Lec 4

CART

- Regression problem with (X_1, X_2) as inputs and Y as continuous response
- Recursive binary partition of space
  - Model Y by the mean of each space
  - Choice of variable and split point to achieve the best split
  - Recursively continue till some stopping rule

$$\hat{f}(x) = \sum_{m=1}^{5} c_m I((X_1, X_2) \in R_m)$$

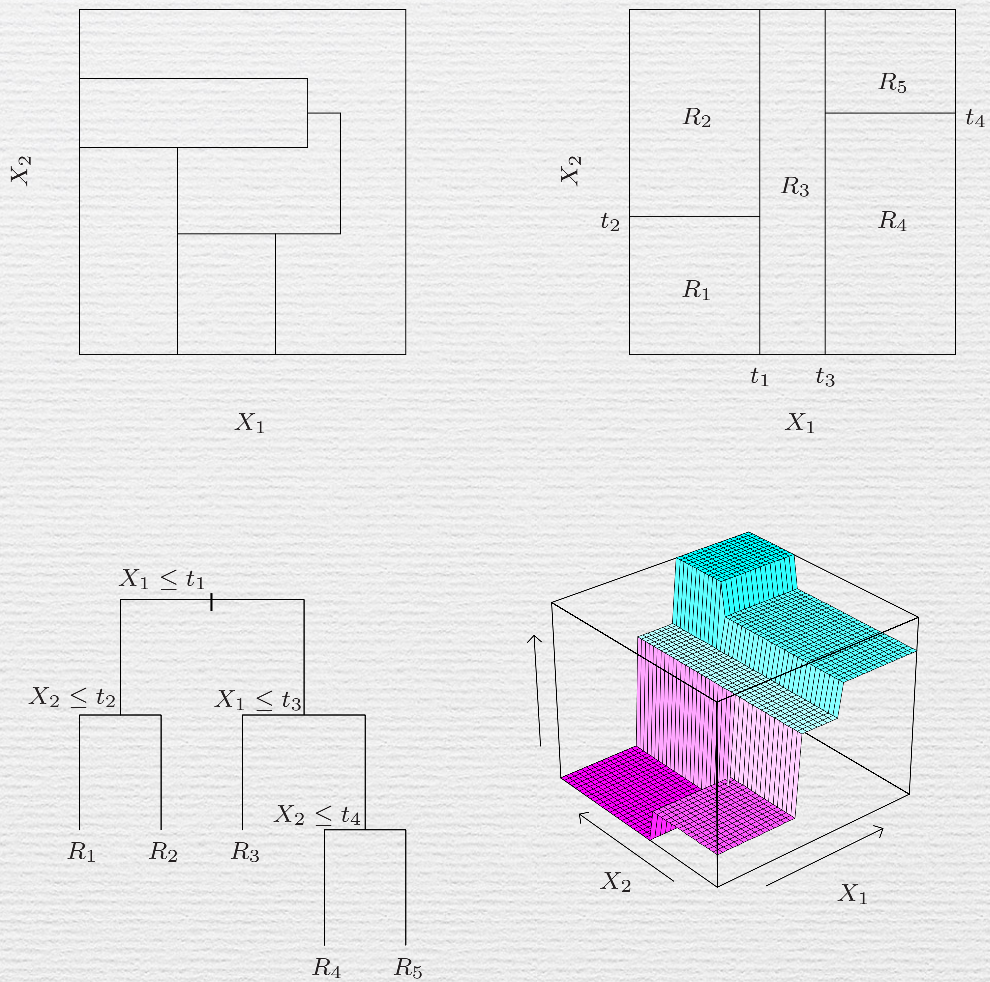
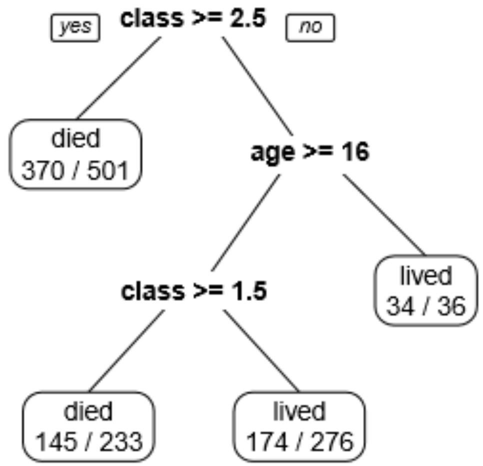


