Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

February 28, 2024

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

Estimation and Validation of Heterogeneous Treatment Effects

15

"You can, for example, never foretell what any one man will do, but you can say with precision what an average number will be up to."

– Sherlock Holmes [1].

We study flexible estimation of heterogeneous treatment effects. We target the construction of an estimate of the true CATE function and not its projection on a simpler function space, with as small root-mean-squared-error as possible. We consider flexible estimation using generic ML techniques and discuss how one can perform model selection and out-of-sample validation of the quality of the learned model of heterogeneity. We conclude with the topic of policy learning, i.e. constructing optimal personalized policies.

15.1 ML Methods for CATE Estimation

We consider the same setting as in Chapter 14 of analyzing the heterogeneous effect of a binary treatment in the presence of a high-dimensional set of observed controls Z, under conditional exogeneity. In this section, we target the construction of an estimate $\hat{\tau}(X)$ of the true CATE function $\tau_0(X)$ and not its best linear approximation, using generic ML techniques, in a manner such that the mean squared error $\mathrm{E}_X(\tau_0(X) - \hat{\tau}(X))^2$, which is also what we used in Chapter 10 to measure the quality of non-linear predictive ML models, is minimized. We will also be interested in the mean squared error of the estimate with respect to the best approximation of the CATE over some flexible, potentially non-linear function space T, i.e. the function τ_* defined as

$$\tau_* = \arg\min_{\tau \in T} \mathrm{E}_X(\tau_0(X) - \tau(X))^2. \tag{15.1.1}$$

As in the previous section, the key is to decompose the estimation of the CATE into a sequence of regression problems. Then generic ML techniques can be used to address each of these regression problems. This approach has been coined *meta-learning* in the literature on CATE estimation, since we are trying to treat ML techniques as a black-box oracle that solves any regression problem and we are trying to build on top of that oracle to learn the CATE. Motivated by the ability to construct confidence intervals, in the previous section, we provided one such choice of a reduction, as we will explain later. However, when one is primarily interested in mean squared error, other decompositions could potentially have better finite sample performance. We present here the multitude of such meta-learning approaches that have been proposed in the literature and we will conclude with a comparative analysis of each of them.

Meta-Learning Strategies for CATE Estimation

To simplify the exposition, and emphasizing the meta-learning aspect of these methods, we will introduce a notation for a regression estimate oracle. We denote with $O_H(\{X_i, Y_i, W_i\}_{i=1}^n)$ an oracle algorithm that takes as input a dataset of n i.i.d. samples, consisting of covariatex X_i, regression labels Y_i and sample weights W_i (where weights are assumed to be independent of Y_i given X_i) and produces an estimate $\hat{h}$ of the function that

minimizes the sample-weighted square loss:

$$\begin{aligned} h_0 &= \arg\min_{h \in H} \mathrm{E} W (Y - h(X))^2 \\ &= \arg\min_{h \in H} \mathrm{E} W (\mathrm{E}(Y \mid X) - h(X))^2 \end{aligned} \tag{15.1.2}$$

over some function space H. When sample weights W_i are omitted, they will be assumed to be equal to 1. This oracle will typically correspond to some ML approach to solving this weighted regression problem and we will be assuming that such an oracle provides an estimate $\hat{h}$ that converges to h_0 at some rate, with respect to the mean-squared-error metric, i.e.

$$\|\hat{h} - h_0\|_{L^2(X)} = r_n \to 0.$$

Single (S)-Learner Starting from the very simple identification formula for the CATE in Equation (14.1.1), we can learn the CATE by first invoking an ML regression method to construct an estimate $\hat{g}$ of the conditional expectation function $g_0(D, Z) := \mathrm{E}[Y \mid D, Z]$, assuming that g_0 lies in some function space G. Then we can construct a model $\hat{\tau}$ of the CATE by invoking an ML regression method to construct an estimate of the conditional expecation function $\mathrm{E}[\hat{g}(1, Z) - \hat{g}(0, Z) \mid X]$, over some function space T. Overall we arrive at the following meta-learning algorithm:

Single Learner (S-Learner)

$$\begin{aligned} \hat{g} &:= O_G(\{(D_i, Z_i), Y_i\}_{i=1}^n) \\ \hat{\tau} &:= O_T\left(\{X_i, \hat{g}(1, Z_i) - \hat{g}(0, Z_i)\}_{i=1}^n\right) \end{aligned} \tag{15.1.3}$$

Even if T does not contain τ_0, as long as $g_0 \in G$ and $\|\hat{g} - g_0\|_{L^2(D,Z)} \to 0$, the S-Learner estimate will be converging to the best approximation of the CATE within the space T, i.e. $\|\hat{\tau} - \tau_*\|_{L^2(X)} \to 0$.

