Yanji Du

Evaluation of neural network-based heterogeneous treatment effect estimation methods in survey experiments

1. Introduction

Understanding the heterogeneity of treatment effect can yield further insight into causal relationships in the social sciences. Dragonnet, a novel architecture that harnesses the predictive power of neural networks, has been shown to be effective in uncovering treatment effects in certain empirical settings with observational data. However, it is unclear how Dragonnet performs with data from survey studies. Survey studies are a ubiquitous and critical experimental archetype in the social sciences, so it is of high interest to compare the performance of Dragonnet with established models, like BART (Bayesian Additive Regression Trees) in this domain. Inspired by Green and Kern's 2012 work on applying BART to survey experiments, in this work we update Green and Kern's BART study with recent data, replicate the analysis using Dragonnet, and compare the results of the two methods.

2. Background

2.1 Bayesian Additive Regression Trees (BART) (Green and Kern 2012)

As an improvement over parametric models such as linear or logistic regression, Green and Kern proposed using Bayesian Additive Regression Trees (BART) as a method for analyzing treatment effect heterogeneity (Green and Kern 2012). Key advantages include avoiding a need for manual specification of nonlinear relationships and interactions, making BART a flexible and robust tool that can be used with minimal researcher discretion. The authors evaluate the performance of BART using an empirical

dataset from the General Social Survey: in this study, we will use the same dataset as a benchmark to evaluate new methods.

BART leverages a sum-of-trees model, whereby a large number of trees are fit, and the predictions are combined. Each tree is constructed by iteratively subsetting the data at dichotomous decision points on a single predictor. Trees are grown, shrunk, or changed along the way. A prior is put on the parameters of the model to limit the influence of individual trees and to limit tree depth. The posterior is computed using Markov Chain Monte Carlo (MCMC): trees are drawn sequentially from the posterior.

CATE estimates are generated through simulation. The fitted BART model generates posterior draws for synthetic observations, where the covariate(s) of interest are varied one-by-one, and generates two vectors of posterior draws (where the treatment variable is set to 1 and 0, respectively). The CATE is computed as the mean of the subtracted vectors.

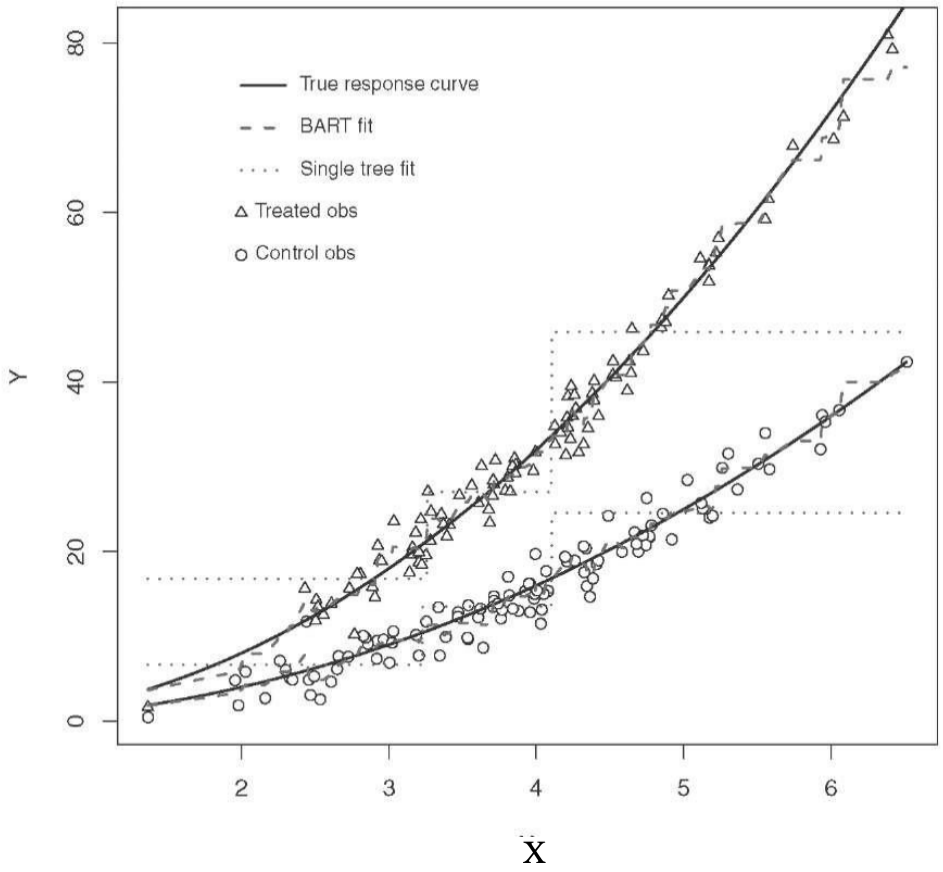

Fig. 1: Single-tree model and BART fits to simulated data (Green and Kern 2012).

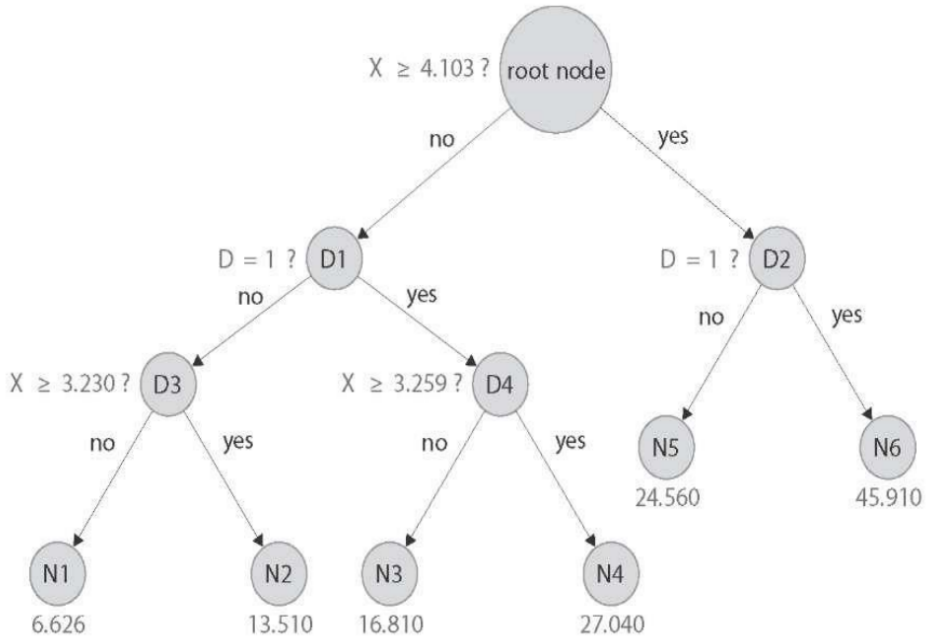

Fig. 2: Single-tree fit (Green and Kern 2012).

2.2 Dragonnet (Shi 2019)

One recent advance adapts a neural network-based architecture, called Dragonnet, to estimate heterogeneous treatment effects. Neural networks have demonstrated strong predictive performance across many domains, and motivates the application to treatment effect estimation. Similar to BART, the model allows for minimal specification of nonlinear relationships and interactions.

The three-headed architecture (Fig. 3) jointly optimizes for propensity score and conditional outcome from covariates and treatment information (e.g. a shared representation of the covariates, Z(X), is used to predict both the treatment and outcome). The rationale for this follows from the Sufficiency of Propensity Score Theorem: “it suffices to adjust for only the information in X that is relevant for predicting the treatment. Consider the parts of X that are relevant for predicting the outcome but not the treatment. Those parts are irrelevant for the estimation of the causal effect, and are effectively noise for the adjustment. As such, we expect conditioning on these parts to hurt finite-sample performance —instead, we should discard this information” (Shi 2019).

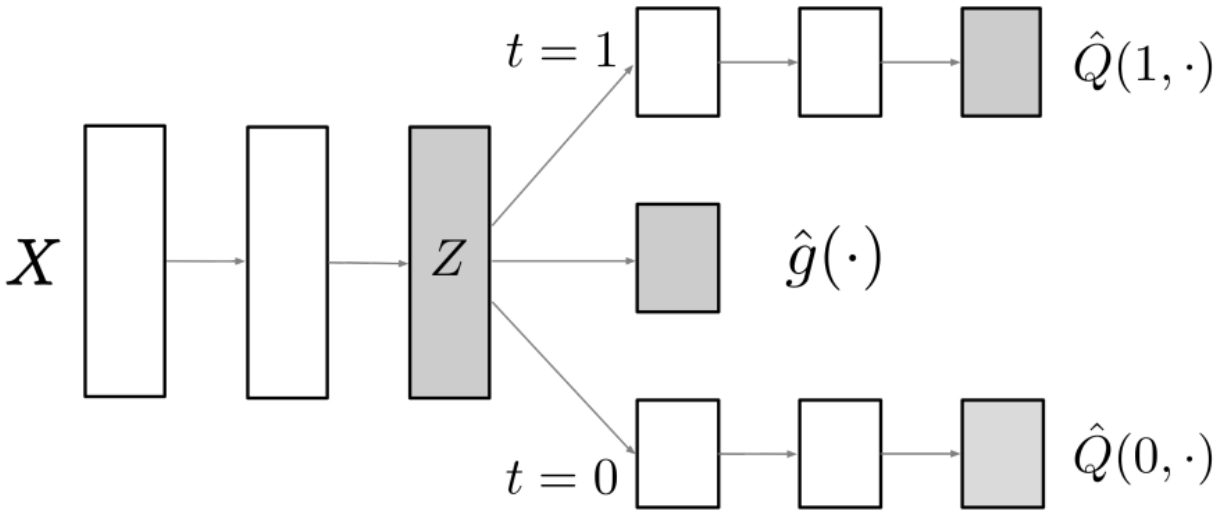

Fig. 3: Dragonnet architecture (Shi 2019).