FIGURE 9.2. *Partitions and CART. Top right panel shows a partition of a two-dimensional feature space by recursive binary splitting, as used in CART, applied to some fake data. Top left panel shows a general partition that cannot be obtained from recursive binary splitting. Bottom left panel shows the tree corresponding to the partition in the top right panel, and a perspective plot of the prediction surface appears in the bottom right panel.*

Example: Who survived the Titantic?



- A key advantage of recursive binary tree is interpretability

Fit a regression tree

- N data points (x_i, y_i), where x_i is P dimensional
- MSE as fit criterion
  - Minimize C over c_m, R_m
  - Finding optimal R_m is computational infeasible
  - Adopt a greedy algorithm
  - Consider a variable *j* and split point *s*, define half planes

$$\hat{f}(x) = \sum_{m=1}^{M} c_m I((X_1, X_2) \in R_m)$$

$$C = \sum (y_i - \hat{f}(x_i))$$

$$\hat{c}_m = ave(y_i | x_i \in R_m)$$

$$R_1(j, s) = \{X | X_j \leq s\} \ and \ R_2(j, s) = \{X | X_j > s\}$$

$$min_{j,s} \ [\ min_{c_1} \sum_{x \in R_1(j,s)} (y - c_1)^2 + min_{c_2} \sum_{x \in R_2(j,s)} (y - c_2)^2 \ ]$$

Stopping, Pruning

- A large tree very every data point is leaf is clear overfit
- Tree size is a model's complexity and the optimal tree should be adaptively chosen from the data
- Grow a large tree, stopping when minimum node size (say 5) is reached
- Prune the tree based on tree size |T|, number of nodes in node N_m, MSE Q_m
    - Successively collapse the internal node that produces the smallest per-node increase in first sum below

$$C = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$$

Classification tree

- In a node m, representing a region R_m with N_m observations let p_mk represent the promotion of class k in node m

- Gini Index differentiable — measure of variance

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k)$$

$$\begin{aligned} &\text{Misclassification error:} && \tfrac{1}{N_m} \textstyle\sum_{i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}. \\ &\text{Gini index:} && \textstyle\sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}). \\ &\text{Cross-entropy or deviance:} && -\textstyle\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}. \end{aligned} \tag{9.17}$$

Spam email

- 4601 email, 48 quantitive features such as address, internet, 6 quantitive features characters that match, ….

TABLE 9.3. *Spam data: confusion rates for the 17-node tree (chosen by cross-validation) on the test data. Overall error rate is 9.3%.*

	Predicted	
True	`email`	`spam`
`email`	57.3%	4.0%
`spam`	5.3%	33.4%

TABLE 9.1. *Test data confusion matrix for the additive logistic regression model fit to the spam training data. The overall test error rate is 5.5%.*

	Predicted Class	
True Class	`email` (0)	`spam` (1)
`email` (0)	58.3%	2.5%
`spam` (1)	3.0%	36.3%

TABLE 9.2. *Significant predictors from the additive model fit to the spam training data. The coefficients represent the linear part of $\hat{f}_j$, along with their standard errors and Z-score. The nonlinear P-value is for a test of nonlinearity of $\hat{f}_j$.*

Name	Num.	df	Coefficient	Std. Error	*Z* Score	Nonlinear *P*-value
Positive effects						
`our`	5	3.9	0.566	0.114	4.970	0.052
`over`	6	3.9	0.244	0.195	1.249	0.004
`remove`	7	4.0	0.949	0.183	5.201	0.093
`internet`	8	4.0	0.524	0.176	2.974	0.028
`free`	16	3.9	0.507	0.127	4.010	0.065
`business`	17	3.8	0.779	0.186	4.179	0.194
`hpl`	26	3.8	0.045	0.250	0.181	0.002
`ch!`	52	4.0	0.674	0.128	5.283	0.164
`ch$`	53	3.9	1.419	0.280	5.062	0.354
`CAPMAX`	56	3.8	0.247	0.228	1.080	0.000
`CAPTOT`	57	4.0	0.755	0.165	4.566	0.063
Negative effects						
`hp`	25	3.9	−1.404	0.224	−6.262	0.140
`george`	27	3.7	−5.003	0.744	−6.722	0.045
`1999`	37	3.8	−0.672	0.191	−3.512	0.011
`re`	45	3.9	−0.620	0.133	−4.649	0.597
`edu`	46	4.0	−1.183	0.209	−5.647	0.000

