# Lecture 23:
# Bagging, Random Forests, & Causal Trees

## Big Data and Machine Learning for Applied Economics
## Econ 4676

Ignacio Sarmiento-Barbieri

Universidad de los Andes

November 3, 2020

# Agenda

1. Recap
   - Bagging and Random Forests
   - Comparisons: Lasso, CART, Random Forests
2. Causal Trees
   - Causality Review: ATE, CATE, HTE
   - Heterogeneous Treatment Effects
   - Empirical Example
3. Review & Next Steps
4. Further Readings

# CART

$$y = f(x) + u$$

- Smart way to represent nonlinearities. Most relevant variables on top.
- Very easy to communicate.
- Reproduces human decision-making process.
- Trees are intuitive and do OK, but
  - They are not very good at prediction
  - If the structure is linear, CART does not work well.
  - Not very robust

# Bagging

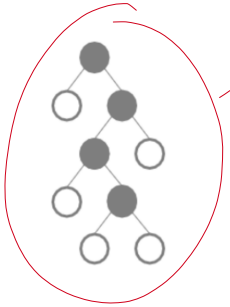
- We can improve performance a lot using either bootstrap aggregation (bagging), random forests, or boosting.
- Bagging & Random Forests:
  - Repeatedly draw bootstrap samples $(X_i^b, Y_i^b)_{i=1}^N$ from the observed sample.
  - For each bootstrap sample, fit a regression tree $\hat{f}^b(x)$
    - Bagging: full sample
    - Random Forests: subset of predictors $\sqrt{(p)}$ (breaks high correlation)
  - Average across bootstrap samples to get the predictor

$$\hat{f}_{bag} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x) \tag{1}$$

  - Basically we are smoothing predictions.
  - Idea: the variance of the average is less than that of a single prediction.

# Random Forests

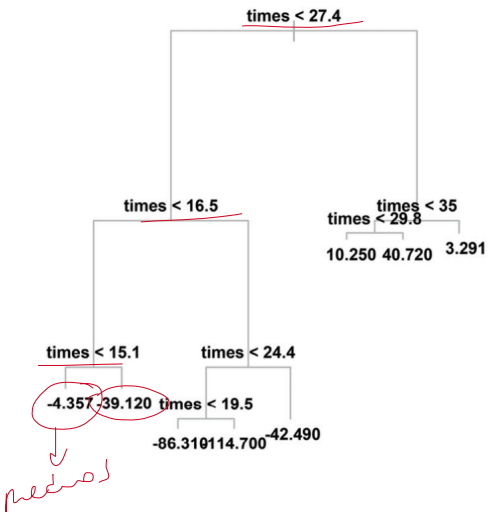
Trees:

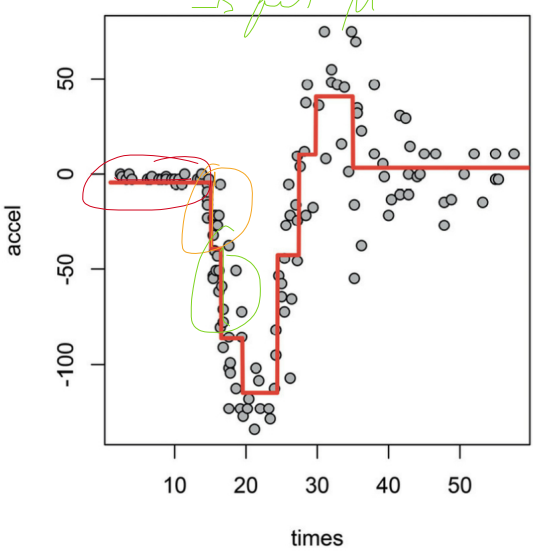


- overfit
- Unbiased
- High variance

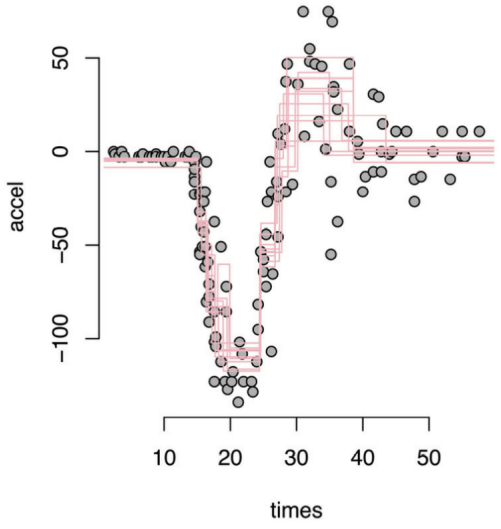
Random Forests:

