaf1:W - leaf2:W = 0
leaf1:W - leaf3:W = 0
leaf1:W - leaf4:W = 0
leaf1:W - leaf5:W = 0
leaf1:W - leaf6:W = 0
leaf1:W - leaf7:W = 0

Model 1: restricted model
Model 2: Y ~ leaf + W:leaf

  Res.Df Df      F    Pr(>F)
1   5272
2   5266  6 4.4771 0.0001575 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

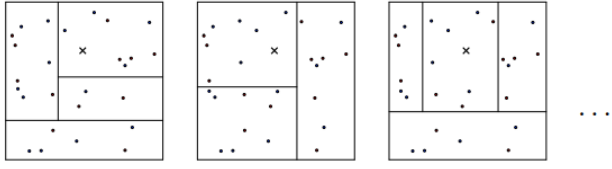
From Trees to Random Forest

- Training set $(X_i, Y_i, W_i)_{i=1}^n$, while tree predictor for a test point (x)

$$\hat{\tau} = T(x; X_i, Y_i, W_{ii=1}^n) \tag{6}$$

- Random Forest: build and average many different trees T^*
- Create alternative tress (T_b^*) by bagging (sampling with replacement) or sub-sampling the training set

$$\hat{\tau} = \frac{1}{B} \sum_{b=1}^{B} T_b^*(x; X_i, Y_i, W_{i i=1}^n) \tag{7}$$



Statistical inference with regression forest

Regresssion forest are asymptotically Gaussian and centered:

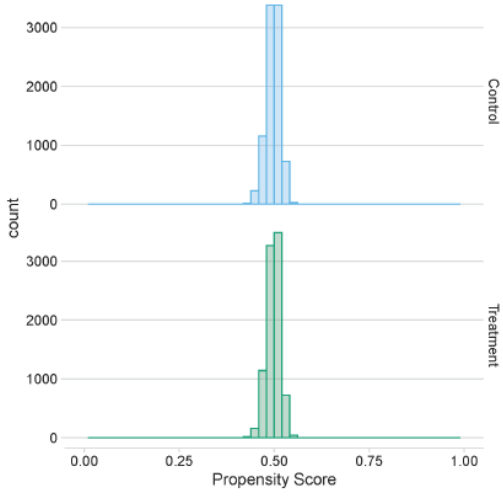
$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \rightarrow \mathbb{N}(0,1), \sigma_n^2(x) \rightarrow_p 0 \tag{8}$$

technical conditions

- Individual trees are honest (**Honesty**)
- Individual trees built random sub-samples of size s n^{β}, where $\beta_{min} < \beta < 1$ (**Subsampling**)
- X_i density from 0 and ∞ (**Continuos features**)
- Conditional mean function $\mu(x) = \mathbb{E}[Y|X = x]$ is Lipschitz continues (**Lipschitz response**)

