Value Of Local Showrooms To Online Competitors

Causal Forest Application with R

Author: Jayarajan Samuel, Zhiqiang (Eric) Zheng, Ying Xie

Group Members: Beyza Celik, Luoying Chen, Yihong Liu, Duc Vu

October 23, 2020

UT DALLAS | Naveen Jindal School of Management

Data Preparation for the Application

- Our analysis is based on data individual transactions.
- For each transaction $i = 1, \ldots, n$,
  - $W_i = \texttt{CCStorePresent}_i \times \texttt{AfterStoreClosing}_i$
  - $Y_i = \log(\texttt{prod\_totprice}, \texttt{PagesPerDollar}, \texttt{MinsPerDollar})$ [1]
  - **10 categorical**: `hoh_most_education`, `census_region`, `household_size`, `hoh_oldest_age`, `children`, `racial_background`, `connection_speed`, `country_of_origin`, `prod_category_type` and `BBStorePresent`
  - **4 real-valued covariates**: `pages_viewed`[2], `duration`[3], `prod_qty`, `household_income`
  - We expanded out categorical random variables via one-hot encoding, thus resulting in covariates $X_i \in \mathbb{R}^p$ with $p = 38$ or $p = 37$.

[1] right-skewed

[2] not `PagesPerDollar` is dependent variable

[3] not `MinsPerDollar` is dependent variable

UT DALLAS | Naveen Jindal School of Management

The potential outcomes framework I

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** $W_i \in \{0, 1\}$

Following the **potential outcomes** framework (Imbens and Rubin, 2015) , 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 $(W_i = 0)$.

The potential outcomes framework II

Goal is to estimate the **conditional average treatment effect**

$$\tau(x) = \mathbb{E}\left[Y(1) - Y(0) \mid X = x\right]$$

However in experiments we only get to see $Y_i = Y_i(W_i)$

The potential outcomes framework III

If we make no further assumptions, estimating $\tau(x)$ is not possible.

- Literature often assumes **unconfoundedness** (Rosenbaum and Rubin, 1983)

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i.$$

- When this assumption holds, methods based on matching or propensity score estimation are usually consistent.

Causal Forests for Observational Studies

All analyses are carried out using the R package **grf**, version 1.2.0 (Tibshirani et al., 2018).

- $e(x) = \mathbb{P}[W_i \mid X_i = x]$ for the propensity score
- $m(x) = \mathbb{E}[Y_i \mid X_i = x]$ for the expected outcome marginalizing over treatment
- An application of causal forests using **grf** (Athey and Wager, 2019):
  1. fitting two separate regression forests to estimate $m(\cdot)$ and $e(\cdot)$ (`Y.forest` and `W.forest`)
  2. It then makes out-of-bag predictions using these two first-stage forests, and uses them to grow a causal forest
  3. training a pilot random forest on all the features, and then train a second forest on only those features that saw a reasonable number of splits in the first step.

UT DALLAS | Naveen Jindal School of Management

on amazon.com Sales

The package **grf** has a built-in function for average treatment effect estimation called `average_treatment_effect`. Using this function we obtain:

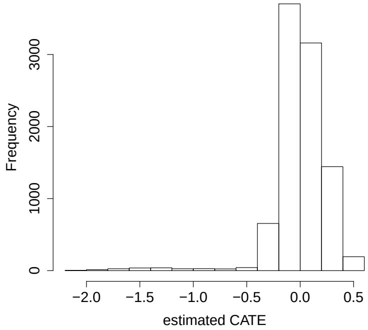

Figure 1: Histogram of out-of-bag CATE estimates from a causal forest

Table 1: 90% CI for the ATT

5%	$\hat{\tau}_t$	95%
-0.42	-0.22	-0.03

on bestbuy.com Sales

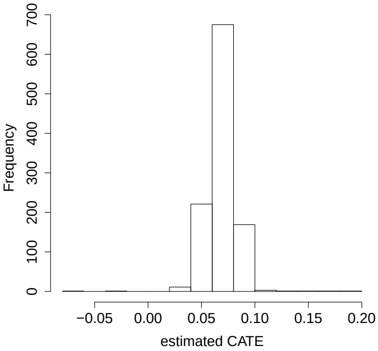

Figure 2: Histogram of out-of-bag CATE estimates from a causal forest

Table 2: 90% CI for the ATT

5%	$\hat{\tau}_t$	95%
-0.39	0.08	0.55

on amazon.com Pages Per Dollar of Sales

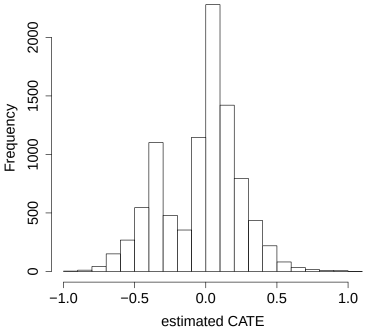

Figure 3: Histogram of out-of-bag CATE estimates from a causal forest

Table 3: 95% CI for the ATT

2.5%	$\hat{\tau}_t$	97.5%
0.02	0.27	0.52

on bestbuy.com Pages Per Dollar of Sales

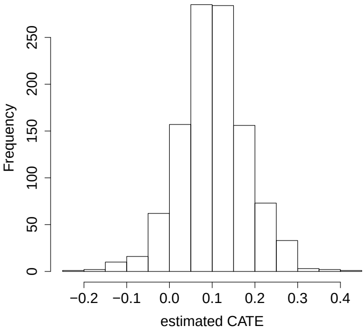

Figure 4: Histogram of out-of-bag CATE estimates from a causal forest

Table 4: 90% CI for the ATT

5%	$\hat{\tau}_t$	95%
-0.47	0.09	0.65

on amazon.com Minutes Per Dollar of Sales

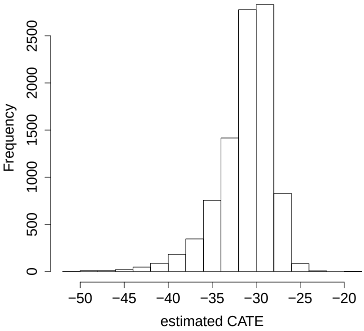

Figure 5: Histogram of out-of-bag CATE estimates from a causal forest

Table 5: 90% CI for the ATT

5%	$\hat{\tau}_t$	95%
-110.76	-24.32	62.13

on bestbuy.com Minutes Per Dollar of Sales

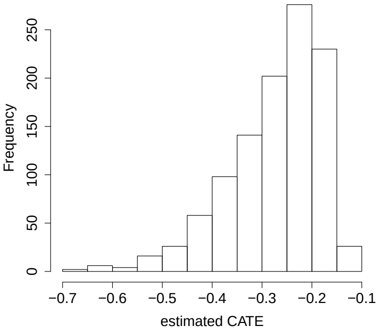

Figure 6: Histogram of out-of-bag CATE estimates from a causal forest

Table 6: 90% CI for the ATT

5%	$\hat{\tau}_t$	95%
-0.63	-0.29	0.05

Reference

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409*, 2019.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Julie Tibshirani, Susan Athey, Stefan Wager, Rina Friedberg, Luke Miner, Marvin Wright, Maintainer Julie Tibshirani, LinkingTo Rcpp, RcppEigen Imports DiceKriging, and GNU SystemRequirements. Package ‘grf’, 2018.