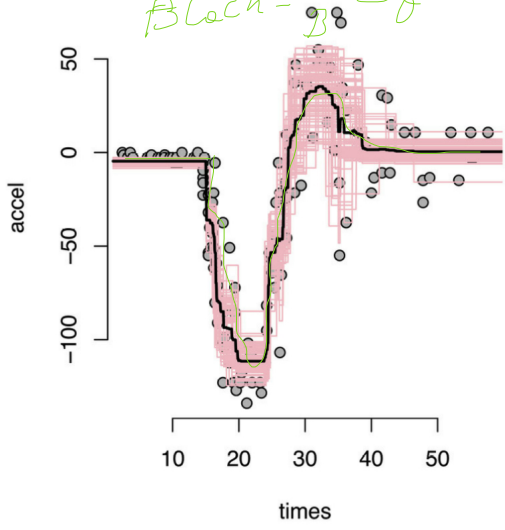
# Random Forests



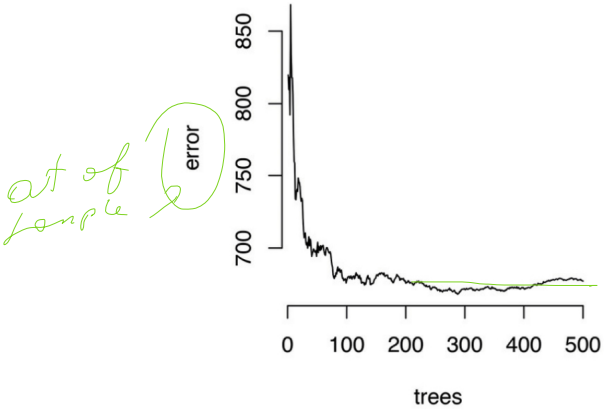



# Random Forests

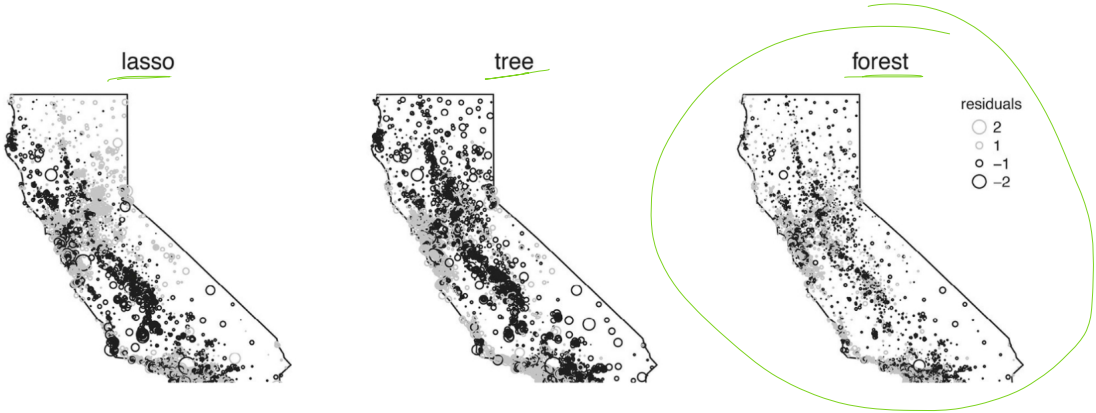


# Random Forests

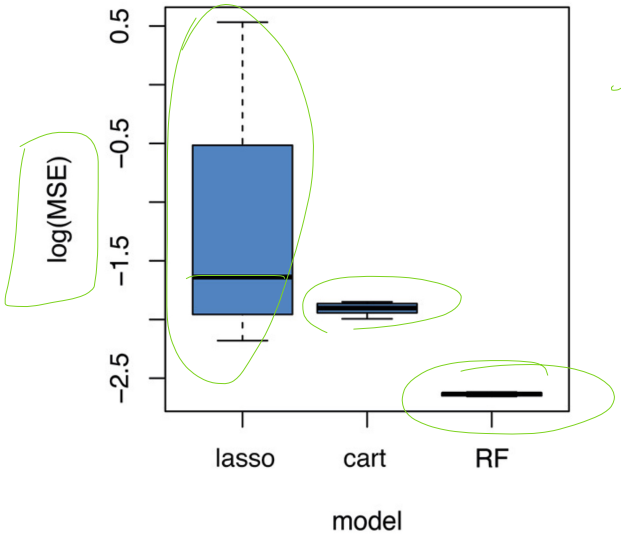


# In sample residuals




```
# Model Matrix for Lasso
XXca <- model.matrix(logMedVal ~ longitude*latitude,
      data=data.frame(scale(CAhousing)))[,-1]
```

# Out of sample MSE



# Treatment Effects

- We observe a sequence of triples $\{(W_i, Y_i, X_i)\}_i^N$, where
  - $W_i \in \{0, 1\}$: is a binary variable indicating whether the individual was treated (1) or not (0)
  - $Y_i^{obs} \in \mathbb{R}$: a real variable indicating the observed outcome for that individual
  - $X_i$: is a $p$-dimensional vector of observable pre-treatment characteristics
- Moreover, in the Neyman-Rubin potential-outcomes framework, we will denote by
  - $Y_i(1)$: the outcome unit $i$ would attain if they received the treatment
  - $Y_i(0)$: the outcome unit $i$ would attain if they were part of the control group

# Treatment Effects

The **individual treatment effect** for subject $i$ can then be written as

$$Y_i(1) - Y_i(0)$$

Unfortunately, in our data we can only observe one of these two potential outcomes.

| $X_i$ | $Y_i(0)$ | $Y_i(1)$ |
|---|---|---|
| $X_1$ | . | $Y_1(1)$ |
| $X_2$ | . | $Y_2(1)$ |
| $X_3$ | $Y_3(0)$ | . |
| ... | ... | ... |
| $X_n$ | $Y_n(0)$ | . |

Using the potential outcome notation above, the observed outcome can also be written as

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$$

