Causal Forest

Alexander Quispe

February 15, 2022

Citation

These notes are based on the Lecture Notes of Susan Athey Machine Learning and Causal Inference Course - 2019.

Baseline method: k-NN matching

Consider the k-**NN matching** estimator for $\tau(x)$:

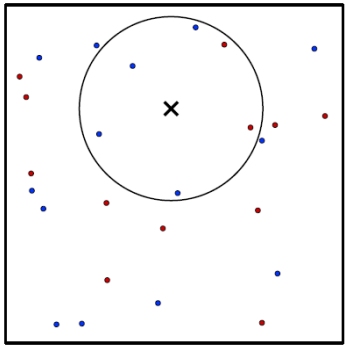
$$\hat{\tau}(x) = \frac{1}{k} \sum_{\mathcal{S}_1(x)} Y_i - \frac{1}{k} \sum_{\mathcal{S}_0(x)} Y_i,$$

where $\mathcal{S}_{0/1}(x)$ is the set of k-nearest cases/controls to x. This is consistent given **unconfoundedness** and regularity conditions.

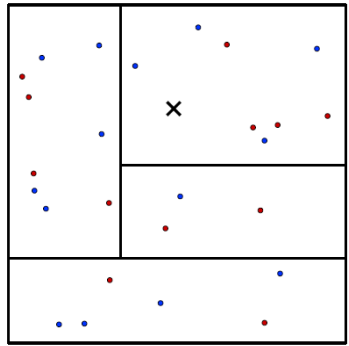
- **Pro:** Transparent asymptotics and good, robust performance when p is small.
- **Con:** Acute curse of dimensionality, even when $p = 20$ and $n = 20k$.

NB: Kernels have similar qualitative issues as k-NN.

Making k-NN matching adaptive



Euclidean neighborhood, for k-NN matching.



Tree-based neighborhood.

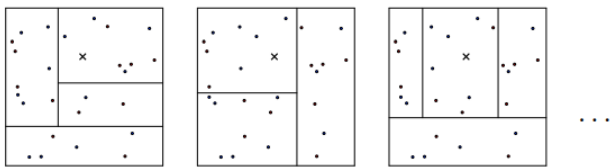
From Trees to Random Forest

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

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

- 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

$$\hat{\tau} = \frac{1}{B} \sum_{b=1}^{B} T_b^*(x; X_i, Y_i, W_{i i=1}^n) \tag{2}$$



Statistical inference with regression forest

Regresssion forest are asymptotically Gaussian and centered:

$$\frac{\hat{\mu}_n(x) - \mu(x)}{\sigma_n(x)} \rightarrow \mathbb{N}(0,1), \sigma_n^2(x) \rightarrow_p 0 \tag{3}$$

technical conditions

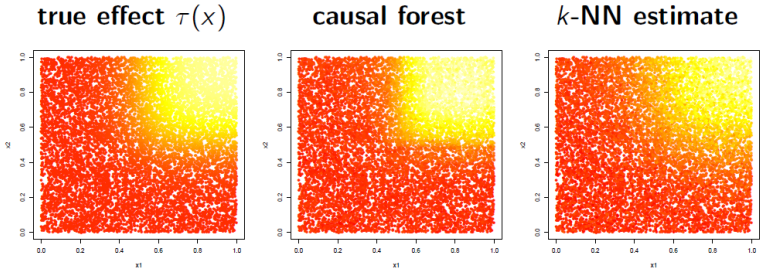
- Individual trees are honest (**Honesty**)
- Individual trees built random sub-samples of size s n^{β}, where $\beta_{min} < \beta < 1$ (**Subsampling**)
- X_i density from 0 and ∞ (**Continuos features**)
- Conditional mean function $\mu(x) = \mathbb{E}[Y|X = x]$ is Lipschitz continues (**Lipschitz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

Causal Forest Example

Figure: True effect, Causal Forest, KNN estimate

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 6$; treatment assignment is random. All **the signal is concentrated along two features**.

The plots below depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.

