**Authors**: Max Bronckers, Veronica Song, and Dustin Zhang

**Summary**

In this paper, we want to compare the use of Bayesian Additive Regression Trees (Green and Kern, 2012) versus Causal Forests (Wager and Athey, 2017) to estimate heterogenous treatment effects among survey responders using an empirical dataset. In doing so, we will assess the benefits and drawbacks of these two machine learning methods that have been touted as best-in-class methods to estimate treatment effect heterogeneity.

**Background**

Many economists are interested in estimating treatment effects for populations or subgroups of populations. They do this by estimating Average Treatment Effect (ATE) or Conditional Average Treatment Effects (CATE). ATE is the average outcome of the treatment group in a selected population minus the average outcome of the control group in the same population. CATEs are the average treatment effects among subgroups of a population as defined by baseline covariates.

Naturally, however, these treatment effects often vary by individual, and these variations may be related to underlying individual characteristics that might not be measurable. This variation in response based on individual characteristics is referred to as \*\*treatment effect heterogeneity\*\*. Understanding and being able to estimate heterogenous treatment effects is especially important for policy makers - knowing how groups of people, given particular characteristics, will respond to a policy helps inform major policy decisions.

**Related work**

In this paper, we draw on the methods and intuition of four works described as follows.

1. Modeling Heterogenous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees (Green and Kern, 2012)

Green and Kern initially proposed the usage of BARTs to estimate treatment effect heterogeneity. They analyzed the heterogeneity in treatment effects of survey data that was intended to measure the stigma of the word "welfare" when asking individuals about public spending. We will extend this work by using their data and want to directly compare the results of BART with those of Causal Forests. Green and Kern's justification for BART’s suitability for uncovering heterogeneity in treatment effect is also important in establishing our comparison criteria. They discuss that BART is an attractive method because it helps eliminating researcher discretion in choosing their data subgroup formation, detecting and modeling high dimension non-linear relationships, discontinuities, and interactions, and relies on minimal parameter tuning and consequently is easier to use.

2. Estimation and Inference of Heterogenous Treatment Effects using Random Forests (Wager and Athey, 2017)

Wager and Athey (2017) proposed Causal Forests, an extension of Random Forests, as a useful ML method for causal inference due to its "tractable asymptotic theory". They directly criticize BART for being difficult to use for statistical inference, claiming that there is no asymptotic normality theory around the posterior of BART that guarantees "concentration around the true conditional mean function". On the other hand, the causal forest predictions are asymptotically Gaussian and unbiased, allowing confidence intervals to be constructed. The construction of adequate confidence intervals is especially relevant in investigating heterogenous treatment effects, as it guides policy decisions around a consistent estimate. The authors do not, however, actually compare BART versus Causal Forests in their empirical evaluation. This is the underlying motivation for our research - we want to focus in on their criticism and provide a comprehensive evaluation of BART versus Causal Forests in their capacity as heterogeneous treatment effect estimator methods using the criteria described below. We aim to offer both a quantitative comparison in estimator performance, computational cost, and robustness, but also a qualitative comparison of model assumptions and ease of use.

(3) Machine Learning Methods Economists Should Know About (Athey and Imbens, 2019)

This paper serves as a very recent reference point of the current approaches machine learning offers in regard to treatment effect heterogeneity. By summarizing the challenges of estimating treatment effect heterogeneity into three questions main questions, Athey and Imbens (2019) note some important criteria for any heterogeneous treatment effect estimator method. Its literature overview of methods for estimation of treatment effect heterogeneity serves as the starting point for this paper. It reports that BART and Causal Forests are both tree-based ML methods that have been applied to causal inference and are claimed to be relatively strong candidates for treatment effect heterogeneity estimation. To our knowledge, this is the first work that directly evaluates the two methods with respect to each other on empirical survey data.

(4) Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (Hahn, Murran and Carvalho, 2019)

To our knowledge, this is the only work that directly compares BART and Causal Forests for causal inference. However, this work assesses the methods only on a synthetic dataset. The authors simulate the performances of different Bayesian Causal Forest (BCF) and BART specifications in estimating CATE for a **synthetic dataset** created by the Atlantic Causal Inference Conferences (ACIC) in 2016 and for the effect of smoking on medical expenditures. They find that the BCF and BART generally "concur on the nonlinear prediction problem", BCF estimates are generally lower and have "much shorter intervals". In the synthetic ACIC data analysis, the results indicated BCF "performed best in terms of estimation error for CATE ... as measured by bias and absolute bias". We will similarly compare the results of BCF and BART applied to the estimation of treatment effect heterogeneity but, in contrast to this paper's usage of a synthetic dataset, we will apply the methods to an empirical dataset used by Green and Kern.

**Methodology**

We adopt the empirical data from Green and Kern (2012) for our comparisons. We use the well-known survey experiment about public support for government spending on welfare to the poor from the GSS to model systematic treatment effect heterogeneity based on the wording of the survey question and its interaction with racial and education perceptions.

**Evaluation criteria**

We will assess the methods on the following criteria:

- Accuracy/performance of estimator based on bias

- Model assumptions (more and stronger assumptions weaken applicability)

- Parameter tuning complexity/ease-of-use

- Computational training cost

- Robustness

**Robustness evaluation**

We will evaluate both methods under different hyperparameter configurations to assess the robustness of these methods. Time permitting, we will also consider a Monte Carlo simulation with a DGP based on another empirical dataset: "Evaluations of Swiss Active Labor Market Policies" (Kraus et al. 2020).