# Average Treatment Effects

- Computing the difference for each individual is impossible.
- But we will try to use the information we have about the distribution of the data to say something about its average,
- This is called the **average treatment effect (ATE)** and denoted here by $\tau$:

$$\tau := E[Y_i(1) - Y_i(0)] \tag{2}$$

# Average Treatment Effects

- For this to work we need a couple of assumptions
  - the data is independently and identically distributed (*iid*)
  - the potential outcome is independent of the treatment:

$$Y_i(1), Y_i(0) \perp W_i$$

  - we are assuming that whether or not a subject received the treatment has nothing to do with how they would respond to this "treatment".
  - in other words, treatment assignment is random.

# Average Treatment Effect (ATE)

- The independence assumption above allows us to produce a simple estimator for the ATE:

$$\tau = E[\tau_i] = E[Y_i(1) - Y_i(0)] \tag{3}$$
$$= E[Y_i(1)] - E[Y_i(0)] \quad \because \text{Linearity of expectations} \tag{4}$$
$$= E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 0] \quad \because \text{Independence assumption} \tag{5}$$

- To know the estimate of the average treatment effect we just need to know the average $Y_i$ for treated and control subjects and compute their difference.
The implied estimator is:

$$\hat{\tau} = \frac{1}{n_1} \sum_{i|W_i=1} y_i - \frac{1}{n_0} \sum_{i|W_i=0} y_i \tag{6}$$

where $n_1$ and $n_0$ are the numbers of subjects in the treatment and control groups, respectively.