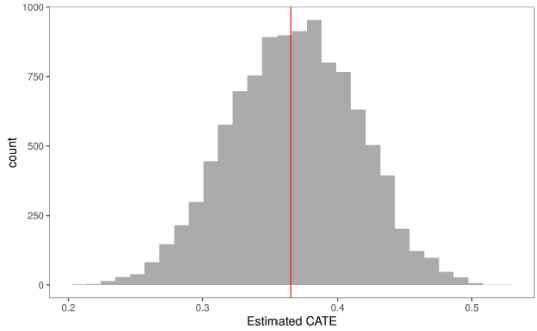
Application: General Social Survey

Figure: Verifying randomization - balance



Application: General Social Survey

Figure: Out-of-bag conditional CATE



Application: General Social Survey

Figure: Quantifying heterogeneity

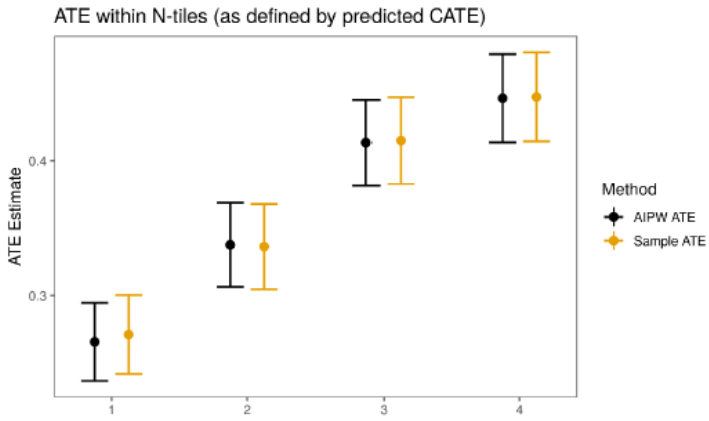
1. Best Linear Predictor (Chernozhukov, Demier, Duflo, and Fernandez-Val, 2018)

```
test_calib_orthog <- grf::test_calibration(cf)

        Best linear fit using forest predictions (on held-out data)
        as well as the mean forest prediction as regressors, along
        with one-sided heteroskedasticity-robust (HC3) SEs:

                                  Estimate Std. Error t value    Pr(>t)
        mean.forest.prediction         0.995229   0.021511 46.2670 < 2.2e-16 ***
        differential.forest.prediction 1.579928   0.164924  9.5797 < 2.2e-16 ***
        ---
        Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Chernozhukov, Demier, Duflo and Fernandez-Val, 2018)

Application: General Social Survey

Figure: Quantifying heterogeneity



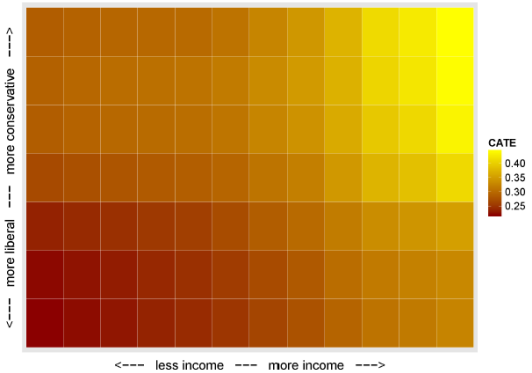
Application: General Social Survey

Figure: bottom vs top deciles

	1	10	p.overall
	N=1057	*N=1056*	
age	3.05 (1.08)	3.43 (0.82)	<0.001
income	6.35 (1.67)	7.23 (0.34)	<0.001
educ	5.68 (1.15)	4.76 (0.72)	<0.001
polviews_X2	0.38 (0.49)	0.02 (0.13)	<0.001
polviews_X3	0.19 (0.39)	0.05 (0.22)	<0.001
polviews_X4	0.26 (0.44)	0.40 (0.49)	<0.001
polviews_X5	0.03 (0.17)	0.23 (0.42)	<0.001
polviews_X6	0.06 (0.23)	0.22 (0.41)	<0.001
polviews_other.values	0.08 (0.27)	0.09 (0.28)	0.634
sex_X1	0.49 (0.50)	0.61 (0.49)	<0.001

Application: General Social Survey

A causal forest analysis uncovers **strong treatment heterogeneity**



Random forest vs Locally linear forest

Forest weaknesses: economic variables have smooth relationships (i.e U shape), forest fit a line as a step function (very inefficient)

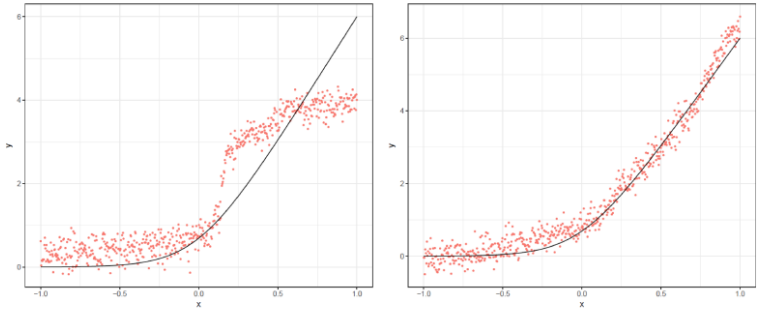
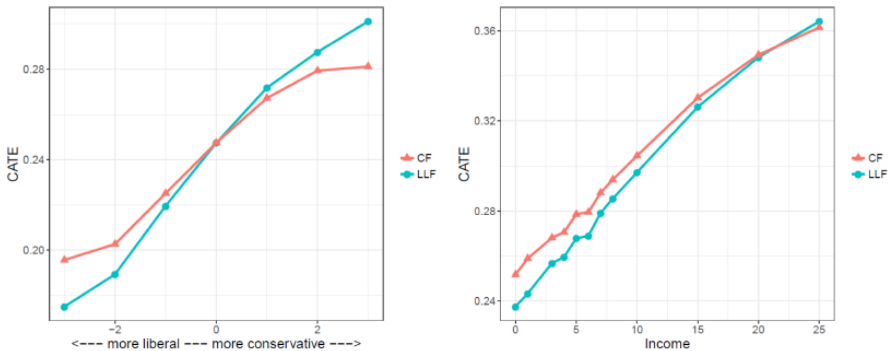