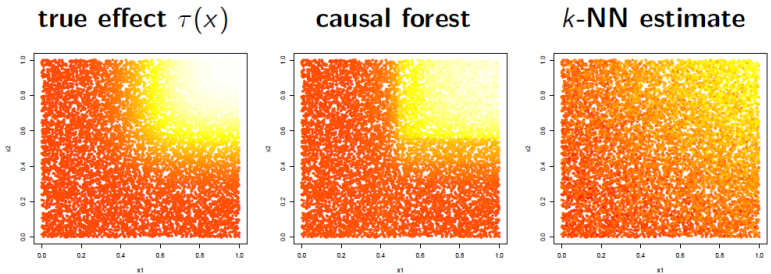


Causal Forest Example

Figure: True effect, Causal Forest, KNN estimate

We have $n = 20$k observations whose features are distributed as $X \sim U([-1, 1]^p)$ with $p = 20$; treatment assignment is random. All **the signal is concentrated along two features**.

The plots below depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.



Application: General Social Survey

Figure: Randomized Experiment

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, ...

Application: General Social Survey

Figure: Code 1

```
## Propensity model
W.hat.mod <- grf::regression_forest(X = as.matrix(select(train_df, -Y, -W))
                                    , Y = train_df$W
                                    , num.trees = 200
                                    , ci.group.size = 1
                                    , tune.parameters = "all")

W.hat.rf <- W.hat.mod$predictions
## Outcome model
Y.hat.mod <- grf::regression_forest(X = as.matrix(select(train_df, -Y, -W))
                                    , Y = train_df$Y
                                    , num.trees = 200
                                    , ci.group.size = 1
                                    , tune.parameters = "all")
Y.hat.rf <- Y.hat.mod$predictions
```

Application: General Social Survey

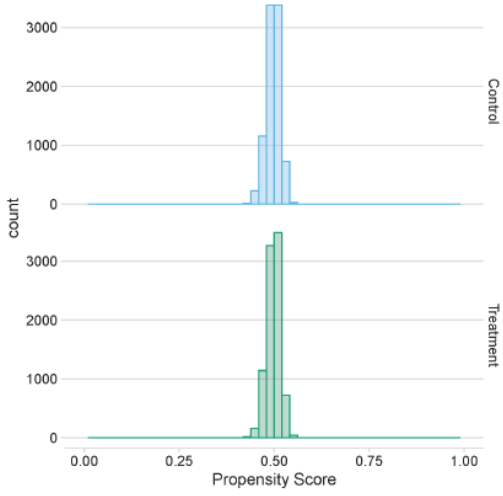
Figure: Code 2

Implement causal forest using `grf`.

```
cf <- grf::causal_forest(
  X = as.matrix(select(train_df, -Y, -W)),
  Y = train_df$Y,
  W = train_df$W,
  Y.hat = Y.hat.rf,
  W.hat = W.hat.rf,
  num.trees=200)
```