Two (T)-Learner Estimating a single regression model that predicts the outcome Y from the treatment D and the controls Z, can overly regularize the treatment variable. Especially in settings where the treatment has a small effect, many ML algorithms will most probably shrink the treatment effect to zero and prioritize the inclusion of other informative covariates in the selected model. For this reason, it seems natural to weaken this regularization bias on the treatment. This can be achieved

by fitting two separate models, one model $\hat{g}^T$ that estimates the relationship between the outcome Y and the covariates Z within the treated group, i.e. $g_0^T := \mathrm{E}[Y \mid Z, D = 1]$ and one model $\hat{g}^C$ that learns the same relationship within the control group, i.e. $g_0^C - := \mathrm{E}[Y \mid Z, D = 0]$. Then the CATE can be estimated by invoking an ML regression method to construct an estimate of the CEF $\mathrm{E}[\hat{g}^T(Z) - \hat{g}^C(Z) \mid X]$. Overall, we arrive at the following meta-learning algorithm:

Two Learner (T-Learner)

$$\begin{aligned} \hat{g}^C &:= O_{G^C}(\{Z_i, Y_i\}_{i \in \{1,\dots,n\}:D_i=0}) \\ \hat{g}^T &:= O_{G^T}(\{Z_i, Y_i\}_{i \in \{1,\dots,n\}:D_i=1}) \\ \hat{\tau} &:= O_T(\{X_i, \hat{g}^T(Z_i) - \hat{g}^C(Z_i)\}_{i=1}^n) \end{aligned} \tag{15.1.4}$$

Similar to the S-Learner, as long as $g_0^C \in G^C$ and $g_0^T \in G^T$, then the result of the T-Learner will always be converging to τ_*, i.e. the best approximation of the CATE within T.

Doubly Robust (DR)-Learner The above approaches rely fully on accurate outcome modelling. If we face settings where the conditional counterfactual outcomes $\mathrm{E}[Y \mid D = 1, Z]$ are complicated functions that are hard to model and estimate, but the CATE function $\tau(X)$ is relatively simple, the aforementioned two meta-learners will suffer from large estimation errors in $\hat{g}, \hat{g}^T, \hat{g}^C$. If for instance, we are in a randomized controlled trial and we know the propensity μ_0, then we also know that the random variable $YH(\mu_0)$ satisfies

$$\begin{aligned} \mathrm{E}[YH(\mu_0) \mid X] &= \mathrm{E}[\mathrm{E}[YH(\mu_0) \mid Z] \mid X] \\ &= \mathrm{E}[\mathrm{E}[Y \mid D = 1, Z] - \mathrm{E}[Y \mid D = 0, Z] \mid X] \\ &= \tau(X). \end{aligned}$$

Thus when solving this regression problem we only need to be accurately approximating the potentially simpler CATE function, as opposed to the response functions under treatment or control.[1]

Beyond randomized control trials, the above approach is too heavily dependent on constructing a good estimate $\hat{\mu}$ of the propensity score. Moreover, even for randomized control trials, the latter method can have very large variance, due to dividing the outcome Y by the inverse propensity. For this reasons, it might be beneficial even when we care solely about mean

1: An estimation strategy based on running a regression of $YH(\hat{\mu})$ on X, is referred to in the literature as the Inverse Propensity Score (IPS)-Learner, but we will omit more details on it.

squared error, to use the doubly robust approach, which combines propensity and regression modelling and can reduce both bias due to errors in estimating the propensity and variance by dividing only the un-explained variation in the outcome by the propensity. This leads to the doubly robust meta-learner (we will describe its two-learner variant, which is advisable in practice):

Doubly Robust Learner (DR-Learner)

$$\begin{aligned}
\hat{g}^C &:= O_{G^C}(\{Z_i, Y_i\}_{i\in\{1,\dots,n\}:D_i=0}) \\
\hat{g}^T &:= O_{G^T}(\{Z_i, Y_i\}_{i\in\{1,\dots,n\}:D_i=1}) \\
\hat{\mu} &:= O_M(\{Z_i, D_i\}_{i\in\{1,\dots,n\}}) \\
\hat{g}(D, Z) &:= \hat{g}^T(Z)\, D + \hat{g}^C(Z)\,(1 - D) \\
Y_i(\hat{\eta}) &:= H_i(\hat{\mu})\,(Y_i - \hat{g}(D_i, Z_i)) + \hat{g}^T(Z_i) - \hat{g}^C(Z_i) \\
\hat{\tau} &:= O_T\left(\{X_i, Y_i(\hat{\eta})\}_{i=1}^n\right)
\end{aligned} \tag{15.1.5}$$

