Honest Random Forests for Heterogeneous Treatment Effect Estimation with Covariate Shift

Colleen Chan[1], Theo Saarinen[2], Jasjeet Sekhon[1]

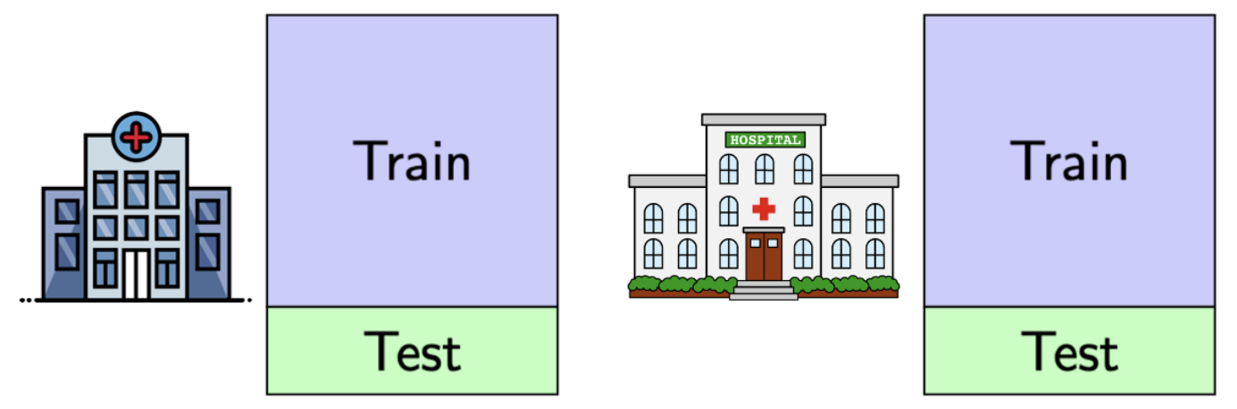
[1]Department of Statistics and Data Science, Yale University, [2]Department of Statistics, University of California, Berkeley

Introduction

In many real world applications, machine learning algorithms are used for prediction in environments that do not share the covariate distribution of the data that the algorithm was trained on. When this is the case, it is important for the algorithm to be robust to the distributional shift of the covariates to avoid harmful results down the line.

For example, the Epic Sepsis Model, a proprietary sepsis prediction model, achieved good prediction accuracy on the three hospitals that it was trained on (Wong et al. 2021). However, the model had failed to predict sepsis in the majority of cases on the hundreds on hospitals that it was later deployed on. This was due to a shift in the data distribution of the hospitals, which had different protocols and equipment and served different patient populations than the hospitals in the training sample.