# Heterogeneous Treatment Effects

- Heterogeneous Treatment Effects: Same treatment may affect different individuals differently
- Conditional Average Treatment Effect (CATE)

$$\tau(x) := E[Y_i(1) - Y_i(0)|X_i = x] \tag{7}$$

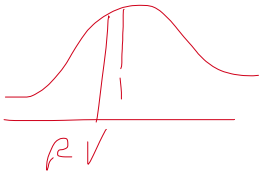
- **Causal Tree (Athey and Imbens, 2016):** A data-driven approach to partition the data into subpopulations that differ in the magnitude of their treatment effects. The approach enables the construction of valid confidence intervals for treatment effects.
- Exploring treatment heterogeneity can provide valuable information about how to improve program targeting and what mechanisms drive results.

# Heterogeneous Treatment Effects

## Motivation I: Experiments and Data-Mining

- Issues:
  - Ad hoc searches for particularly responsive subgroups may mistake noise for a true treatment effect.
  - Concerns about ex-post "data-mining" or p-hacking
    - preregistered analysis plan can protect against claims of data mining
    - But may also prevent researchers from discovering unanticipated results and developing new hypotheses
- But how is researcher to predict all forms of heterogeneity in an environment with many covariates?
- Athey and Imbens to the rescue
  - Allow researcher to specify set of potential covariates
  - Data-driven search for heterogeneity in causal effects with valid standard errors

# Heterogeneous Treatment Effects

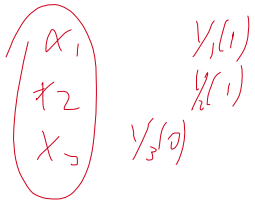


- Before proceeding we need to make a couple of assumptions
- Assumption 1: Unconfoundedness

$$Y_i(1), Y_i(0) \perp W_i \mid X_i \tag{8}$$

- The *unconfoundedness* assumption states that, once we condition on observable characteristics, the treatment assignment is independent to how each person would respond to the treatment.
- i.e., the rule that determines whether or not a person is treated is determined completely by their observable characteristics.
- This allows, for example, for experiments where people from different genders get treated with different probabilities,
- **rules out** experiments where people self-select into treatment due to some characteristic that is not observed in our data.

# Heterogeneous Treatment Effects



- Assumption 2: Overlap

$$\forall\, x \in \text{supp}\,(X), \qquad 0 < P\,(W = 1 \mid X = x) < 1 \tag{9}$$

  - The *overlap* assumption states that at every point of the covariate space we can always find treated and control individuals.
  - i.e., in order to estimate the treatment effect for a person with particular characteristics $X_i = x$, we need to ensure that we are able to observe treated and untreated people with those same characteristics so that we can compare their outcomes.

# Causal Trees: Empirical Example (Green and Kern)

- To illustrate how it works let me use this experiment from the General Social Survey (GSS)
- GSS conducts surveys regular surveys on Americans think feel about different issues
- For decades, scholars studying Americans' support for social welfare spending have noted the special disdain that americans harbor for programs labeled "welfare"
- This phenomenon became the subject of sustained experimental inquiry in the mid-1980s, when the GSS included a question-wording experiment in its national survey of adults.

# Causal Trees: Empirical Example

- Respondents in each survey were randomly assigned to one of two questions about public spending.
- *"too much" money is spent on assistance to the Poor (control) or Welfare (treatment)*
- Various explanations put forward: stereotypes associated with welfare recipients and poor people, particularly racial stereotypes, and to political orientations such as individualism and conservatism .
- Some authors consider the interaction between the treatment and attributions, e.g.
  - Federico (2004) examines a complicated three-way interaction between the treatment, education, and racial perceptions.
  - Jacoby (2000) suggests that party and ideology may make some respondents especially receptive to the more specific program (should strong and weak Democrats be treated as separate subgroups or should they be combined?)