Empirical studies proposed by the authors have confirmed that the Dragonnet model appears to trade off quality of predictions and propensity scores. We evaluate the Dragonnet architecture on simulation studies and compare its results on an empirical survey dataset with those of BART (Green and Kern 2012).

3. Research method and Data

3.1 Simulation Study

We compared the performance of BART and Dragonnet in three simulation scenarios of heterogeneous treatment effects with data-generating processes of various complexity:

1. Simple linear treatment effect by covariate interaction
2. 3-way linear treatment effect by covariate interaction
3. Non-linear treatment effect by covariate interaction

Yanji Du

In all scenarios we simulate randomized control trials, so $D_i \sim$ Bernoulli(0.5). In each simulation there are p=5 dimensions and n=10,000 samples. Covariates are generated from a standard normal distribution. A baseline treatment effect is computed as the mean of three covariates. Treatment effects are computed as follows:

1. $3.\tau = 5*X_1 + 0.5$
2. $\tau = 5*X_1 - 3*X_2 + X_3^2$
3. $\tau = X_1 + \ln(\text{abs}(X_2)) + 3*\text{sqrt}(\text{abs}(X_3*X_4)) - 3*(X_0>0)$

For each simulation scenario, the dataset was split into train and validation sets (80%/20% split). The train set was used to fit the estimators; the fitted estimators were used to predict CATEs on the train and validation set separately. Results from the validation set are reported.

The Dragonnet model appears to perform similarly well in estimating the distribution of CATEs at a low MSE and high area-under-uplift-curve (AUUC) compared to our BART benchmark in all simulation scenarios.

Simulation 1 Results

Dragonnet performed comparatively well with BART (lower MSE, higher AUUC) (Table 1). The true distribution of CATE predictions by observation in the validation set are more closely mirrored in Dragonnet than BART (Fig. 4).

	ATE	MSE	Abs % Error of ATE	AUUC
Actuals	0.420685	0.000000	0.000000	NaN
DragonNet	0.458209	0.014815	0.089197	3.820838
BART	0.441197	0.121989	0.048760	3.816819

Table 1: Summary of results by estimator compared with oracle data for validation set.

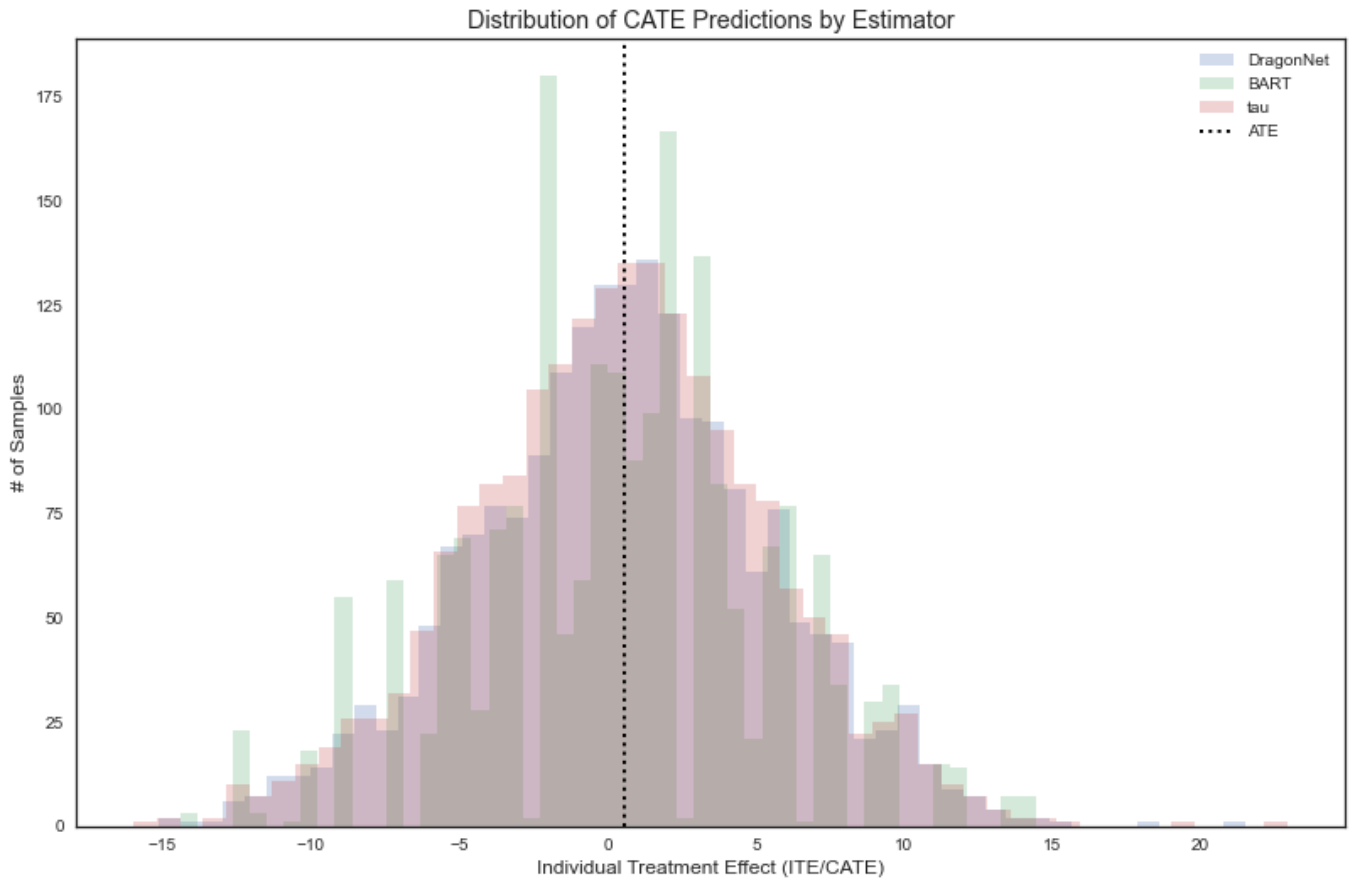

Fig. 4: Distribution of CATE predictions by estimator compared with oracle (tau) for validation set.

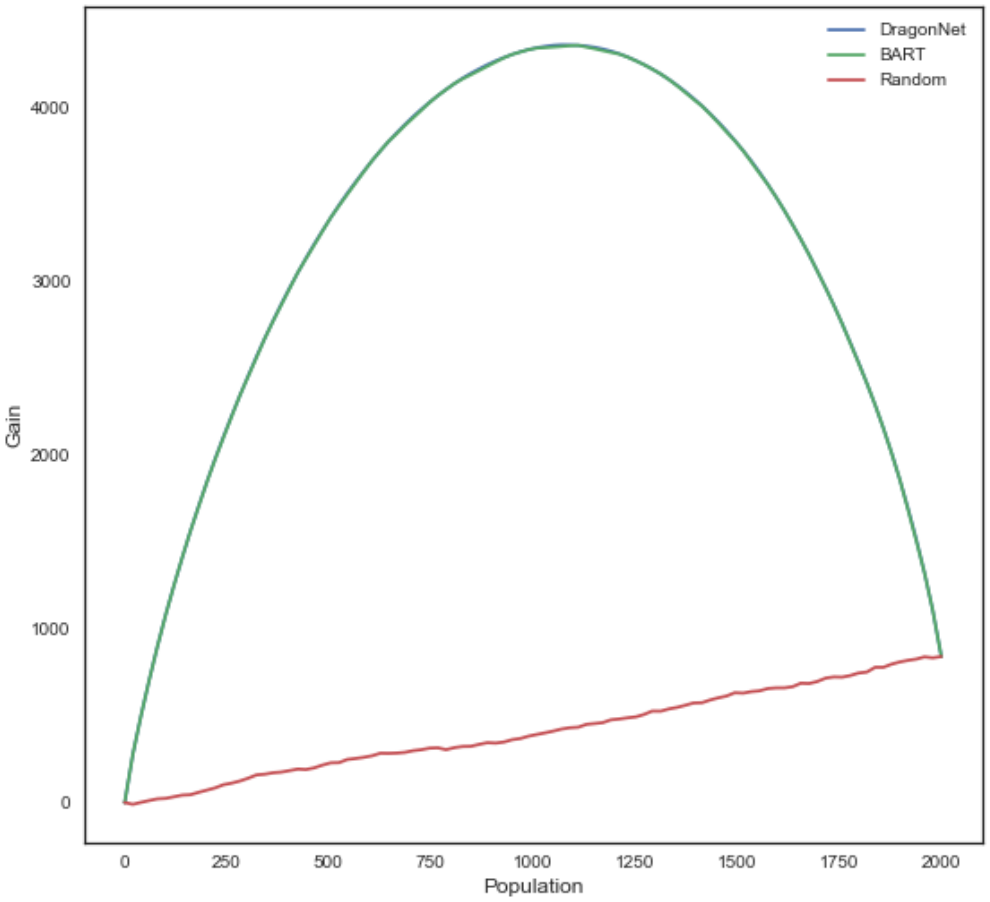

Fig 5: Area Under Uplift Curve by Estimator (Validation Set).