How models are evaluated in papers



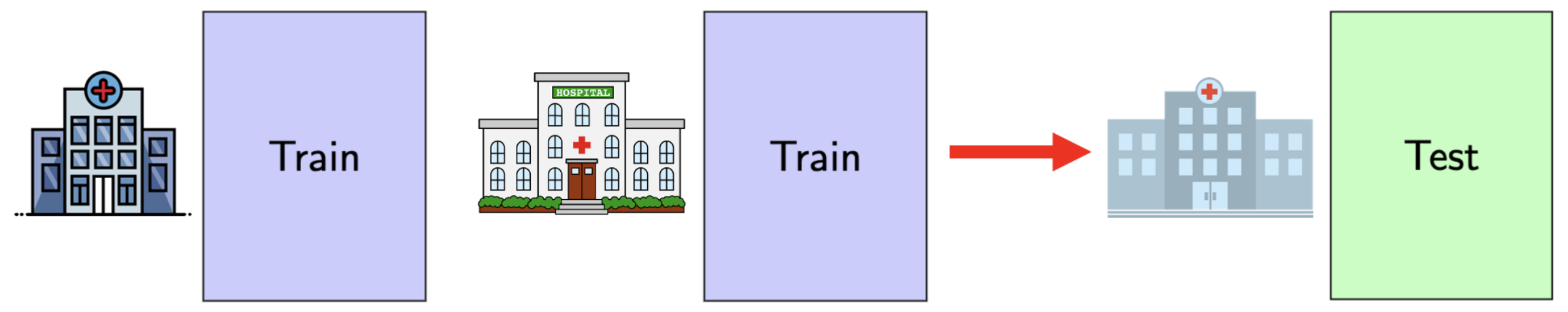
Train

Test

Train

Test

How models are deployed



Train

Train

Test

Background

In the standard decision tree algorithm, prediction is done by going down the tree structure and giving the mean outcome of the leaves. The overall prediction is then a simple weighted average of the training observations. This can lead to overfitting because the *same data* was used for tree partitioning and for prediction. A tree is considered *honest* if it does not use the same information for selecting the model structure as for estimation given a model structure (Biau, 2010). In an honest tree, the data is split into two disjoint sets: a *splitting set* $\mathcal{I}$ to determine the tree structure and an *averaging set* $\mathcal{J}$ to estimate the mean values within each leaf.

Notation

We observe the tuple $\mathcal{D} = (Y, S, Z, X)$ where we denote:

- Y outcome of interest
- $S \in \mathcal{S} = \{1, \dots, K\}$ group number
- $Z \in \{0, 1\}$ treatment
- $\mathbf{X} \in \mathcal{X}$ vector of covariates.

For treatment effect estimation, we also assume that a separate RCT was conducted on each group. The estimand of interest is the group-specific conditional average treatment effect (CATE):

$$\tau_s(x_i) = \mathbb{E}[Y(1) - Y(0) \mid X = x_i, S = s].$$

While we can directly estimate a site-specific CATE using only data from a given site, we want to leverage data from the other sites for a more efficient CATE estimate.

Methods

We propose a novel sample splitting procedure during training in order to induce robustness to distribution shifts at prediction time. If we expect a distribution shift with respect to a certain group later on, we train each tree in the forest so that the observations in that group are left out during splitting and averaging. B trees are grown for each group for a total of $K \cdot B$ trees in the forest.

Algorithm 1 Honest Random Forest with Groups

```
Define $\mathcal{D}_{-k}$ as the subset of the data $\mathcal{D}$ excluding the $k$th group.
for $k = 1$ to $K$ do
  for $b = 1$ to $B$ do
    Randomly split $\mathcal{D}_{-k}$ into two disjoint sets $\mathcal{I}$ and $\mathcal{J}$
    Grow a tree via recursive partitioning using the $\mathcal{I}$-sample
    Estimate leafwise responses using only the $\mathcal{J}$-sample observations
    Optional (double tree): Repeat and switch the roles of $\mathcal{I}$ and $\mathcal{J}$ and average over the two predictions.
  end for
end for
```

Heterogeneous Treatment Effect Estimation

Assumptions:

❶ Consistency of potential outcomes: If $Z_i = z$, then $Y_i(z) = Y_i$

❷ Unconfoundedness over Z:

$$Y(0), Y(1) \perp Z \mid X, S = s \text{ for all } s \in \mathcal{S}$$

❸ Positivity of treatment assignment:

$$0 < P(Z = 1 \mid X = x, S = s) < 1 \text{ for all } x \in \mathcal{X}, s \in \mathcal{S}$$

❹ Treatment effect unconfoundedness over S:

$$Y(1) - Y(0) \perp S = s \mid X \text{ for all } s \in \mathcal{S}$$

❺ Positivity of group participation:

$$0 < P(S = s \mid X = x) < 1 \text{ for all } x \in \mathcal{X}, s \in \mathcal{S}$$