# Causal Trees

```
#load packages
library(fBasics)      # Summary statistics
library(rpart)        # Classification and regression trees
library(rpart.plot)   # Plotting trees
library(treeClust)    # Predicting leaf position for causal trees
library(car)          # linear hypothesis testing for causal tree
library(kableExtra)   # Tables
library(causalTree) # For causal trees (Athey and Imbens, 2016)
library(dplyr) # For data wrangling
# Set seed for reproducibility
set.seed(201911)
# Load Data
df<-readRDS("welfare.rds")
str(df)
```

```
'data.frame':   13198 obs. of  34 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Y        : num  0 0 1 1 1 0 0 0 1 0 ...
 $ W        : num  1 1 1 0 0 1 1 0 0 1 ...
 $ hrs1     : num  40 35 30 40 35 38 27 40 32 50 ...
 $ partyid  : num  4 1 2 2 1 2 0 1 3 3 ...
 $ income   : num  12 12 12 12 11 12 12 11 12 12 ...
 ...
```

# Causal Trees

## ATE

```r
difference_in_means <- function(dataset) {
  treated_idx <- which(dataset$W == 1)
  control_idx <- which(dataset$W == 0)

  # Filter treatment / control observations, pulls outcome variable as a vector
  y1 <- dataset[treated_idx, "Y"] # Outcome in treatment grp
  y0 <- dataset[control_idx, "Y"] # Outcome in control group

  n1 <- sum(df[,"W"])     # Number of obs in treatment
  n0 <- sum(1 - df[,"W"]) # Number of obs in control

  # Difference in means is ATE
  tauhat <- mean(y1) - mean(y0)

  # 95% Confidence intervals
  se_hat <- sqrt( var(y0)/(n0-1) + var(y1)/(n1-1) )
  lower_ci <- tauhat - 1.96 * se_hat
  upper_ci <- tauhat + 1.96 * se_hat

  return(c(ATE = tauhat, lower_ci = lower_ci, upper_ci = upper_ci))
}

tauhat_rct <- difference_in_means(df)
tauhat_rct
```

```
       ATE   lower_ci   upper_ci
-0.3697802 -0.3841123 -0.3554481
```

# Causal Trees

## ATE

```
outcome_variable_name <- "Y"
treatment_variable_name <- "W"
covariate_names <- c("hrs1", "partyid", "income", "rincome",
                     "wrkstat", "wrkslf","age", "polviews",
                     "educ", "earnrs", "race","wrkslf",
                     "marital","sibs","childs", "occ80",
                     "prestg80", "indus80","res16","reg16",
                     "mobile16", "family16", "parborn",
                     "maeduc","degree","sex","race",
                     "born","hompop","babies",
                     "preteen","teens","adults")

fmla <- paste("Y ~ W +",paste(covariate_names, collapse = " + "))
print( fmla)
```

```
[1] "Y ~ W + hrs1 + partyid + income + rincome + wrkstat + wrkslf + age
+ polviews + educ + earnrs + race + wrkslf + marital + sibs + childs
+ occ80 + prestg80 + indus80 + res16 + reg16 + mobile16 + family16
+ parborn + maeduc + degree + sex + race + born + hompop + babies + preteen + teens + adults"
```

# Causal Trees

## ATE

```
reg_simple<-lm(Y~W,data=df)
reg_controls<-lm(fmla,data=df)
stargazer::stargazer(reg_simple,reg_controls,type="latex")
```

Table 1

| | *Dependent variable:* | |
|---|---|---|
| | Y | |
| | (1) | (2) |
| W | −0.370*** | −0.368*** |
| | (0.007) | (0.007) |
| Constant | 0.481*** | 0.223*** |
| | (0.005) | (0.069) |
| Controls | No | Yes |
| Observations | 13,198 | 13,198 |
| $R^2$ | 0.166 | 0.215 |

# Causal Trees

## HTE

- We need to proceed in steps
- Step 1: Split the dataset. Why? $\rightarrow$ Athey and Imbens innovation
  - In order to ensure valid estimates of the treatment effect within each subgroup, Athey and Imbens propose a sample-splitting approach that they refer to as honesty:
  - a method is honest if it uses one subset of the data to estimate the model parameters, and a different subset to produce estimates given these estimated parameters.
  - In the context of causal trees, honesty implies that the asymptotic properties of treatment effect estimates within leaves are the same as if the tree partition had been exogenously given, and it is one of the assumptions required to produce unbiased and asymptotically normal estimates of the treatment effect.