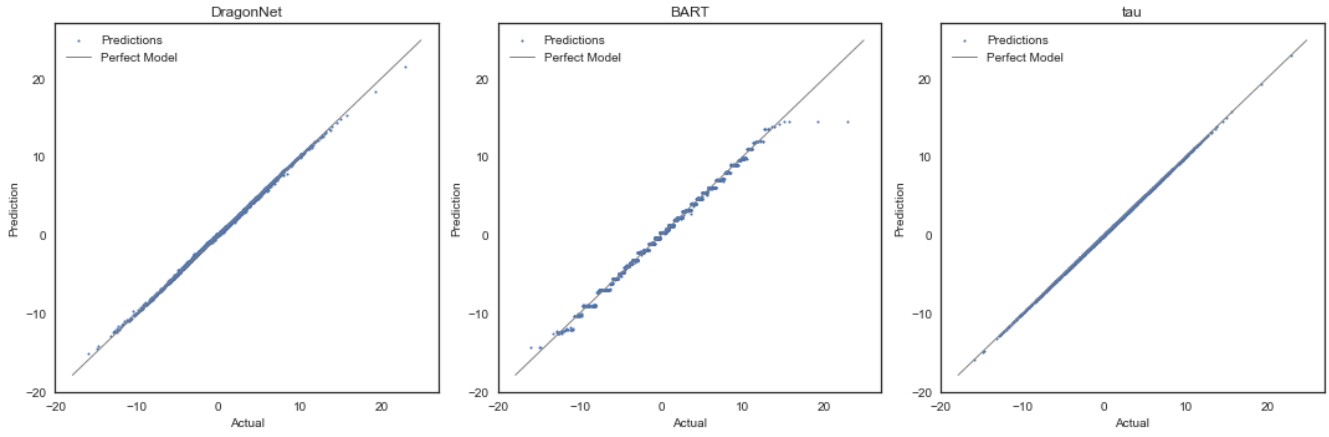

Fig 6.: Actual vs. Predicted CATEs (Validation Set).

Simulation 2 Results

Dragonnet performed comparatively well with BART (lower MSE, higher AUUC) (Table 2). The true distribution of CATE predictions by observation in the validation set are closely mirrored by both Dragonnet and BART (Fig. 7).

	ATE	MSE	Abs % Error of ATE	AUUC
Actuals	0.956819	0.000000	0.000000	NaN
DragonNet	0.991163	0.025387	0.035894	2.257193
BART	0.982249	0.229463	0.026578	2.253589

Table 2: Summary of results by estimator compared with oracle data for validation set.

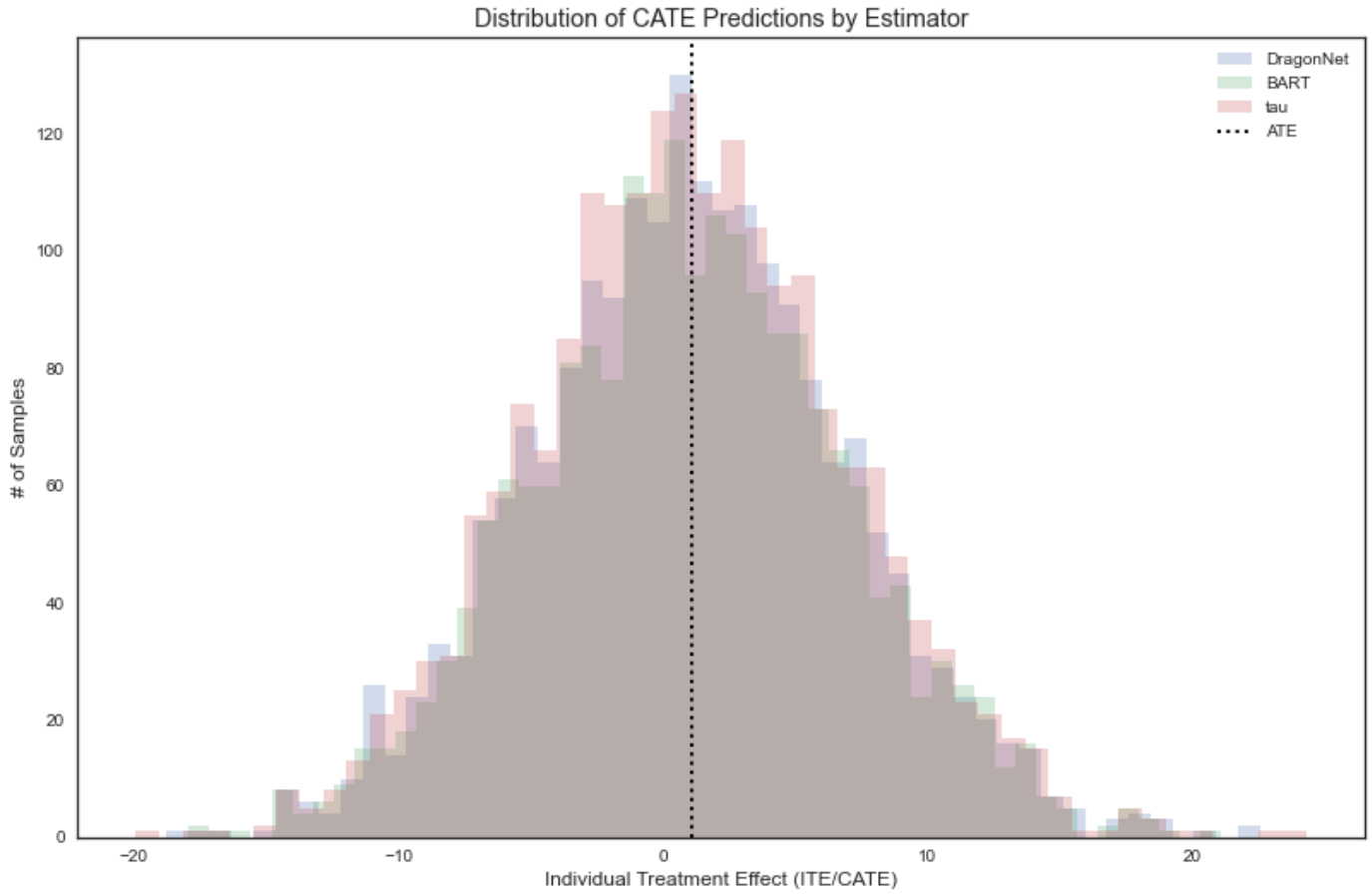

Fig. 7: Distribution of CATE predictions by estimator compared with oracle (tau) for validation set.

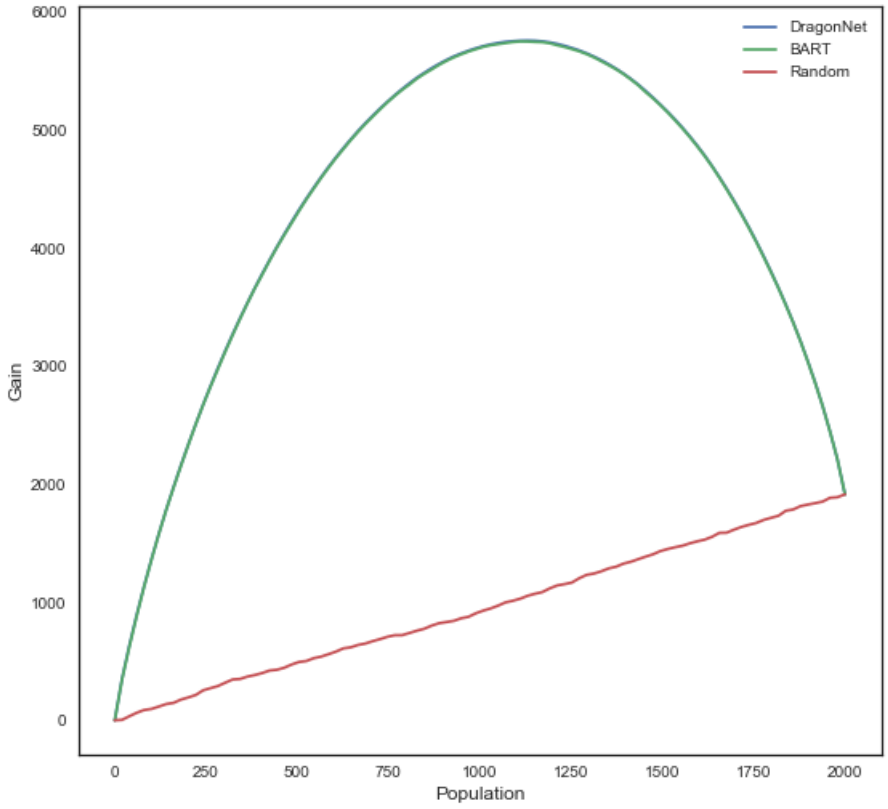

Fig. 8: Area Under Uplift Curve by Estimator (Validation Set).