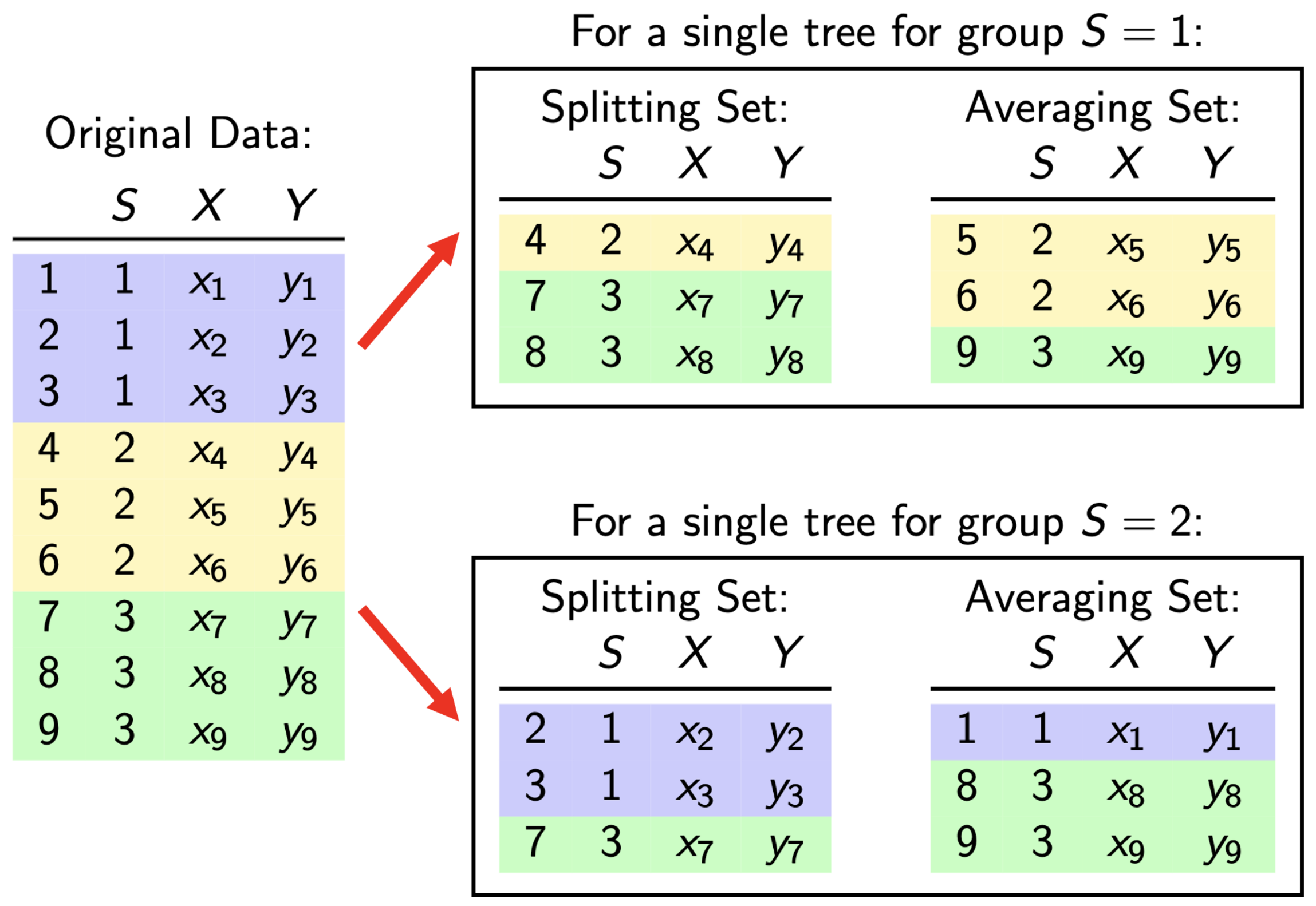
Under Assumptions 1-5, the CATE in a target group s can be identified using the following functional of the observed data distribution:

$$\begin{aligned}\tau(x_i) = {} & \mathbb{E}[Y \mid X = x_i, Z = 1, S \in \mathcal{S}] \\ & - \mathbb{E}[Y \mid X = x_i, Z = 0, S \in \mathcal{S}].\end{aligned} \tag{1}$$

Proof:

$$\begin{aligned}
&\tau(x_i) \\
&= \mathbb{E}[Y \mid X = x_i, Z = 1, S \in \mathcal{S}] - \mathbb{E}[Y \mid X = x_i, Z = 0, S \in \mathcal{S}] \\
&= \mathbb{E}[Y(1) \mid X = x_i, Z = 1, S \in \mathcal{S}] \\
&\qquad - \mathbb{E}[Y(0) \mid X = x_i, Z = 0, S \in \mathcal{S}] \quad \text{by Asm. 1} \\
&= \mathbb{E}[Y(1) \mid X = x_i, S \in \mathcal{S}] \\
&\qquad - \mathbb{E}[Y(0) \mid X = x_i, S \in \mathcal{S}] \quad \text{by Asm. 2, 3} \\
&= \mathbb{E}[Y(1) \mid X = x_i, S = s] \\
&\qquad - \mathbb{E}[Y(0) \mid X = x_i, S = s] \quad \text{by Asm. 4, 5} \\
&= \mathbb{E}[Y(1) - Y(0) \mid X = x_i, S = s] \text{ by linearity of expectation.}
\end{aligned}$$