The previous section can be seen as a special of this meta-learner, where we use OLS as our regression oracle in the final step and were we use cross-fitting when we estimate the first three steps and calculate the proxy labels $\hat{Y}_i$.

The DR-Learner inherits certain type of double robustness properties that we have, even when analyzing the mean squared error. In particular, suppose that we knew the true regression functions g_0^C, g_0^T, μ_0 and based on some appropriate argument, we could show that the regression oracle in the last step, when ran with the ideal labels $Y(\eta_0)$, achieves a mean squared error rate of the order of r_n^2. Then, assuming $g_0^C \in G^C$, $g_0^T \in G^T$ and $\mu_0 \in M$, one can argue, under benign regularity conditions, that the mean squared error of the DR-Learner, can be upper bounded as:

$$\mathrm{E}(\hat{\tau}(X) - \tau_*(X))^2 \lesssim r_n^2 + \mathrm{Error}(\hat{g})^2 \cdot \mathrm{Error}(H(\hat{\mu}))^2$$

where the error of these nuisances can always be taken to be the fourth moment of the prediction error[2] and under further regularity conditions on the function space T used in the estimation for τ, it can be taken to be the root-mean-squared-error.[3] Thus as long as the product of the errors in modelling the regression and the propensity function are small, then the mean squared error for $\hat{\tau}$ will not be significantly impacted by these first stage estimation errors. For formal versions of variants of such results, we defer the reader to the following papers [2–6].

2: For any estimate $\hat{f}$ of a function f_0, that takes as input some random variable W, the fourth moment of the prediction error $\|\hat{f} - f_0\|_{L^4(X)}$ is defined as

$$\left(\mathrm{E}_W(\hat{f}(W) - f_0(W))^4\right)^{1/4}$$

and is a slightly strong measure of performance that the root-mean-squared-error, i.e.

$$\sqrt{\mathrm{E}_W(\hat{f}(W) - f_0(W))^2}$$

.

3: In fact, one can always use the slightly better error metric:

$$\left(\mathrm{E}_X\left[\mathrm{E}_W[(\hat{f}(W) - \hat{f}(W))^2 \mid X]^2\right]\right)^{1/4}$$

where X is the set of variables that enter the CATE function τ. When X is the empty set, as in the case of average causal effects this boils down to the mean squared error and otherwise this requires better control, on average, of the conditional mean squared error of the nuisance functions, conditional on the variables X that enter the CATE function.

Residual (R)-Learner If we know that the true CATE model lies in a simple function space and even if we knew the true nuisance parameters η_0, the labels that are used in the final stage of the DR-Learner can still have a large magnitude, due to the division by the propensity. In settings were the overlap assumption is almost violated at particular regions of the covariate space, the regression labels $Y_i(\eta_0)$ will be taking very large values in absolute magnitude. This can lead to a high-variance estimate. For instance, if we knew that the treatment effect is constant, then we are essentially assuming the partially linear regression model and we shouldn't be using the doubly robust method, but rather the residual-on-residual method, which minimizes the loss $\mathrm{E}(\hat{Y}_i - \tau \hat{D}_i)^2$, where $\hat{Y} = Y - \mathrm{E}[Y \mid Z]$ and $\hat{D} = D - \mathrm{E}[D \mid Z]$. Similarly, if we are willing to assume that the CATE function is linear in some engineered features of only the variables X, i.e. $\delta(Z) = \tau(X) = \beta' p(X)$, then we should instead be estimating a linear interactive model, where we interact the treatment with the engineered features and apply the residualization apporach to arrive at the loss function

$$\mathrm{E}(\hat{Y}_i - \beta' p(X)\, \hat{D}_i)^2$$

since $p(X)D - \mathrm{E}[p(X)D \mid Z] = p(X)\,(D - \mathrm{E}[D \mid Z]) = p(X)\,\hat{D}$.