# Causal Trees

## HTE

- Divide the data 40%-40%-20% for honest estimation and validation.

```
train_fraction <- 0.80  # Use train_fraction % of the dataset to train our models

df_train <- sample_frac(df, replace=F, size=train_fraction)
df_test <- anti_join(df,df_train, by = "ID")#need to check on larger datasets


split_size <- floor(nrow(df_train) * 0.5)
df_split <- sample_n(df_train, replace=FALSE, size=split_size)

# Make the splits
df_est <- anti_join(df_train,df_split, by ="ID")
```

# Causal Trees

- Step 2: Fit the tree

```
fmla_ct <- paste("factor(Y) ~", paste(covariate_names, collapse = " + "))

ct_unpruned <- honest.causalTree(
  formula = fmla_ct,              # Define the model
  data = df_split,                # Subset used to create tree structure
  est_data = df_est,              # Which data set to use to estimate effects

  treatment = df_split$W,         # Splitting sample treatment variable
  est_treatment = df_est$W,       # Estimation sample treatment variable

  split.Rule = "CT",              # Define the splitting option
  cv.option = "TOT",              # Cross validation options
  cp = 0,                         # Complexity parameter

  split.Honest = TRUE,            # Use honesty when splitting
  cv.Honest = TRUE,               # Use honesty when performing cross-validation

  minsize = 25,                   # Min. number of treatment and control cases in each leaf
  HonestSampleSize = nrow(df_est)) # Num obs used in estimation after building the tree
```

# Causal Trees

- Step 3: Crossvalidate

```
# Table of cross-validated values by tuning parameter.
ct_cptable <- as.data.frame(ct_unpruned$cptable)
# Obtain optimal complexity parameter to prune tree.
selected_cp <- which.min(ct_cptable$xerror)
optim_cp_ct <- ct_cptable[selected_cp, "CP"]
# Prune the tree at optimal complexity parameter.
ct_pruned <- prune(tree = ct_unpruned, cp = optim_cp_ct)
ct_pruned
```

```
n= 5279

node), split, n, deviance, yval
      * denotes terminal node

 1) root 5279 912.78610 -0.3753160
   2) partyid>=1.5 3530 654.60930 -0.3822570
     4) polviews>=3.5 2826 532.11460 -0.3997024
       8) reg16>=0.5 2658 500.98290 -0.4043439
        16) hrs1>=44.5 1063 203.85490 -0.4271320
          32) wrkslf< 1.5 182  35.70208 -0.4264330
            64) indus80< 526 81  15.81056 -0.3757764 *
            65) indus80>=526 101  19.59552 -0.4610849 *
          33) wrkslf>=1.5 881 167.91090 -0.4283712 *
        17) hrs1< 44.5 1595 295.01770 -0.3896395 *
       9) reg16< 0.5 168  30.09444 -0.3315372 *
     5) polviews< 3.5 704 115.34030 -0.3167446 *
```

# Causal Trees

- Step 4: Predict point estimates (on estimation sample)

```
tauhat_ct_est <- predict(ct_pruned, newdata = df_est)
head(tauhat_ct_est)
```

```
         1          2          3          4          5          6
-0.3843850 -0.4283712 -0.3843850 -0.3896395 -0.3896395 -0.3843850
```

# Causal Trees

- Step 5: Compute standard errors
- The `causalTree` package does not compute standard errors by default, but we can compute them using the following trick.
  - First, define $L_l$ to indicate assignment to leaf $l$
  - Second, consider the following linear model.

$$Y = \sum_l L_l \alpha_l + W L_l \beta_l \tag{10}$$

  - The interaction coefficients in this regression recover the average treatment effects in each leaf, since

$$E[Y|W=1, L=1] - E[Y|W=0, L=1] = (\alpha_1 + \beta_1) - \alpha_1 = \beta_1 \tag{11}$$

  - Therefore, the standard error around the coefficients is also the standard error around the treatment effects.
  - We will also use these statistics to test hypothesis about leaf estimates.