□



Original Data:

	S	X	Y
1	1	x_1	y_1
2	1	x_2	y_2
3	1	x_3	y_3
4	2	x_4	y_4
5	2	x_5	y_5
6	2	x_6	y_6
7	3	x_7	y_7
8	3	x_8	y_8
9	3	x_9	y_9

For a single tree for group $S = 1$:

Splitting Set:

	S	X	Y
4	2	x_4	y_4
7	3	x_7	y_7
8	3	x_8	y_8

Averaging Set:

	S	X	Y
5	2	x_5	y_5
6	2	x_6	y_6
9	3	x_9	y_9

For a single tree for group $S = 2$:

Splitting Set:

	S	X	Y
2	1	x_2	y_2
3	1	x_3	y_3
7	3	x_7	y_7

Averaging Set:

	S	X	Y
1	1	x_1	y_1
8	3	x_8	y_8
9	3	x_9	y_9

We also train the forest with out-of-bag (OOB) honesty, which uses the OOB observations as the averaging set for each tree. For prediction for an in-sample observation, we aggregate only over the trees for which that observation's respective group is left out. For prediction out-of-sample, all trees are used. These methods are implemented in the `Rforestry` R package, available on CRAN.

Simulation Studies

Monte Carlo simulations are conducted to assess the proposed method. We assume there are $K = 5$ sites, each with sample size 300. For each simulation, we simulate features $\mathbf{X}_i \in \mathbb{R}^5$ from a multivariate normal $N(\mathbf{0}, \mathbf{I})$. The outcome model is $Y = \frac{1}{2}x_1 + \Sigma_{d=2}^{4} x_d + 5(x_1 - 3) \cdot U_k$, where $U_k \sim \text{Unif}(0, 1)$, which represents the site-level heterogeneity; Y is also then scaled to have mean 0 and standard deviation 1 for stability. These simulation settings are motivated by designs in Tan, Chang, and Tang (2021). Without loss of generality, we hold out site 1 to be the test site.

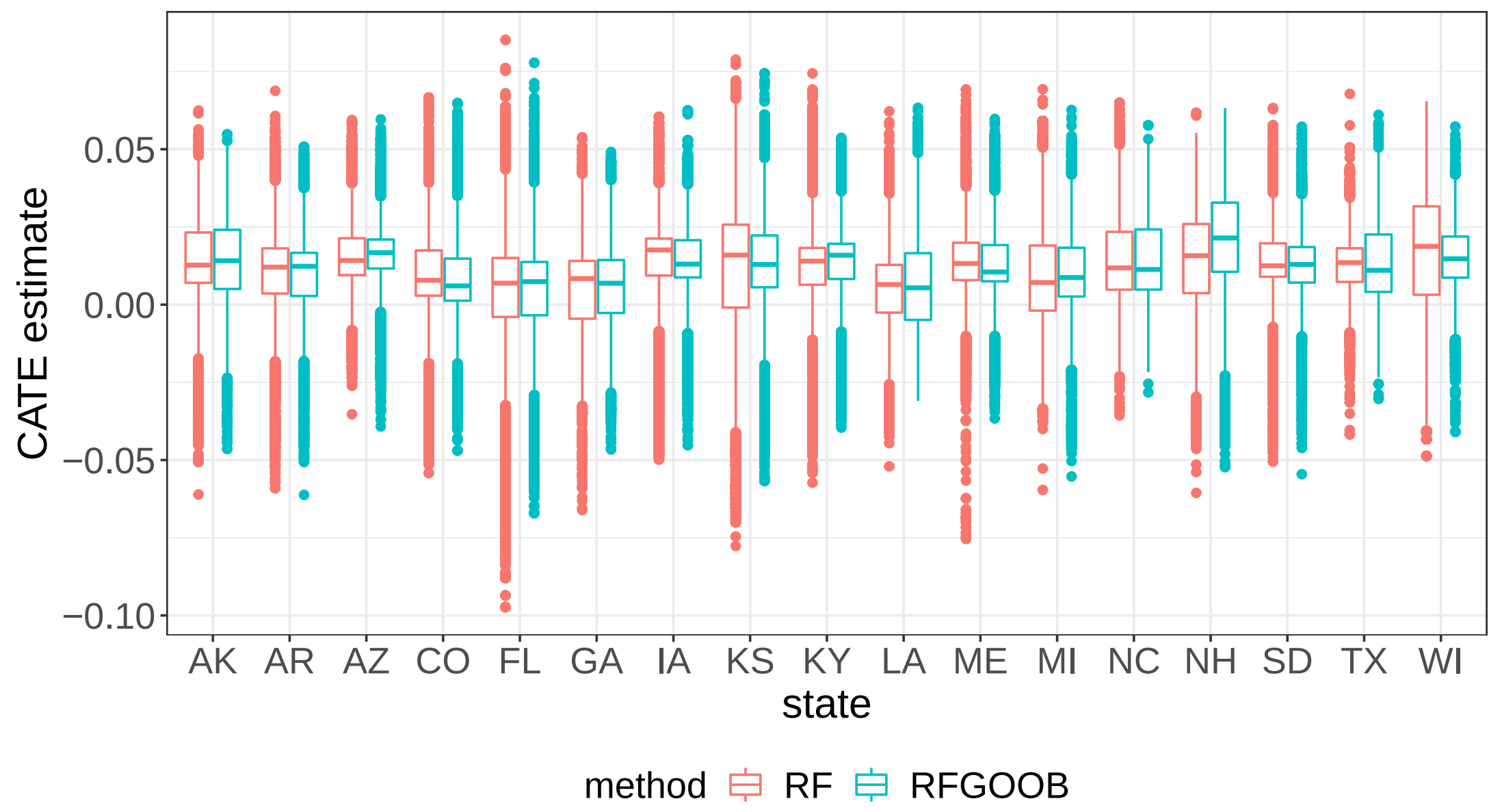
We evaluate the MSE on the test set with other tree-based machine learning methods: Bayesian additive regression trees (BART); random forest (RF); random forest with groups and OOB honesty (RFGOOB); XGBoost (XGB). To tune parameters, we use the OOB set for validation for RFGOOB. For the remaining methods, we use 5-fold cross-validation. We report the mean MSE on the test set and the mean difference between the out-of-sample (OOS) MSE (i.e., the error on the validation set) and test MSE across the 250 simulations conducted.