Analogously, if we know that the high-dimensional CATE function $\delta_0(Z) = \mathrm{E}[Y(1) - Y(0) \mid Z]$, is only a function of the variables X, i.e. $\delta_0(Z) = \tau_0(X)$ and τ_0 lies in some simple space T, a lower variance loss function, than the doubly robust loss function would be:

$$\min_{\tau \in T} \mathrm{E}(\hat{Y}_i - \tau(X)\hat{D}_i)^2$$

As we already showed in Equation (14.4.4) in Section 14.4, under the aforementioned assumptions we can write the regression equation:

$$Y = \tau_0(X)\, D + g_0(Z) + \epsilon, \qquad \mathrm{E}[\epsilon \mid D, Z] = 0$$

Thus, we are faced with a non-linear regression equation, regressing $\hat{Y}$ on $\hat{D}, X$, where we know that the CEF is of the form $\mathrm{E}[\hat{Y} \mid \hat{D}, X] = \tau(X)\, \hat{D}$, for some function τ in some simple function space T. To estimate this regression problem, we should thus minimize the square loss, over the space of such CEFs, i.e.

$$\min_{\tau \in T} \mathrm{E}(\hat{Y} - \tau(X)\hat{D})^2, \tag{15.1.6}$$

which is exactly the R-Learner loss.

When taken to estimation, the residuals $\hat{Y}, \hat{D}$ will be replaced by the estimated residuals $\check{Y}, \check{D}$, where $\check{Y} = Y - \hat{h}(Z)$ and $\check{D} = D - \hat{\mu}(Z)$, with $\hat{h}$ being an estimate of the CEF $\mathrm{E}[Y \mid Z]$ (e.g. one could use the two-learner based estimate

$$\hat{h}(Z) := \hat{g}^T(Z)\,\hat{\mu}(Z) + \hat{g}^C(Z)\,(1 - \hat{\mu}(Z)).$$

or a direct regression, regressing Y on Z. Moreover, note that minimizing the R-Learner loss, is equivalent to minimizing a sample-weighted square loss, where the covariates are X, the labels are $\hat{Y}/\hat{D}$ and the weights are $\hat{D}^2$:

$$\mathrm{E}(\hat{Y} - \tau(X)\hat{D})^2 = \mathrm{E}\hat{D}^2\,(\hat{Y}/\hat{D} - \tau(X))^2,$$

Thus the final step in the R-Learner also corresponds to a sample-weighted regression oracle problem. This leads to the following meta-learner algorithm:

Residual Learner (R-Learner)

$$\begin{aligned}
\hat{h} &:= O_H(\{Z_i, Y_i\}_{i\in\{1,\dots,n\}}) \\
\hat{\mu} &:= O_M(\{Z_i, D_i\}_{i\in\{1,\dots,n\}}) \\
\check{Y}_i &:= Y_i - \hat{h}(Z_i) \\
\check{D}_i &:= D_i - \hat{\mu}(Z_i) \\
\hat{\tau} &:= O_T\left(\left\{X_i, \check{Y}_i/\check{D}_i, \check{D}_i^2)\right\}_{i=1}^n\right)
\end{aligned} \tag{15.1.7}$$

Under the assumption that $\delta_0 = \tau_0 \in T$, and that $h_0 \in H$, $\mu_0 \in M$, then the R-Learner converges to the true CATE τ_0. Moreover, this approach inherits similar robustness properties as the partialling out approach for the case of estimating average causal effects. In particular, if we let r_n^2 denote the mean squared error that the final regression oracle would have achieved had we known the true nuisance parameters h_0, μ_0, then under regularity conditions, one can show that:

$$\mathrm{E}_X(\hat{\tau}(X) - \tau_0(X))^2 \lesssim r_n^2 + \mathrm{Error}(\hat{\mu})^4 + \mathrm{Error}(\hat{\mu})^2\,\mathrm{Error}(\hat{h})^2$$

Unlike the DR-Learner, we see here that accurate estimation of the propensity is more important and cannot be compensated by more accurate estimation of the outcome regression problem. Similar to the DR-Learner, the error function in the above claim can always be taken to be the fourth moment of the prediction error and under further restrictions on the function space T, it can be taken to be the root-mean-squared-error. For formal versions of this claim see [4, 6, 7].