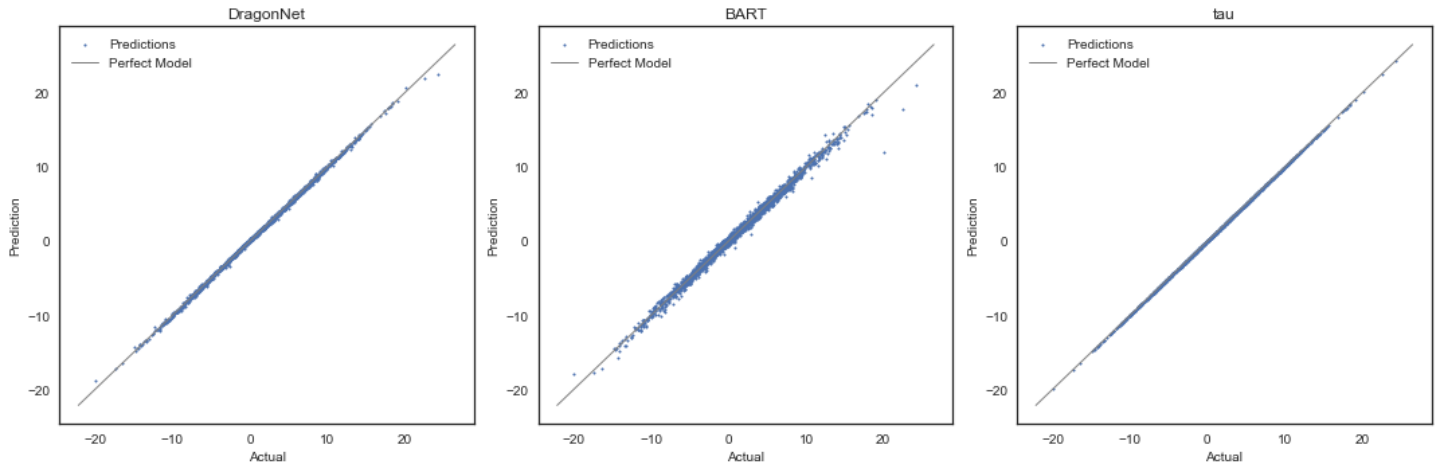

Fig. 9: Actual vs. Predicted CATEs (Validation Set).

Yanji Du

Simulation 3 Results

Dragonnet performed comparatively well with BART (lower MSE, higher AUUC) (Table 3). The true distribution of CATE predictions by observation in the validation set are not perfectly mirrored by either estimator, but they are reasonably close (Fig. 10).

	ATE	MSE	Abs % Error of ATE	AUUC
Actuals	1.091060	0.000000	0.000000	NaN
DragonNet	1.098646	0.151146	0.006953	1.054937
BART	1.083671	0.238134	0.006773	1.049644

Table 3: Summary of results by estimator compared with oracle data for validation set.

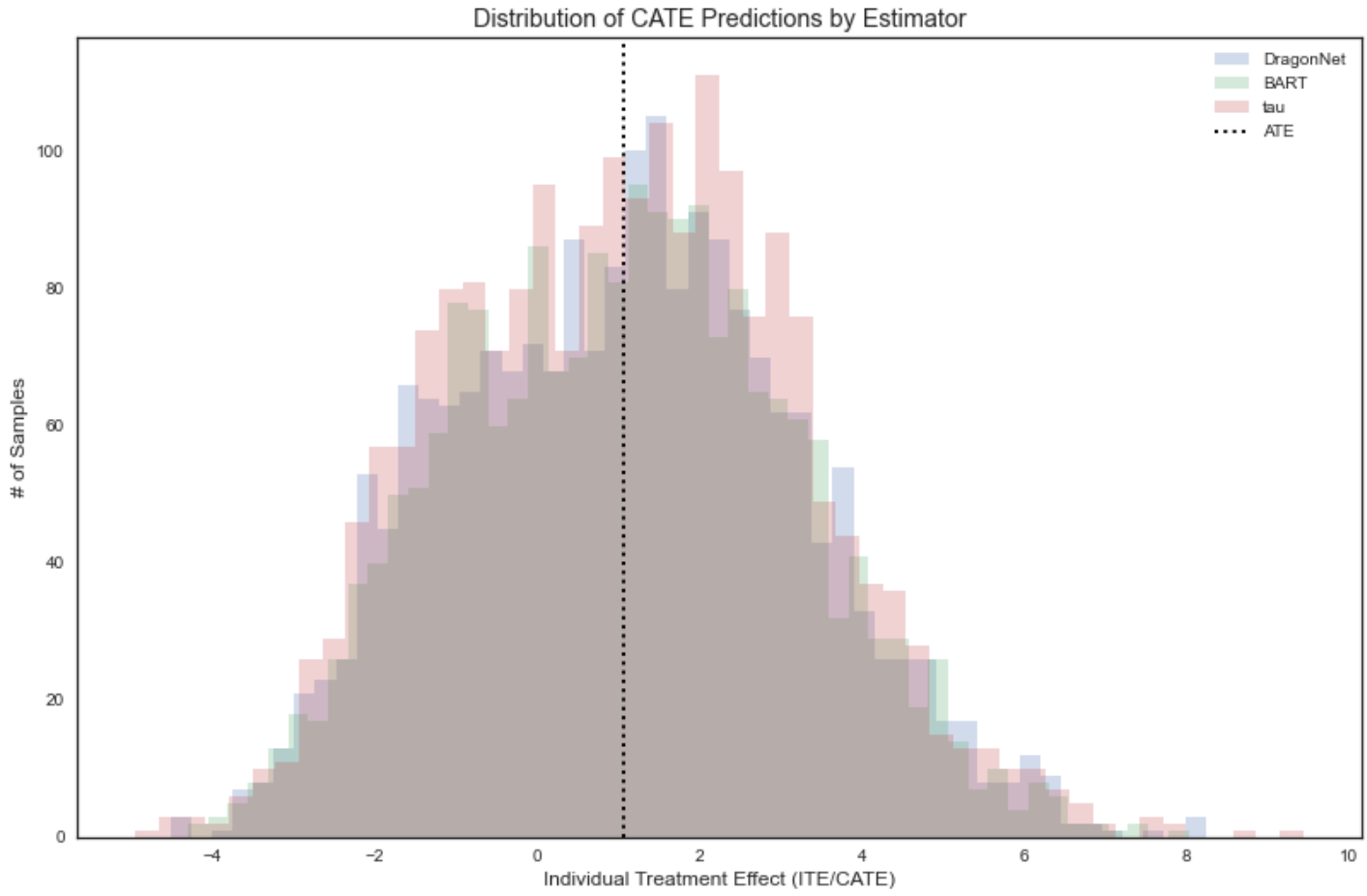

Fig. 10: Distribution of CATE predictions by estimator compared with oracle (tau) for validation set.