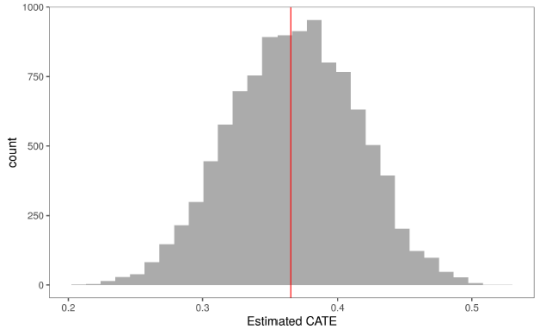
Application: General Social Survey

Figure: Verifying randomization - balance



Application: General Social Survey

Figure: Out-of-bag conditional CATE



Application: General Social Survey

Figure: Quantifying heterogeneity

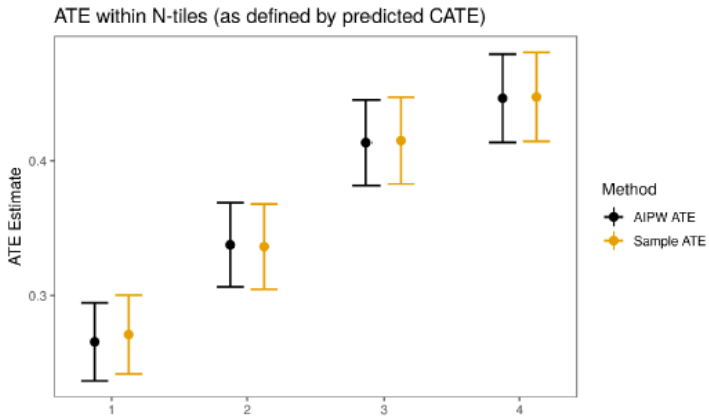
1. Best Linear Predictor (Chernozhukov, Demier, Duflo, and Fernandez-Val, 2018)

```
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



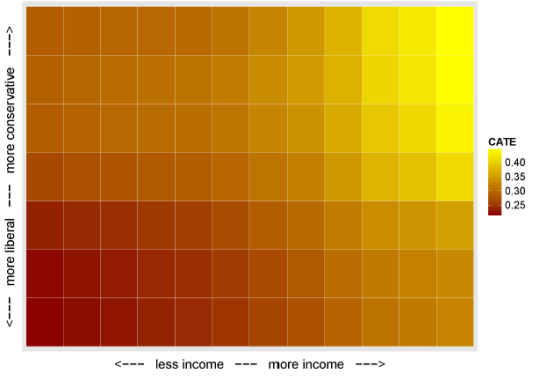
Application: General Social Survey

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)

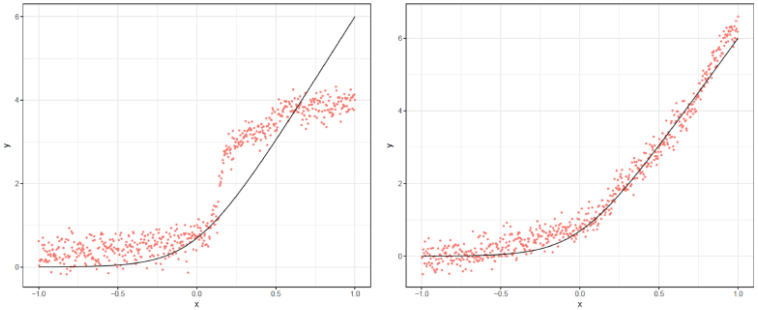
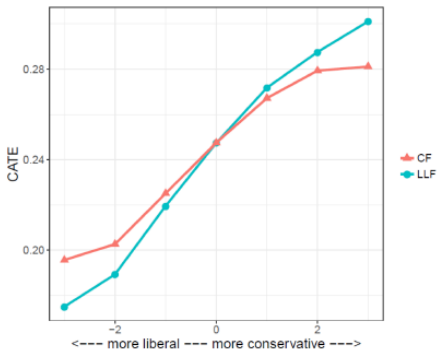
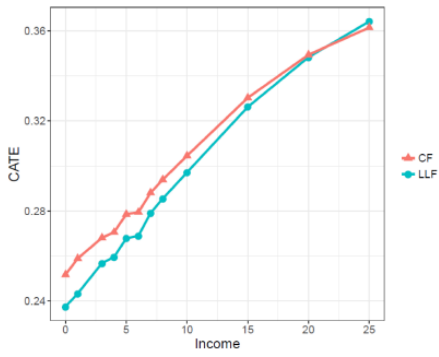


Figure 1: Predictions from random forests (left) and locally linear forests (right) on 600 test points. Training and test data were simulated from equation (1), with dimension $d = 20$ and errors $\epsilon \sim N(0, 20)$. Forests were trained also on $n = 600$ training points and tuned via cross-validation. 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

$$\begin{pmatrix} \hat{\mu}(x) \\ \hat{\theta}(x) \end{pmatrix} = \arg\min_{\mu,\theta} \sum_{i=1}^{n} \alpha_i(x)(Y_i - \mu(x) - (X_i - x)\theta(x))^2 + \lambda||\theta(x)||_2^2 \tag{4}$$

$$\begin{pmatrix} \hat{\mu}(x) \\ \hat{\theta}(x) \end{pmatrix} = (X^T A X + \lambda J)^{-} 1 X^T A Y \tag{5}$$

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

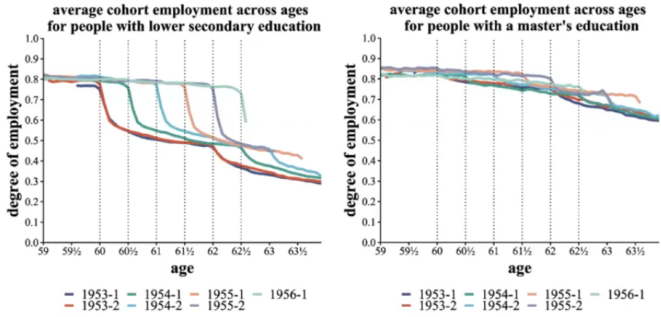


Figure: Average cohort employment for different ages by education level

Causal Forest example

- Descriptive evidence of treatment effect heterogeneity

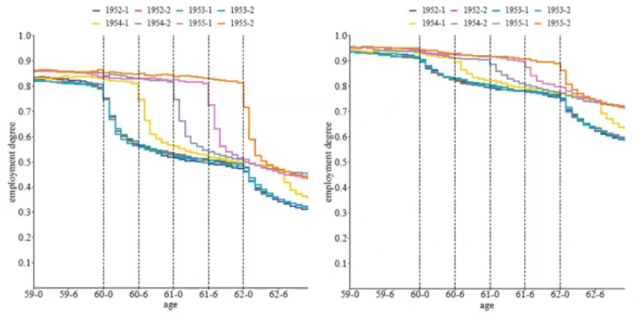


Figure: Sample split by median income from 50-60. Left: below median.

Causal Forest vs OLS

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

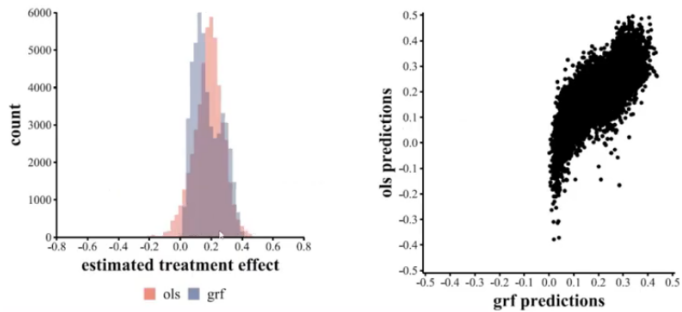


Figure: Distribution of predicted treatment effects

Causal Forest example

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

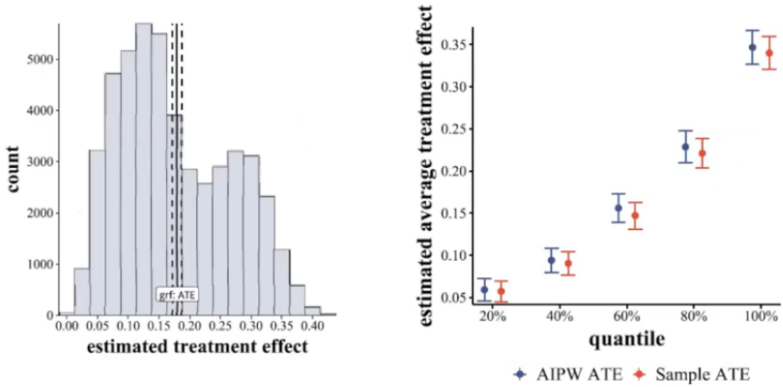
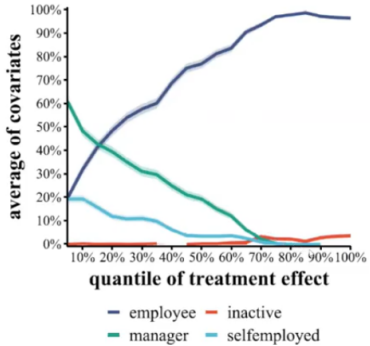
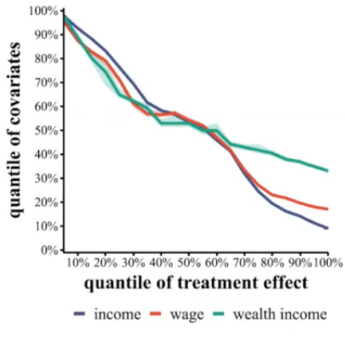


Figure: Distribution of estimated out-of-bag treatment effects

Causal Forest example

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





Causal Forest Application: Labor Markets and Crime