Method	Test MSE	Test MSE − OOS MSE
BART	1.669	0.382
RF	0.982	0.385
RFGOOB	**0.964**	**-0.038**
XGB	0.967	0.393

RFGOOB achieves the lowest MSE on the test set and the smallest difference between the OOS error and test error. RFGOOB is also the only method that provides a conservative OOS error estimate.

Application

We apply our method on a large-scale field experiment in which a nonpartisan campaign randomly sent mailers to encourage people to vote in the 2014 general election across 17 states (Gerber et al. 2017). The states are heterogeneous with respect to each state's population. We hold out each state and estimate its CATE by estimating each term in Equation 1 using RF and RFGOOB, setting the group option to the individuals' state and weighting the trees in each state proportionate to the sample size. The features used are age, sex, race, marital status, and the proportion of eligible general elections that an individual voted in since 2006. We report the estimated CATE's for each state below.



The most unbiased model for the CATE for state s is the local model, i.e., one that uses only data from state s. Assuming the local model using the T-learner with random forests is the true model (Künzel et. al. 2019), we find the RFGOOB estimates achieve lower bias and a lower MSE than the RF estimates.

Discussion

Real world shifts are ubiquitous in deployments, whose environments are often different than the ones on which they are trained. By training a random forest excluding a pre-specified group for which we might expect a distribution shift later on, our proposed method produces predictions that are more robust than those of the standard random forest. Given that the test set shares a similar covariate distribution to those of the groups during training, we also expect the out-of-sample error to be an unbiased estimate of the test set error. We note that for *within-sample* prediction, our method produces less efficient estimates than those of the standard random forest since less data is used to train the forest. Nonetheless, our method can be especially useful for heterogeneous treatment effect estimation, where researchers often want to transport causal inferences learned from multiple RCT's to a different population of interest. Future research includes extending the groups option setting to other machine learning algorithms, such as gradient boosting.

Acknowledgements

CC and JS thank the Adobe Data Science Research Award for funding.

Contact Information

[1]colleen.chan@yale.edu [2]theo_s@berkeley.edu
[3]jasjeet.sekhon@yale.edu