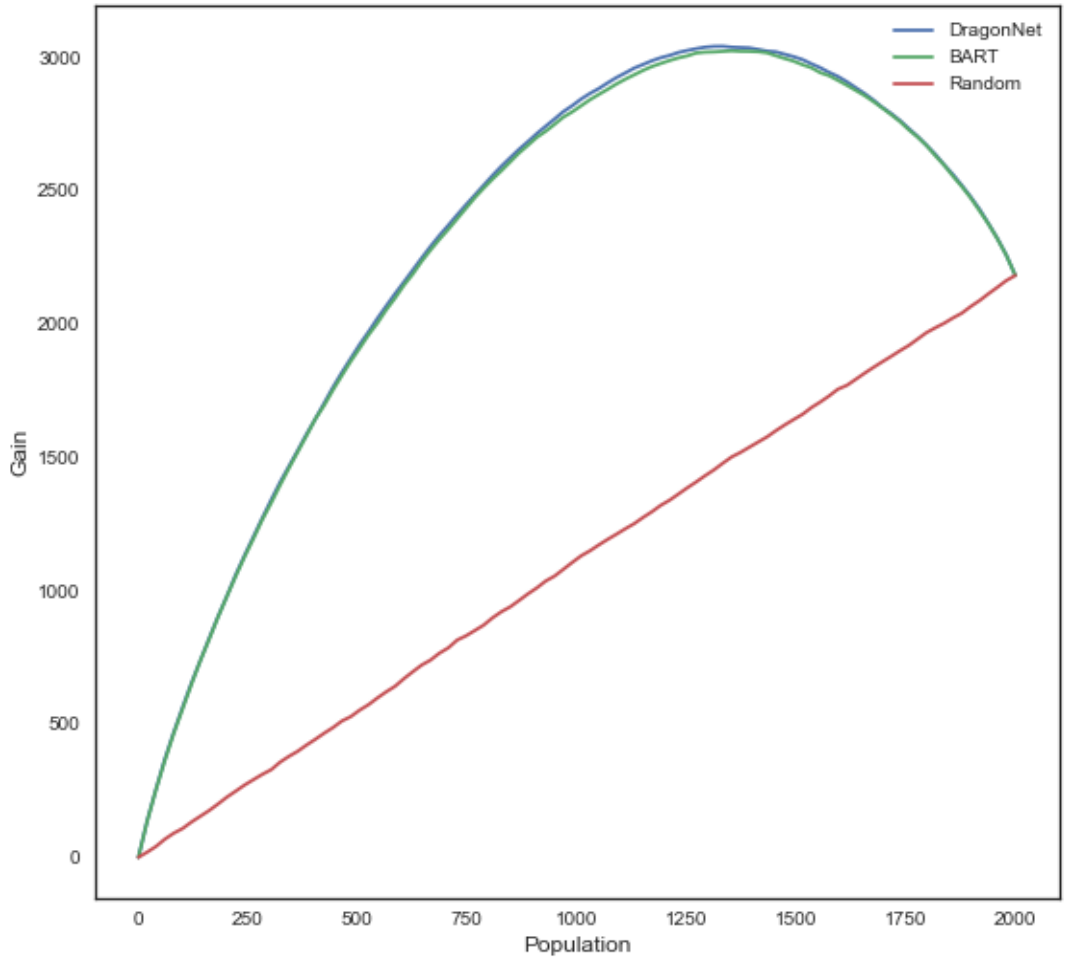

Fig. 11: Area Under Uplift Curve by Estimator (Validation Set).

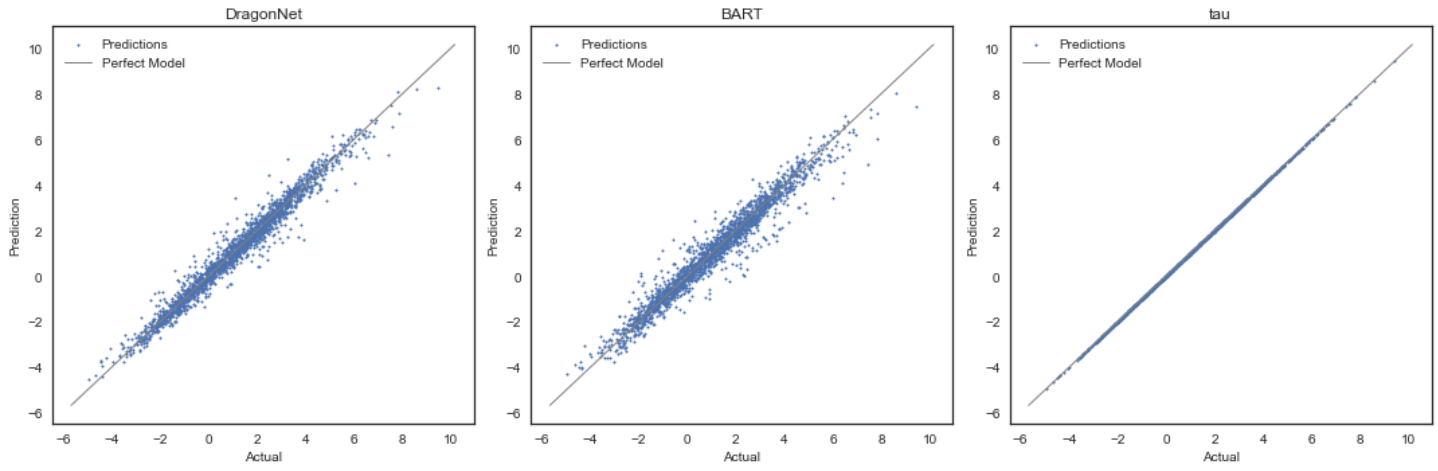

Fig. 12: Actual vs. Predicted CATEs (Validation Set).

4. Empirical Example: General Social Survey Experiment

4.1 Data description

We leverage the General Social Survey (GSS) empirical dataset as a benchmark for estimating heterogeneous treatment effects. The advantages of this dataset include: “the ATE estimate in this experiment is sizable and well established by replication studies. The sample size is large enough to allow us to investigate systematic treatment effect heterogeneity with ample statistical power” (Green and Kern 2012).

4.2 Treatment effect measurement

In this survey experiment, respondents were asked whether they supported the U.S. government spending for social support programs when labeled as either “welfare” or “assistance to the poor”. Despite the simple change of phrasing, the effect is strong and statistically significant: estimates of the treatment effect range from 23.0 to 48.6 percentage points.

year	N (Assistance)	N (Welfare)	Mean (Assistance)	Mean (Welfare)	ATE
1986	678	661	0.096	0.424	0.328
1988	470	426	0.070	0.423	0.352
1989	460	427	0.093	0.431	0.337
1990	630	598	0.073	0.401	0.328
1991	463	451	0.112	0.381	0.269
1993	498	488	0.131	0.584	0.453
1994	871	898	0.148	0.634	0.486
1996	858	860	0.181	0.587	0.407
1998	847	806	0.117	0.444	0.327
2000	819	824	0.111	0.396	0.285
2002	425	401	0.087	0.436	0.349
2004	416	416	0.060	0.440	0.380
2006	914	907	0.084	0.362	0.277
2008	615	586	0.073	0.370	0.297
2010	626	659	0.096	0.408	0.312
2012	569	580	0.104	0.464	0.360
2014	785	747	0.113	0.474	0.361
2016	874	832	0.077	0.452	0.375
2018	680	701	0.074	0.392	0.319
2021	1204	1154	0.091	0.321	0.230

Table 4: Average treatment effect on percentage of respondents indicating whether “too much” was spent on social programs labeled as “welfare” (y1) as compared to “assistance to the poor” (y0).

Similar to the approach Green and Kern used with BART, we use the fitted Dragonnet model to produce CATE predictions for each observation based on its covariates. The distribution of the CATE predictions are shown in Fig. 13. We see that though the ATE is 0.338, there is a noteworthy spread of CATE depending on the specific characteristics of each observation (the treatment effect is not the same for each respondent).

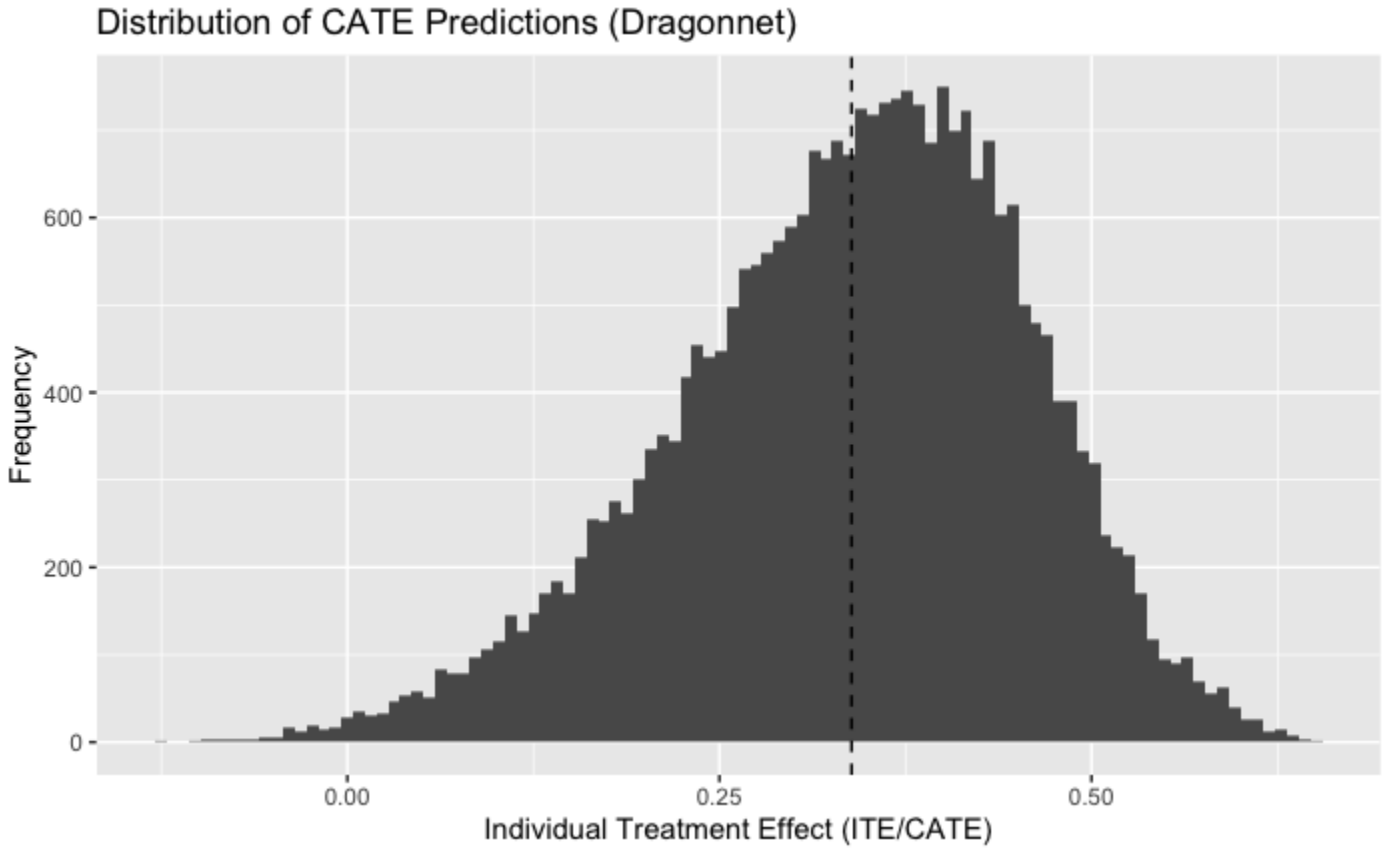

Fig. 13: Distribution of Dragonnet CATE predictions on GSS data.