Davis, Jonathan M.V., and Sara B. Heller. 2017. ”Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs.” American Economic Review, 107 (5): 546-50.

Causal Forest Application

motivation

- Treatment heterogeneity provides valuable information to improve program targeting and figure out what mechanisms drive results

Justification

- Contribution to youth program design
- Understand how different subpopulations respond to youth program

Objective

- Apply Causal Forest algorithm to data from two large-scale randomized control trials of the summer jobs program.
- Test to identify distinct treatment heterogeneity

Causal Forest Application

Chicago´s One Summer Plus (OSP) program conducted in 2012 - 2013.

- Target subpopulation: disadventage youth ages 14-22
- The program Provided 25 hours a week of employment. Program paid Chicago´s minimun wage ($8.25)
- The program Provided an Adult mentor and other programming

Causal Forest Application

Pooled data with 6850 observation

Year	Treatment (T)	Control (C)	Total
2012	790	904	1634
2013	2634	2582	5216

Outcome of interest

Variable	Description	Number of observation
Crime	The number of violent-crime arrests within two years of random assignment	N = 6850
Employment	An indicator for ever being employed during the six quarters after the program	N = 4894

¿How splits are made?

- Standard MSE in a context of Causal effects is not feasible.
- Athey and imbens (2016) show that minimizing the expected MSE of predicted treatment effect, rather than infeasible MSE, is equivalent to maximizing the variance of treatment effects across leaves minus a penalty for within-leaf variance

"Objective function O".

$$O = (n_T + n_C)\hat{\tau}_l^2 - 2\left(\frac{\hat{Var}(Y_{Tl})}{n_T} + \frac{\hat{Var}(Y_{Cl})}{n_C}\right) \tag{6}$$

Implementing Road Map

Steps	Description
(i)	Draw a subsample b without replacement $n_b = 0.2N$
(ii)	n_{tr} training sample and n_e estimation sample ($\frac{n_{tr}}{2} = \frac{n_e}{2} = n_b$)
(iii)	1. Calculate objective function on unpartitioned data O 2. Randomly select v covariates as candidates for a split 3. Unique value of each covariate $X_j = x$ form a candidate split ($X_j \leq x$ *and* $X_j > x$) 4. At least 10 treatment and 10 control observations in both leaves. Calculate $O' = O_{left} + O_{right}$ 5. One candidate split if $O' > O$, then implement the single split that maximizes O'. 6. Repeat this process in each child leaf. Terminal leaf where no split increase O.