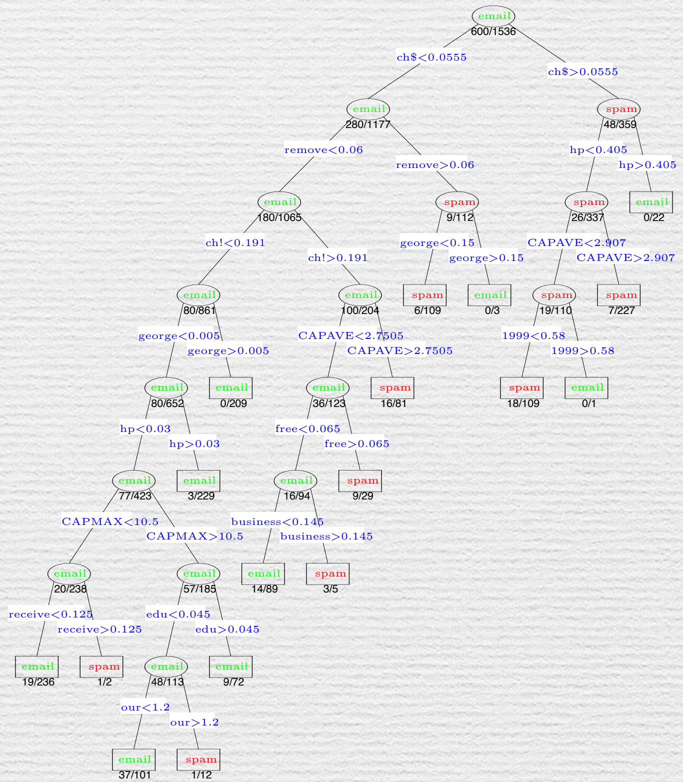


FIGURE 9.5. *The pruned tree for the* `spam` *example. The split variables are shown in blue on the branches, and the classification is shown in every node. The*

Causal Inference for Average Treatment effects

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, \ldots, n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

Following the **potential outcomes** framework (Holland, 1986, Imbens and Rubin, 2015, Rosenbaum and Rubin, 1983, Rubin, 1974), we posit the existence of quantities $Y_i^{(0)}$ and $Y_i^{(1)}$.

- These correspond to the response we **would have measured** given that the i-th subject received treatment ($W_i = 1$) or no treatment ($W_i = 0$).
- **NB:** We only get to see $Y_i = Y_i^{(W_i)}$

The potential outcomes framework

For a set of i.i.d. subjects $i = 1, ..., n$, we observe a tuple (X_i, Y_i, W_i), comprised of

- A **feature vector** $X_i \in \mathbb{R}^p$,
- A **response** $Y_i \in \mathbb{R}$, and
- A **treatment assignment** $W_i \in \{0, 1\}$.

- Define the **average treatment effect (ATE)**, the **average treatment effect on the treated (ATT)**

$$\tau = \tau^{\text{ATE}} = \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right]; \tau^{\text{ATT}} = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \,\middle|\, W_i = 1\right];$$

- and, the **conditional average treatment effect (CATE)**

$$\tau(x) = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \,\middle|\, X = x\right].$$

"Moving the Goalpost": What is Question?