Additionally, similar to Green and Kern, we also use the Dragonnet model to produce CATE estimates by a few selected covariates in Fig. 14 (partyid = party identification where higher is stronger publican and lower is strong democrat; polviews = political views scale where higher is more conservative and lower is more liberal; age = age of the respondent; educ = number of years of education of the respondent; attblack = negative attitudes towards blacks scale where higher means more negative; year = year that the survey was given).

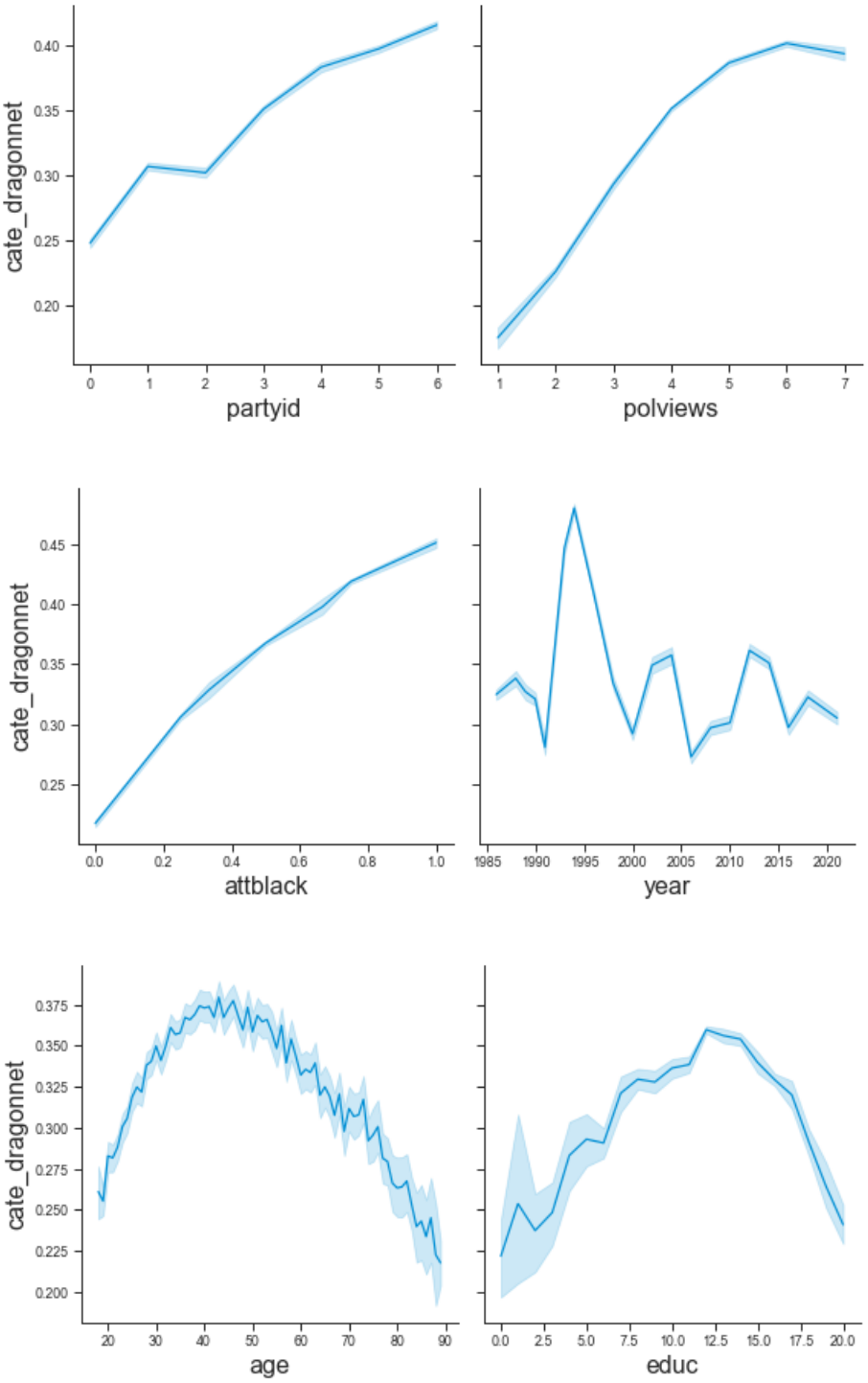

Fig. 14: CATE estimates by selected covariates.

Fig. 14 demonstrates substantial treatment effect heterogeneity. On the top row, the figures show that Democrats and those with liberal views are more likely to have a lower treatment effect, while

Yanji Du

Republicans and those with conservative views are more likely to have a stronger treatment effect. In the middle row, we see a strong treatment effect heterogeneity with the "negative attitudes towards blacks" scale: since many associate welfare with minorities, it follows that the treatment effect is stronger when the respondent has a more negative view of blacks. There appears to be a cyclical pattern with the year variable, peaking during Bill Clinton's years of Presidency. In the bottom row, the age of the respondent interestingly shows a non-monotonic treatment effect heterogeneity: the treatment effect starts low in early adulthood, appears to peak at around age 40-50, and then decreases from there as the respondent ages. Adults are typically most self-sufficient socially and financially in ages 40-50 and may explain the aversion towards welfare. A similar non-monotonic trend is seen in years of education: there is a peak at 12 years of education, which typically corresponds to completing high school.

Below Fig. 15 plots the "area under uplift curve" (AUUC), where respondents are sorted based on predicted individual treatment effect and evaluated how much the top group differs from the bottom ones in terms of treatment effect compared to random.

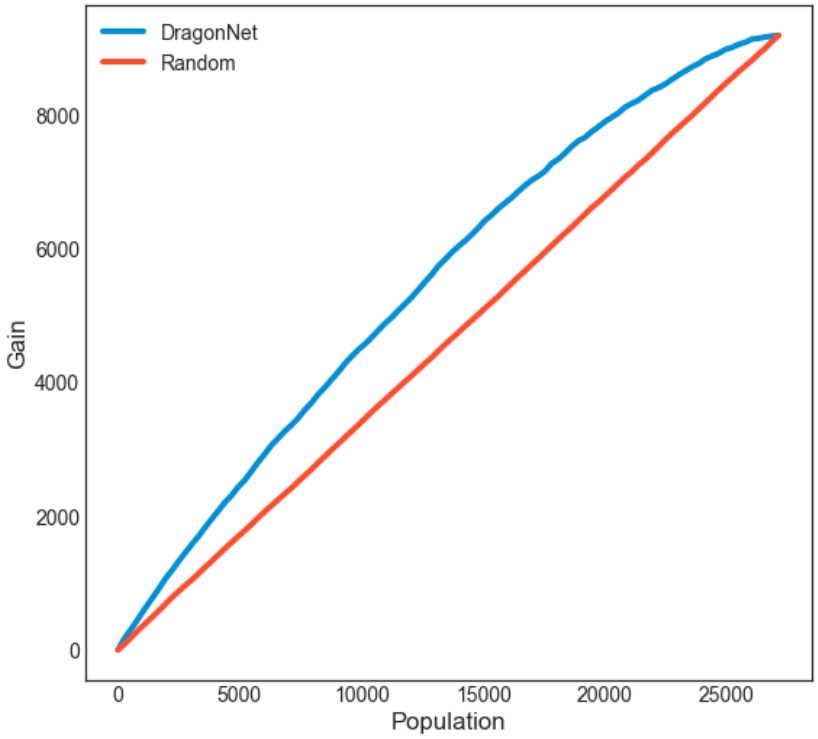

Fig. 15: Dragonnet area-under-uplift-curve (AUUC).

We also compare the CATE estimates of Dragonnet with those produced by BART. Fig. 16 below shows a replication of Green and Kern's application of BART on the GSS dataset, including data up until the year 2021.