One may also wonder what does the R-Learner estimate when the assumption that $\delta_0 = \tau_0$ or that $\tau_0 \in T$ is violated. Unlike all prior meta-learners, the R-Learner does not converge necessarily to the best approximation τ_* of the CATE within T. For instance, consider the extreme case where T contains only constant functions. Then we are estimating an average treatment effect based on a partialling out approach, while the partial linear response function does not hold and there exists treatment effect heterogeneity. In this case, the partialling out approach will not be converging to the average causal effect and similarly for any T, the R-Learner will not be converging to τ_*.

To understand the limit point of the R-Learner, let us examine the R-Learner loss as defined in Equation (15.1.6). By construction, $\hat{\tau}$ will be converging to the solution to that minimization problem. As we have already argued, under conditional exogeneity, we can always write $\hat{Y} = \delta_0(Z)\hat{D} + \epsilon$, with $\mathrm{E}[\epsilon \mid \hat{D}, Z] = 0$. Thus we can re-write the R-Learner loss as:

$$\begin{aligned}\mathrm{E}(\hat{Y} - \tau(X)\hat{D})^2 &= \mathrm{E}(\delta_0(Z)\hat{D} - \tau(X)\hat{D})^2 + \mathrm{E}\epsilon^2 \\ &= \mathrm{E}\left[(\delta_0(Z) - \tau(X))^2 \operatorname{Var}(D \mid Z)\right] + \mathrm{E}\epsilon^2\end{aligned}$$

where we used the fact that $\mathrm{E}[\hat{D}^2 \mid Z] = \operatorname{Var}(D \mid Z)$. Thus minimizing the R-Learner loss is equivalent to minimizing a treatment-variance-weighted square loss and the estimate will be converging to the best treatment-variance-weighted approximation of the high-dimensional CATE function, i.e.

$$\tilde{\tau} = \arg\min_{\tau \in T} \mathrm{E}\left[(\delta_0(Z) - \tau(X))^2 \operatorname{Var}(D \mid Z)\right] \tag{15.1.8}$$

This solution is essentially putting more weight on regions of the covariate space Z, where the treatment was more randomly assigned. If for instance parts of the population were almost always treated or almost always not treated, then these parts of the population will not be considered when constructing the best approximation. We will refer to this solution as the *best overlap-weighted approximation*, since it assigns weights to parts of the population, dependent on the degree of "overlap" (i.e. whether both treatments were observed for this part of the population). For instance, suppose that T is the space of constant functions and that the treatment is randomly assigned for some parts of the population and is essentially deterministic for other parts. Then $\tilde{\tau}$ will recover the average treatment effect of the subset of the population for which treatment was randomly assigned. On the contrary, in this case the doubly robust estimate will try to recover the average causal effect of the overall population, but because of that it will inadvertently

be very high variance and unstable, since for some parts of the population it barely ever sees one of the two treatments.

Cross (X)-Learner The Cross Learner tries to combine propensity to improve on outcome modelling in a manner qualitatively very different from the DR- or R-learner and not with the target of reducing the sensitivity to errors in the nuisance models. Rather it does so primarily motivated from an *accuracy and covariate-shift consideration*. Moreover, it begins with a very different starting point and idea. As a first one realizes that the high-dimensional CATE $\delta_0(Z)$ is the same, whether we measure it on the treated or on the control! In other words, the Conditional Average Treatment Effect on the Treated (CATT) is equal to the Conditional Average Treatment Effect on the Control (CATC), unlike the average treatment effect, which can be different due to different distributions of Z in treatment and control. This can be easily seen as, by conditional exogeneity:

$$\begin{aligned}\delta_0^T(Z) &:= \mathrm{E}[Y(1) - Y(0) \mid Z, D = 1] \\ &= \mathrm{E}[Y(1) - Y(0) \mid Z] = \delta_0(Z)\end{aligned}$$

and similarly for τ^0.[4]

Moreover, when we try to measure the CATT, then we actually observe the counterfactual under treatment and therefore we do not need to impute this counterfactual outcome (e.g. by learning g_0^1). Similarly for the CATC. Thus we can identify the CATT and CATC as:

$$\begin{aligned}\delta_0^T(Z) &= \mathrm{E}[Y - \mathrm{E}[Y \mid Z, D = 0] \mid Z, D = 1] \\ \delta_0^C(Z) &= \mathrm{E}[\mathrm{E}[Y \mid Z, D = 1] - Y \mid Z, D = 0]\end{aligned}$$

This yields two ways of identifying the CATE $\delta_0(Z)$ and any convex combination of these two solutions, would also be a valid identification strategy for the CATE. This approach, allows us to avoid having to model both response models well for all regions of the covariate space (which would be the case for the S-, T-, or DR-Learners). This can be powerfull if we know that the CATE is a much simpler function to learn than a mean counterfactual response model.