- Estimate $\tau(x) = E[\tau_i | X_i = x]$ as well as possible
  - Why? Want to hold some covariates fixed and look at the effect of others.
- Estimate $\mathrm{BLP}[\tau_i | X_i = x]$
  - Why? "Interpretable"? The best linear predictor is a bit hard to interpret without the whole variance-covariance matrix of nonlinear functions and interactions; you have omitted variable bias on the coefficients you are explaining, relative to $\tau(x)$. My view is that simple models can be more "mis-interpretable" than interpretable.
- Causal Tree: Find partition of covariate space and estimate $E[\tau_i | X_i \in S]$ for each element of partition
  - Why? Easier to interpret than BLP, but still important to report mean, median, percentiles of all covariates for each leaf to understand how leaves are different, when covariates are correlated.
- Which units have highest or lowest treatment effects?
  - Why? Helps understand who could be treated. Can be estimated directly or can draw inferences based on output of causal tree or non-parametric estimates of $\tau(x)$
  - Common practice to display differences between covariates; see Chernozhukov and Duflo (2018)
- What is the best policy mapping from X to treatments W?
  - Why? Sometimes this is the direct object of interest.
  - Fully nonparametric? See e.g. Hirano and Porter (2009)
  - With limited complexity or other constraints? See e.g. Kitagawa and Tetenov (2015), Athey and Wager (2017).
- What is the full set of covariates for which there is statistically significant heterogeneity?
  - List, Shaikh, and Xu (2016) (multiple testing)
- Tradeoffs: More personalization, reliable confidence intervals, role of assumptions, interpretability

Using Trees to Estimate Causal Effects

Model:

$$Y_i = Y_i(W_i) = \begin{cases} Y_i(1) & if \;\; W_i = 1, \\ Y_i(0) & otherwise. \end{cases}$$

- Suppose random assignment of W_i
- Want to predict individual *i*'s treatment effect
  - $\tau_i = Y_i(1) - Y_i(0)$
  - This is not observed for any individual
  - Not clear how to apply standard machine learning tools
- Let

$$\mu(w, x) = \mathbb{E}[Y_i | W_i = w, X_i = x]$$
$$\tau(x) = \mu(1, x) - \mu(0, x)$$

Using Trees to Estimate Causal Effects

$$\mu(w, x) = \mathbb{E}[Y_i | W_i = w, X_i = x]$$
$$\tau(x) = \mu(1, x) - \mu(0, x)$$

- Approach 1: Analyze two groups separately
  - Estimate $\hat{\mu}(1, x)$ using dataset where $W_i = 1$
  - Estimate $\hat{\mu}(0, x)$ using dataset where $W_i = 0$
  - Use propensity score weighting (PSW) if needed
  - Do within-group cross-validation to choose tuning parameters
  - Construct prediction using
    $$\hat{\mu}(1, x) - \hat{\mu}(0, x)$$

- Observations
  - Estimation and cross-validation not optimized for goal
  - Lots of segments in Approach 1: combining two distinct ways to partition the data

- Problems with these approaches

Another Approach: Transform the Outcome

- Suppose we have 50-50 randomization of treatment/control
  - Let $Y_i^* = \begin{cases} 2Y_i & if\ W_i = 1 \\ -2Y_i & if\ W_i = 0 \end{cases}$
  - Then $E[Y_i^*] = 2 \cdot \left(\frac{1}{2}E[Y_i(1)] - \frac{1}{2}E[Y_i(0)]\right) = E[\tau_i]$

- Suppose treatment with probability p_i
  - Let $Y_i^* = \frac{W_i - p}{p(1-p)} Y_i = \begin{cases} \frac{1}{p} Y_i & if\ W_i = 1 \\ -\frac{1}{1-p} Y_i & if\ W_i = 0 \end{cases}$
  - Then $E[Y_i^*] = \left(p\frac{1}{p}E[Y_i(1)] - (1-p)\frac{1}{1-p}E[Y_i(0)]\right) = E[\tau_i]$

- Selection on observables or stratified experiment
  - Let $Y_i^* = \frac{W_i - p(X_i)}{p(X_i)(1-p(X_i))} Y_i$
  - Estimate $\hat{p}(x)$ using traditional methods

Critique of Approach: Transform the Outcome

$$Y_i^* = \frac{W_i - p}{p(1-p)} Y_i = \begin{cases} \frac{1}{p} Y_i & if\ W_i = 1 \\ -\frac{1}{1-p} Y_i & if\ W_i = 0 \end{cases}$$