Implementing Road Map

Steps	Description
(iv)	Group n_e observations into the same tree based on their *Xs*
(v)	Using estimation sample to calculate $\hat{\tau}_l = \bar{y}_{Tl} - \bar{y}_{Cl}$ within each terminal leaf.
(vi)	Save prediction $\hat{\tau}_{l,b} = \hat{\tau}_l$ to each observation in full sample whose *Xs* would place it in leaf *l*
(vii)	Repeat steps (i) and (vi) B = 25 000 times *l*.
(viii)	- Predictive CATE for observation (i) : $\hat{\tau}_i^{CF}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_{l,b}$. - Average across trees for observation

Parameters

- Three parameters: number of trees, minimun number of treatment and control observations on each leaf and subsample size.
- Number of trees: A large number of trees to reduce Monte Carlo error
- Increase minimun observation in each leaf trades off bias and variance, bigger leaves makes results consistent across different samples but predict less heterogeneity.
- Smaller sub-sample reduce dependence across trees but increase variance of each estimate

De - correlating trees

- Random subset of covariates at each split ensures that a few string predictors are not used over and over again in the same way across trees
- Use square root of the total number of covariates $v = \sqrt{|X|}$
- De-correlating trees improve predictions

Parameters

- Three parameters: number of trees, minimun number of treatment and control observations on each leaf and subsample size.
- Number of trees: A large number of trees to reduce Monte Carlo error
- Increase minimun observation in each leaf trades off bias and variance, bigger leaves makes results consistent across different samples but predict less heterogeneity.
- Smaller sub-sample reduce dependence across trees but increase variance of each estimate

Unconfoundedness assumption

- Treatment probabilities vary across blocks, then we have to condition on randomization block.
- Use inverse probability weights in calculations of treatment effects and variances.

$$weight_i = \frac{T_i}{P_{block(i)}} + \frac{1 - T_i}{1 - P_{block(i)}} \tag{7}$$

- $P_{block(i)}$ is the probability of being treated in observations i's block