# Causal Trees

```
# Create a factor column 'leaf' indicating leaf assignment
num_leaves <- length(unique(tauhat_ct_est))  #There are as many leaves as there are predictions
df_est$leaf <- factor(tauhat_ct_est, labels = seq(num_leaves))
# Run the regression
ols_ct <- lm(as.formula("Y ~ 0 + leaf + W:leaf"), data= df_est)
ols_ct_summary <- summary(ols_ct)
```

Table 2: Average treatment effects per leaf

| | Estimate | Std. Error |
|---|---|---|
| leaf1:W | -0.4611 | 0.0817 |
| leaf2:W | -0.4284 | 0.0276 |
| leaf3:W | -0.3896 | 0.0205 |
| leaf4:W | -0.3844 | 0.0214 |
| leaf5:W | -0.3758 | 0.0920 |
| leaf6:W | -0.3315 | 0.0633 |
| leaf7:W | -0.3167 | 0.0309 |
| leaf8:W | -0.2124 | 0.0497 |

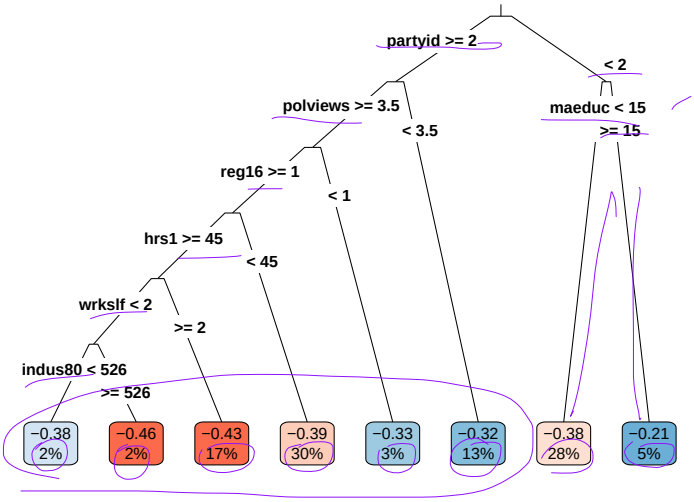
# Causal Trees

- Step 6: Predict point estimates (on test set)

```
tauhat_ct_test <- predict(ct_pruned, newdata = df_test)


rpart.plot(
  x = ct_pruned,          # Pruned tree
  type = 3,               # Draw separate split labels for the left and right directions
  fallen = TRUE,          # Position the leaf nodes at the bottom of the graph
  leaf.round = 1,         # Rounding of the corners of the leaf node boxes
  extra = 100,            # Display the percentage of observations in the node
  branch = 0.1,           # Shape of the branch lines
  box.palette = "RdBu") # Palette for coloring the node
```

# Causal Trees



# Causal Trees

```
# Null hypothesis: all leaf values are the same
hypothesis <- paste0("leaf1:W = leaf", seq(2, num_leaves), ":W")
ftest <- linearHypothesis(ols_ct, hypothesis, test="F")

kable_styling(kable(data.frame(ftest, check.names = FALSE, row.names = NULL)[2,],
                    "latex", digits = 4,
                    caption="Testing null hypothesis: Average treatment effect is same across leaves"),
              bootstrap_options=c("striped", "hover", "condensed", "responsive"),
              full_width=FALSE)
```

Table 3: Testing null hypothesis: Average treatment effect is same across leaves

| | Res.Df | RSS | Df | Sum of Sq | F | Pr($> F$) |
|---|---|---|---|---|---|---|
| 2 | 5263 | 884.921 | 7 | 3.4114 | 2.8984 | 0.005 |

# Review & Next Steps

- Bagging and Random Forests
- Comparisons: Lasso, CART, Random Forests
- Causality Review: ATE, CATE, HTE
- Heterogeneous Treatment Effects Empirical Example
- Next class: More on causal trees, and causal forests
- Questions? Questions about software?

# Further Readings


- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
- Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Public opinion quarterly, 76(3), 491-511.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
- Taddy, M. (2019). Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions. McGraw Hill Professional.