- Within a leaf, sample average of Y_i^* is not most efficient estimator of treatment effect
  - The proportion of treated units within the leaf is not the same as the overall sample proportion
- This motivates preferred approach: use sample average treatment effect in the leaf

$$\hat{\mu}(R_m) = \frac{1}{|i \in R_m|} \sum_{i \in R_m} Y_i$$

$$\hat{\mu}(w, R_m) = \frac{1}{|i \in R_m \,\&\, i \in S_w|} \sum_{i \in R_m \& i \in S_w} Y_i$$

$$\hat{\tau}(R_m) = \hat{\mu}(1, R_m) - \hat{\mu}(0, R_m)$$

TOT Error

$$C = \sum (\hat{Y}_i^* - Y_i^*)^2$$

Causal Tree Error

$$C = \sum_{x_i \in R_m} (\hat{\tau}_i - Y_i^*)^2$$

Causal Trees

- What are you estimating? Within a leaf estimate treatment effect rather than a mean
  - Difference in average outcomes for treated and control group
  - Weight by normalized inverse propensity score in observational studies
- What is your goal? MSE of *treatment effects:* $-E_{S^T}\left[\sum_{i \in S^T}(\tau_i - \hat{\tau}(X_i))^2\right]$
- Problem: this is infeasible (true treatment effect unobserved)
  - We show we can estimate the criteria
- We also modify existing methods to be "honest." We decouple model selection from model estimation.
  - Split sample, one sample to build tree, second to estimate effects.
  - This changes criteria—novel idea for the literature.

$$-E_{S^T,S^E}\left[\sum_{i \in S^T}(\tau_i - \hat{\tau}(X_i; S^E))^2\right]$$

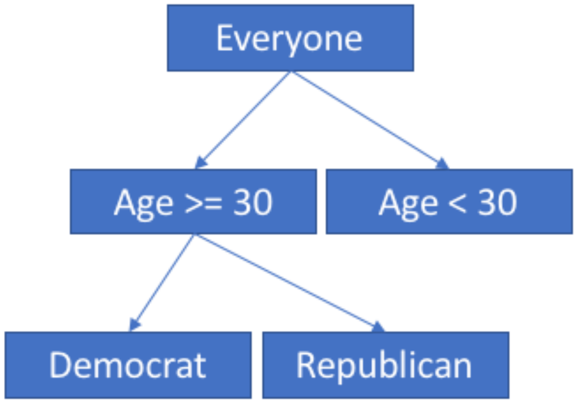
  - Tradeoff:
    - COST: sample splitting means build shallower tree, less personalized predictions, and lower MSE of treatment effects.
    - BENEFIT: Valid confidence intervals with coverage rates that do not deteriorate as data generating process gets more complex or more covariates are added.

Honest Estimation

- To get estimate of \hat{tau} do not use training samples (samples that have been used for construction of tree)
- Instead have an estimation split before training and use that for estimation of \hat{\tau}
- This way the estimation is not overfitted
- Athey. Et.al have have shown that these “honest” treatment effect estimates are asymptotically normal distributed
  - Hence they can be used for confidence interval etc.

Sample from Randomized Experiment

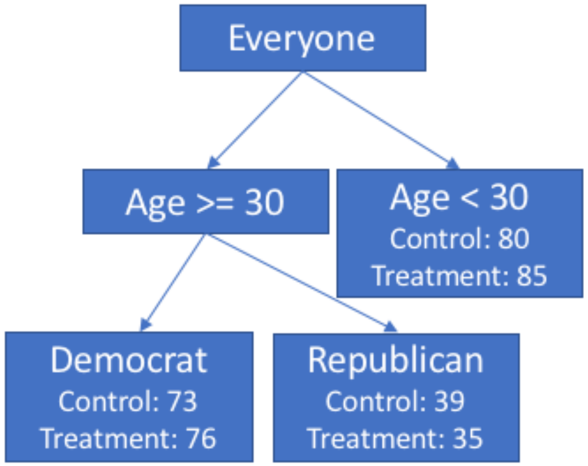
Splitting Subsample



Using the splitting criteria for a causal tree on this subsample, we find three groups in the data:

- People under 30
- Democrats 30 or older
- Republicans 30 or older

Estimating Subsample



We drop everyone in this subsample down the tree and find the percent favorable toward our candidate in each condition in each node. The differences are treatment effects:

- People under 30 = +5 points
- Democrats, 30 and older: +3 points
- Republicans, 30 and older: -4 points

Actual People We Are Trying to Target

We can only afford to target two of these people:

1. 19 year-old Republican
2. 25 year-old Democrat
3. 64 year-old Republican
4. 31 year-old Democrat

Using tree fit by splitting subsample and treatment effects from estimating subsample, we predict the following effects on these people:

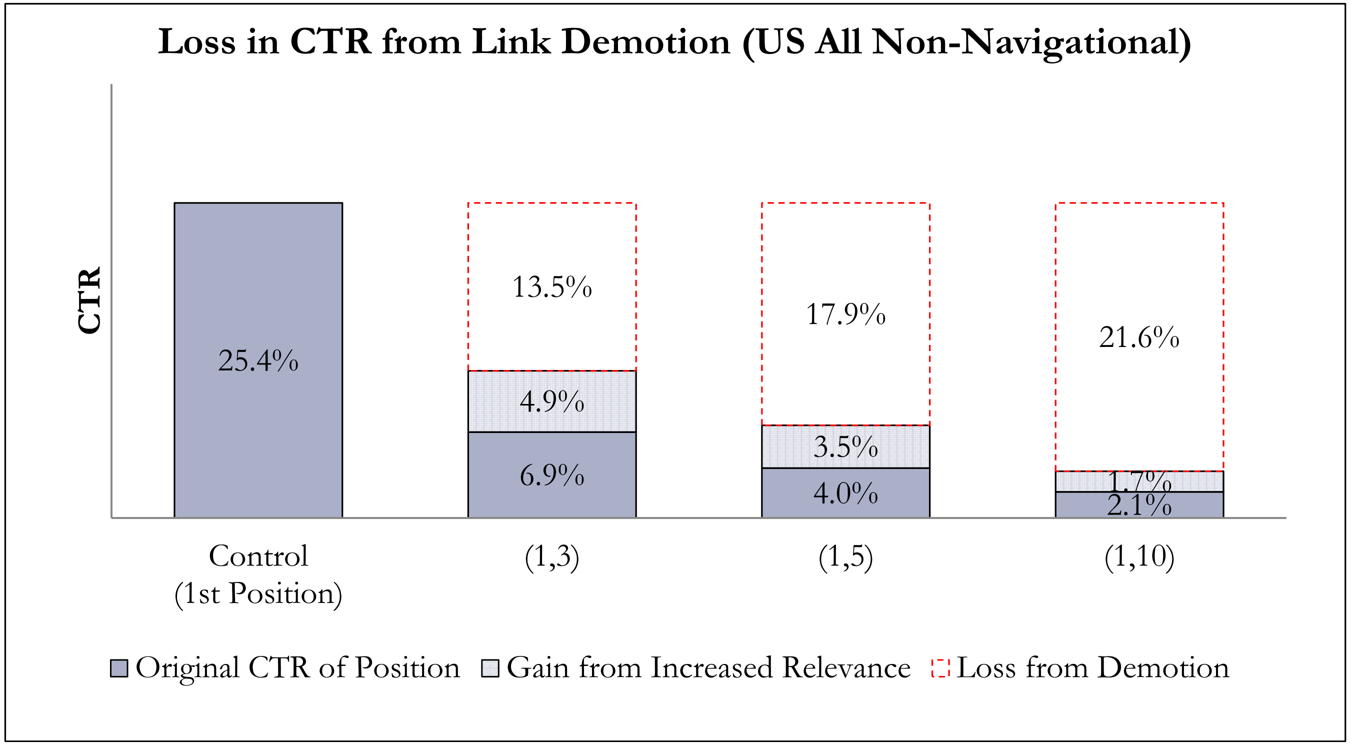
1. +5 points
2. +5 points
3. -4 points
4. +3 points