- T_i is the treatment assigment

Test of the Predictions

- Test to identify distinct heterogeneous treatment effects by subgroups
- Including observations used in building tree or estimation might cause fail conclusions to identify distinct heterogeneous treatment effects

Application:

- Split 6850 observations in a half to create in and out of sample groups S_{in} and S_{out}.
- Run causal forest using S_{in} and make predictions for both sample
- For each sample predictions, group by positive $(\hat{\tau}_i^{CF}(x) > 0)$ or negative treatment effect $(\hat{\tau}_i^{CF}(x)) < 0$

Test of the Predictions

- Estimate separate treatment effects for these two subgroups by regressing each outcome on the indicator $\mathbf{1}[\hat{\tau}_j^{CF}(x) > 0]$, $T_i \ x \ \mathbf{1}[\hat{\tau}_i^{CF}(x) > 0]$, $T_i \ x \ (1 - \mathbf{1}[\hat{\tau}_i^{CF}(x) > 0])$, baseline covariates used in CF and block fixed effects.
- Null Hypothesis: the treatment effects are equal across the two subgroups.
- Adjusted in sample: in step (viii), average across trees in which observation was not part of either the tree-building or estimation sample.

Treatment effects by predictive response

Subgroup	No. of violent crime arrests	Any formal employment
Panel A. In sample		
$\hat{\tau}_i^{CF}(x) > 0$	0.22	0.19
	(0.05)	(0.03)
$\hat{\tau}_i^{CF}(x) < 0$	−0.05	−0.14
	(0.02)	(0.03)
H_0: subgroups equal, $p =$	0.00	0.00
Panel B. Out of sample		
$\hat{\tau}_i^{CF}(x) > 0$	−0.01	0.08
	(0.05)	(0.03)
$\hat{\tau}_i^{CF}(x) < 0$	−0.02	−0.01
	(0.02)	(0.03)
H_0: subgroups equal, $p =$	0.77	0.02
Panel C. Adjusted in sample		
$\hat{\tau}_i^{CF}(x) > 0$	−0.06	0.05
	(0.04)	(0.03)
$\hat{\tau}_i^{CF}(x) < 0$	−0.02	−0.04
	(0.02)	(0.03)
H_0: subgroups equal, $p =$	0.41	0.02

Treatment effects by predictive response

- The Causal Forest successfully identifies two subgroups with distinct employment effects but not in violence impacts, which could happen for a few reasons:

1. The treatment effects may not vary with observed covariates
2. The greedy algorithm may fail to identify the true functional form of the treatment effect, or our subgroup test may not isolate the true form of the heterogeneity
3. Treatment heterogeneity could be obscured by sampling error; the CF may need bigger datasets.