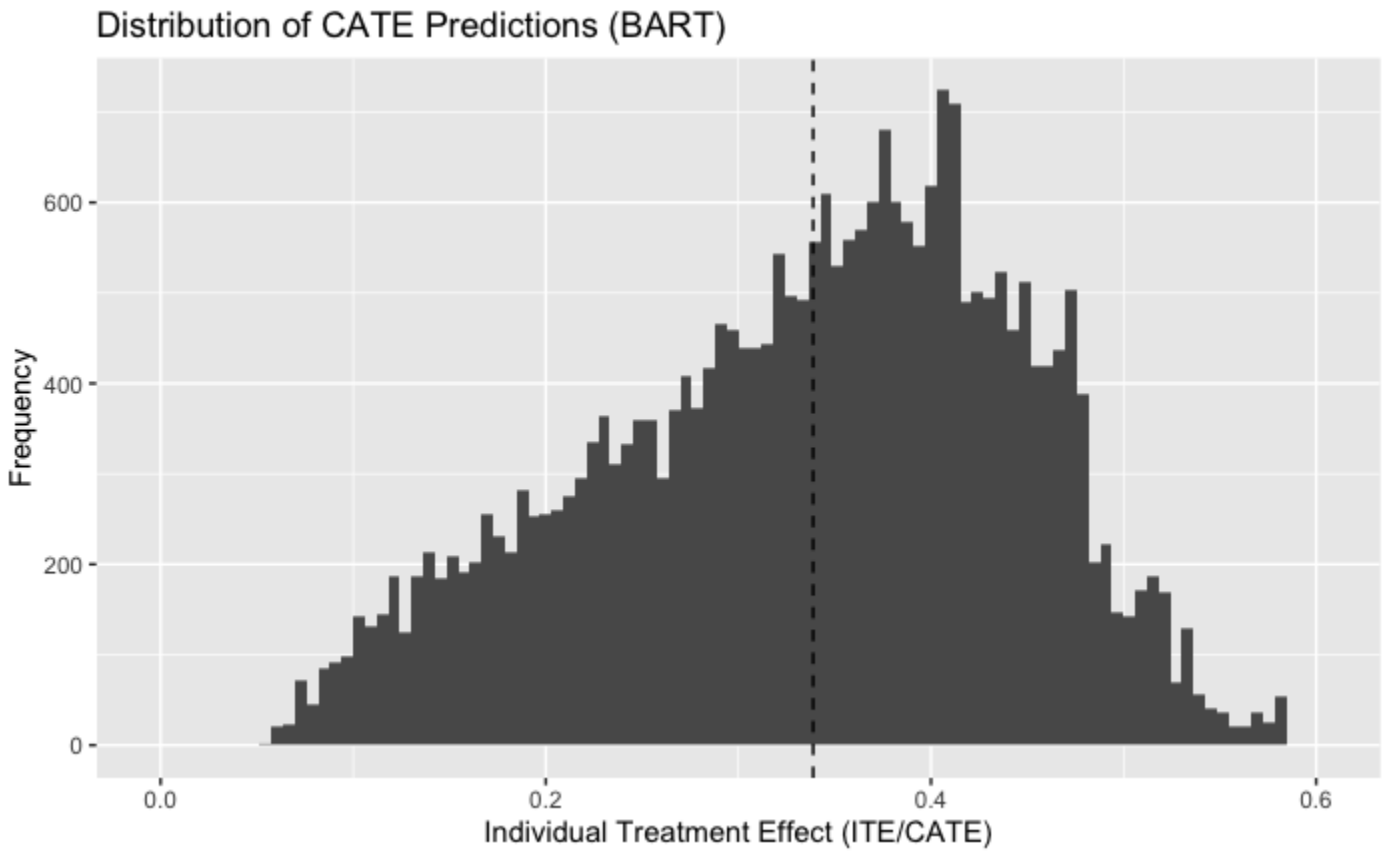

Fig. 16: Distribution of BART CATE predictions on GSS data.

Fig. 17 below shows the correlation between the Dragonnet and BART CATE estimates on the GSS dataset. With a correlation of 0.860, the estimates are quite close to each other. Additionally, differences between the model estimates appear to be evenly distributed across the range of estimates (i.e. there does not appear to be bias in certain regimes of treatment effects).

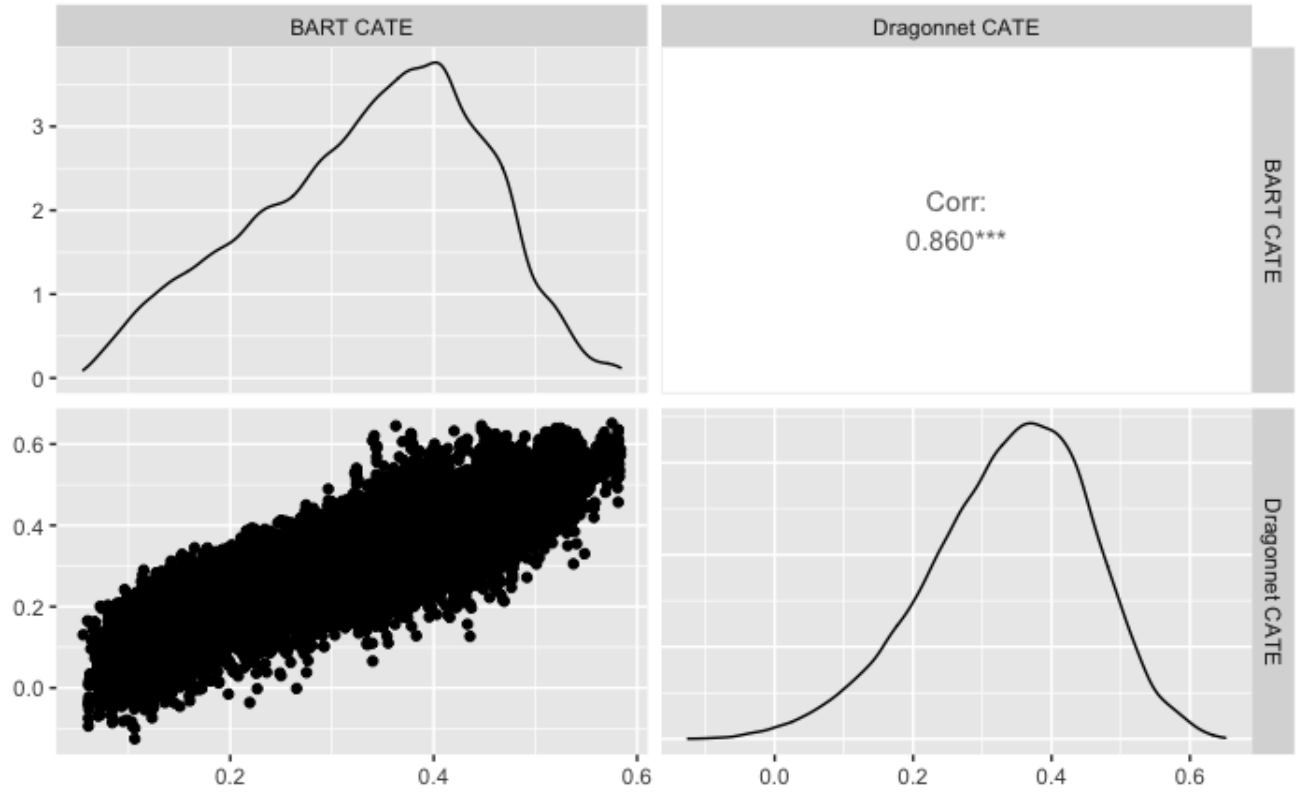

Fig. 17: Correlation between BART and Dragonnet CATE estimates on GSS dataset.

4. Further analysis on empirical differences between BART and Dragonnet CATE estimates

We compare the differences of BART and Dragonnet CATE estimates on GSS across different covariate ranges. We first ran a regression of the absolute difference between BART and Dragonnet CATE estimates on the GSS study covariates. From Table 5, we see that all covariates (except for 'attblack') are statistically significant ($p < 0.05$).

Table 5. Regression Analysis: Absolute CATE difference between BART and Dragonnet (dependent variable) on the GSS covariates (independent variables).

Term	*b*	*SE*	*t*	*p*	95% CI

(Intercept)	-1.01634	0.04238	-23.98	**< .001**	[-1.10, -0.93]
year	0.00052	0.00002	24.61	**< .001**	[0.00048, 0.00056]
age	0.00005	0.00001	3.63	**< .001**	[0.00002, 0.00007]
educ	0.00046	0.00008	5.79	**< .001**	[0.00030, 0.00061]
partyid	0.00128	0.00013	10.05	**< .001**	[0.00103, 0.00152]
polviews	0.00060	0.00018	3.30	**.001**	[0.00024, 0.00095]
attblack	0.00035	0.00082	0.43	.665	[-0.00124, 0.00195]

- Half sample analysis

  We also conducted a half sample analysis. We split the data randomly in half and retrained both models on each half. We sorted the predictions from each model from low to high, and compared the correlation between BART vs. BART (Figure 21) and Dragonnet vs. Dragonnet (Figure 22). The split data predictions have very high correlation (0.992 - 0.993) and shows both BART and Dragonnet have high reliability, giving approximately the same CATEs each time.