Target people 1 and 2

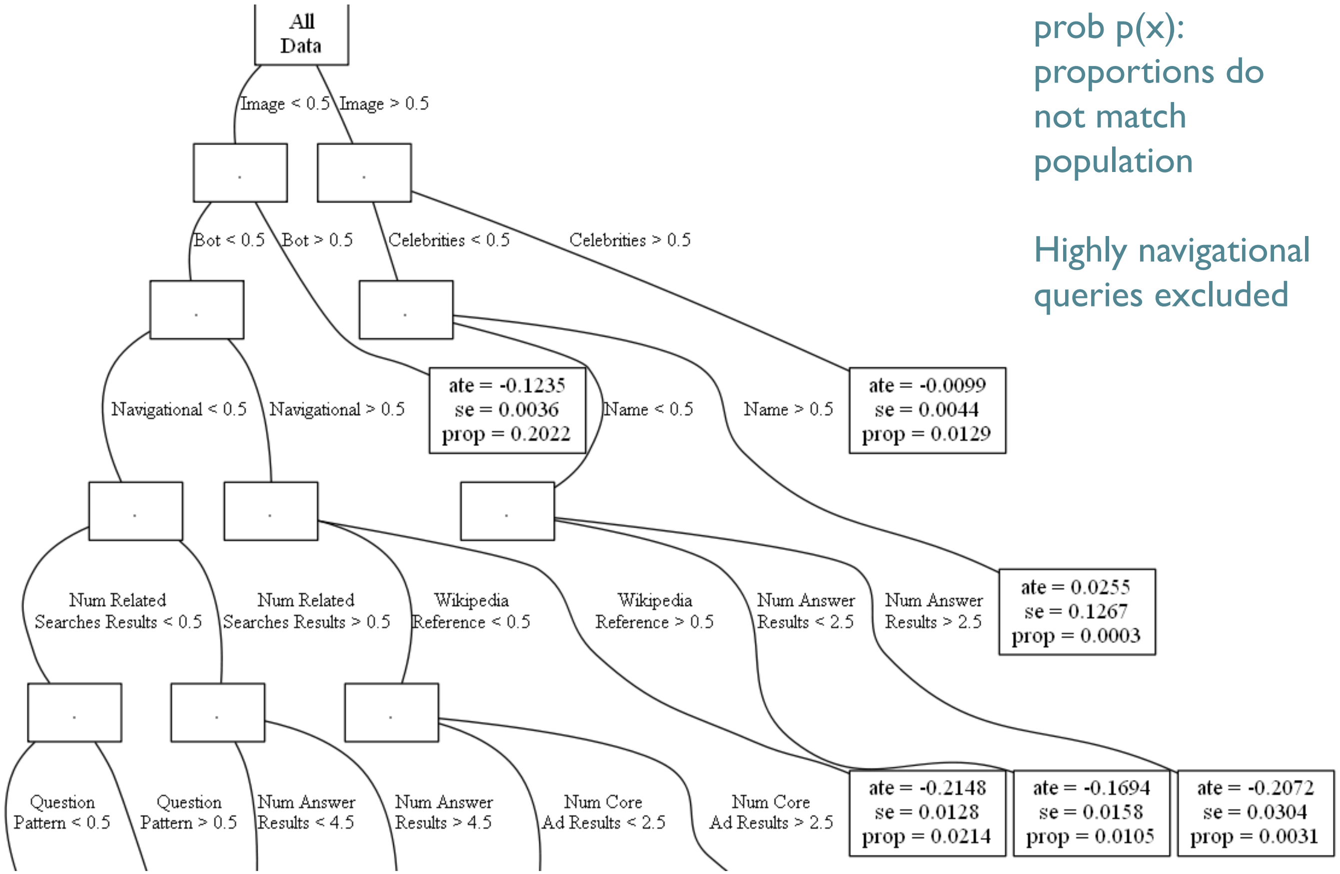
Application: Treatment Effect Heterogeneity in Estimating Position Effects in Search

- Queries highly heterogeneous
  - Tens of millions of unique search phrases each month
  - Query mix changes month to month for a variety of reasons
  - Behavior conditional on query is fairly stable
- Desire for segments.
  - Want to understand heterogeneity and make decisions based on it
  - "Tune" algorithms separately by segment
  - Want to predict outcomes if query mix changes
    - For example, bring on new syndication partner with more queries of a certain type

Relevance v. Position



Search Experiment Tree: Effect of Demoting Top Link (Test Sample Effects)



Some data excluded with prob p(x): proportions do not match population

Highly navigational queries excluded