Figure 1: Predictions from random forests (left) and locally linear forests (right) on 600 test points. Training and test data were simulated from equation (1), with dimension $d = 20$ and errors $\epsilon \sim N(0, 20)$. Forests were trained also on $n = 600$ training points and tuned via cross-validation. Here the true conditional mean signal $\mu(x)$ is in black, and predictions are shown in red.

Causal random forest vs Causal locally linear forest



Locally linear regression with ridge penalty

$$\begin{pmatrix} \hat{\mu}(x) \\ \hat{\theta}(x) \end{pmatrix} = \arg\min_{\mu,\theta} \sum_{i=1}^{n} \alpha_i(x)(Y_i - \mu(x) - (X_i - x)\theta(x))^2 + \lambda||\theta(x)||_2^2 \tag{9}$$

$$\begin{pmatrix} \hat{\mu}(x) \\ \hat{\theta}(x) \end{pmatrix} = (X^T A X + \lambda J)^{-} 1 X^T A Y \tag{10}$$

Weights are determine from forest a la GRF, accounting for regression in splitting for efficiency.

Causal Forest example

- The Retirement Reform increased the early retirement age (ERA) gradually by (1/2) year annually from 2014 for cohorts born after 1954
- Descriptive evidence of treatment effect heterogeneity

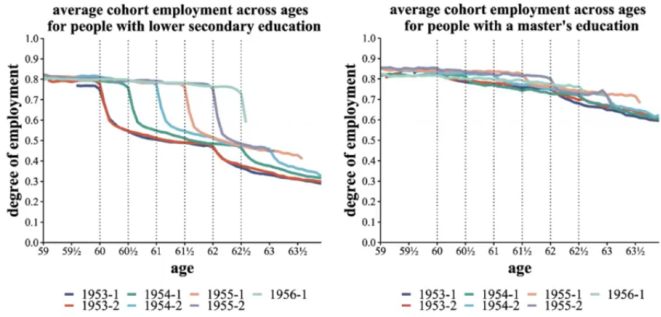


Figure: Average cohort employment for different ages by education level

Causal Forest example

- Descriptive evidence of treatment effect heterogeneity

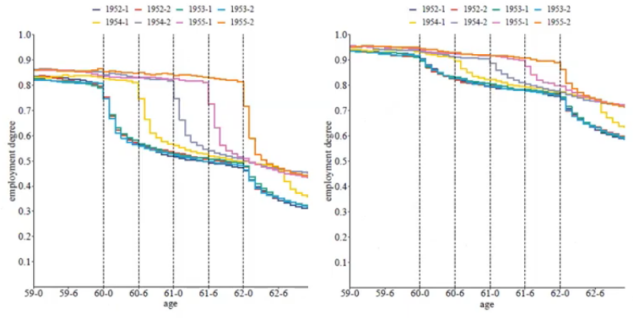


Figure: Sample split by median income from 50-60. Left: below median.

Causal Forest vs OLS

- Machine learning method get a better fit to the sign of treatment effect heterogeneity

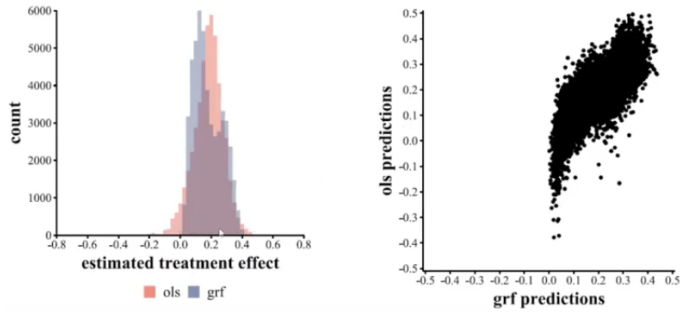


Figure: Distribution of predicted treatment effects

Causal Forest example

- Causal forest finds significant treatment effect heterogeneity as evaluated ”out of bag”

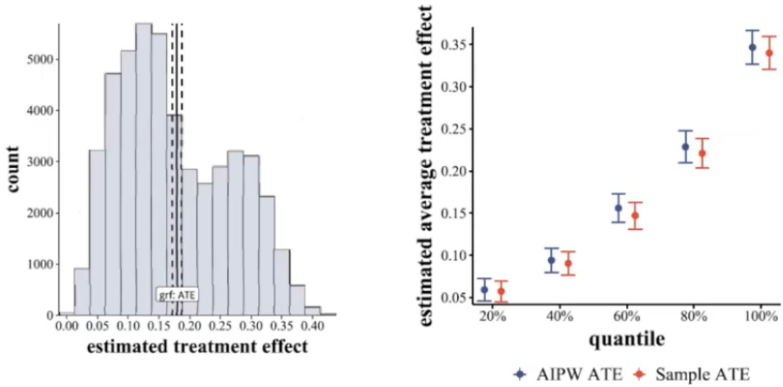


Figure: Distribution of estimated out-of-bag treatment effects

Causal Forest example

- Average values of covariates for different quantiles of estimated treatment effects