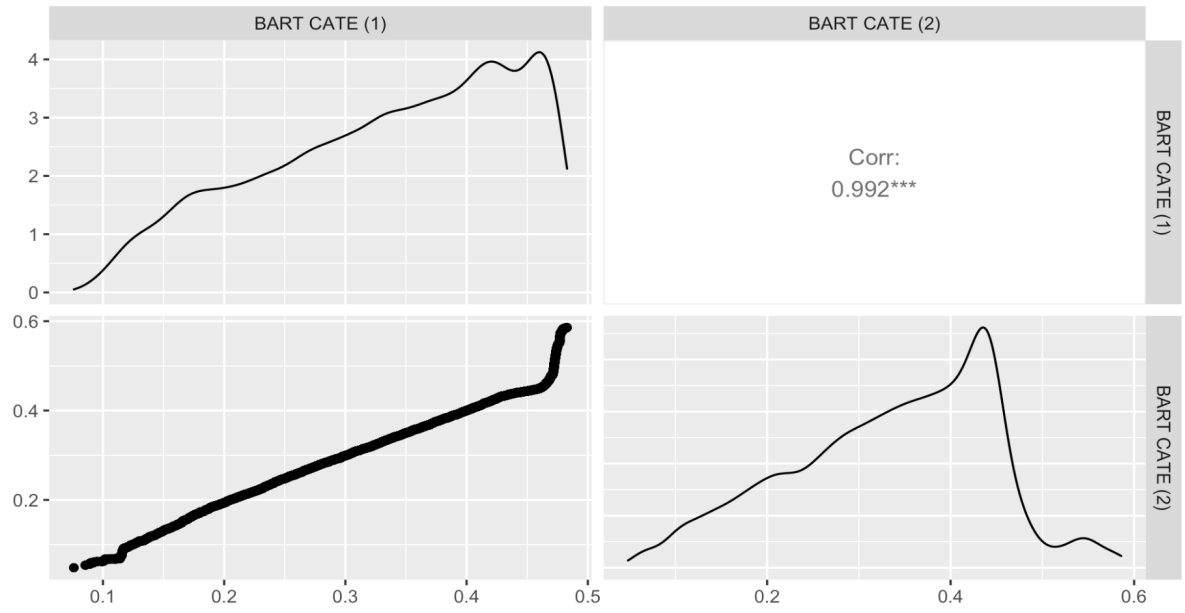

Fig 21: BART vs BART correlation

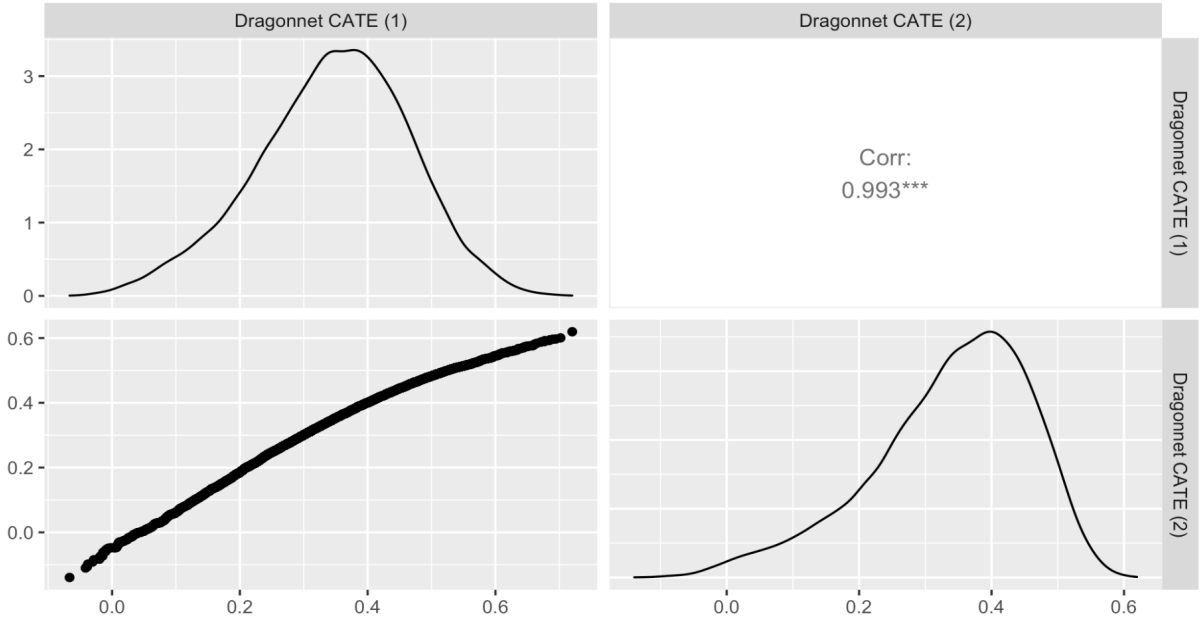

Fig. 18: Dragonnet vs Dragonnet correlation

- Residual vs residual plot

In this case, our outcome Y is defined as the difference in CATE between the BART- and Dragonnet-predicted CATEs for each observation. For each covariate in our model, we plot the residuals of Y when regressing on all other covariates against the residuals of the predictor

variable after regressing it on all other covariates. The plots in Figure 19 suggest a very negligible relationship between the differences caused by two models.

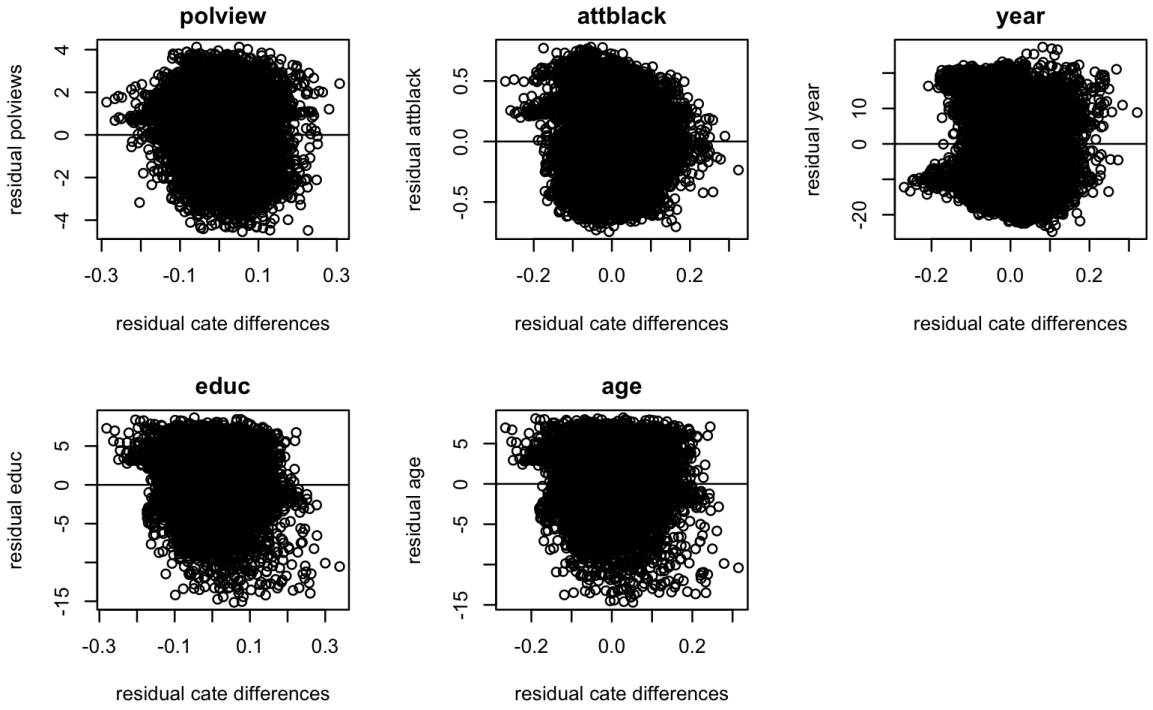

Fig. 19: residual vs residual plot

5. Conclusion and future study

Our simulation studies and empirical example shows that Dragonnet can effectively recover treatment effect heterogeneity under scenarios of varying complexity. The performance of Dragonnet was comparable to BART in the simulation studies when evaluating the estimators based on MSE and AUUC, with Dragonnet slightly outperforming BART in all simulation scenarios. The distribution of CATE predictions appeared to closely mirror the true treatment effect distribution in all simulation scenarios.

In the empirical example using GSS data, Dragonnet recovered a CATE distribution similar to the one Green and Kern produced with BART. The treatment effect heterogeneity by covariates were closely

mirrored as well. The results confirm the promise of neural-network based approaches for the purposes of treatment effect estimation.

Further, in both simulation and empirical studies, we did not use extensive parameter tuning to produce the CATE estimates, also following the "off-the-shelf" nature of BART to minimize researcher discretion. It's possible that parameter tuning may be necessary to produce good CATE predictions with more complex treatment-by-covariate interactions; we leave this area for future research.

We explored the differences between BART CATE and Dragonnet CATE at different covariate levels. Though CATE estimate differences are not uniform across covariate ranges, these estimation differences are negligible. Dragonnet proved to be a reliable alternative for the GSS survey study and maybe other similar experiment settings. In order to seek a more clear understanding of the differences between BART and Dragonnet, we may need to explore other simulation scenarios.

References

Green, D.P., & Kern, H.L. 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian Additive Regression Trees. Public Opinion Quarterly 76: 491-511.

Shi, C., Blei, D.M., & Veitch, V. 2019. Adapting Neural Networks for the Estimation of Treatment Effects. NeurIPS.