If we believe that the hard part is modelling the mean counterfactual response under some treatment but not the treatment effect, then we can use the following strategy: for parts of the covariate space Z, where we have more control data (i.e. $\mu_0(Z)$ is small), we can use the CATT strategy, which only requires estimating the mean counterfacutal response under control, i.e.

4: Note that the same wouldn't be true if we condition a subset X of Z:

$$\begin{aligned}\tau_0^T(X) &:= \mathrm{E}[Y(1) - Y(0) \mid X, D = 1] \\ &= \mathrm{E}[\mathrm{E}[Y(1) - Y(0) \mid Z] \mid X, D = 1] \\ &= \mathrm{E}[\delta(Z) \mid X, D = 1] \\ &\neq \mathrm{E}[\delta(Z) \mid X] =: \tau_0(X)\end{aligned}$$

$\mathrm{E}(Y \mid Z, D = 0)$, but not under treatment. Of course, we still have to learn the effect function using only the treated data, which we don't have that many in this part of Z, but since we believe that the effect function is simple, this is a more benign problem. Similarly, if for parts of the covariate space Z, we have more treated data (i.e. $\mu_0(Z)$ is large), we can use the CATC strategy, which only requires estimating the mean counterfactual response under treatment, i.e. $\mathrm{E}(Y \mid Z, D = 1)$, but not under control. This motivates using the following convex combination as our final identification formula for the CATE:

$$\delta_0(Z) = \delta_0^T(Z)\,(1 - \mu_0(Z)) + \delta_0^C(Z)\,\mu_0(Z)$$

Subsequently, for any subset X of Z, we can use the fact that $\tau_0(X) = \mathrm{E}[\delta_0(Z) \mid X]$. This identification strategy leads to the following meta-learning estimation strategy:

Cross Learner (X-Learner)

$$\begin{aligned}
\hat{g}^C &:= O_{G^C}(\{Z_i, Y_i\}_{i\in\{1,\dots,n\}:D_i=0}) \\
\hat{g}^T &:= O_{G^T}(\{Z_i, Y_i\}_{i\in\{1,\dots,n\}:D_i=1}) \\
\hat{\mu} &:= O_M(\{Z_i, D_i\}_{i\in\{1,\dots,n\}}) \\
\hat{\delta}^C &:= O_\Delta(\{Z_i, \hat{g}^T(Z_i) - Y_i\}_{i\in\{1,\dots,n\}:D_i=0}) \\
\hat{\delta}^T &:= O_\Delta(\{Z_i, Y_i - \hat{g}^C(Z_i)\}_{i\in\{1,\dots,n\}:D_i=1}) \\
\hat{\delta}^X(Z) &:= \hat{\delta}^T(Z)\,(1 - \hat{\mu}(Z)) + \hat{\delta}^C(Z)\,\hat{\mu}(Z) \\
\hat{\tau} &:= O_T\left(\left\{X_i, \hat{\delta}^X(Z_i)\right\}_{i=1}^n\right)
\end{aligned} \tag{15.1.9}$$

Assuming that the function spaces used in the nuisance oracles contain the true functions, the final step of the X-learner will converge to the best approximation of the CATE τ_*, within the space T. Moreover, this estimation strategy can have substantial benefits when the CATE function δ_0 is much simpler than the response functions g_0^C, g_0^T and when there are substantial imbalances in the treatment across the population (i.e. the propensities substantially deviate from $1/2$). The latter many times arises in digital experimentation, where only a small fraction of the population receives the treatment. In this case, the response under control can be much more accurately estimated. In fact, in many such settings we have a lot of historical data, prior to running an experiment, where the treatment was un-available and which can be used as auxiliary datasets for learning the baseline response; with the small treated data from the experiment only being used to estimate the heterogeneous effect function δ_0^T.

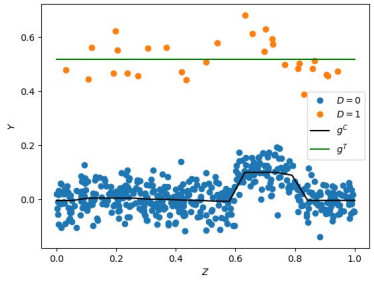


Figure 15.1: DGP 1. Imbalanced dataset were baseline response is more complex than heterogeneous effect.

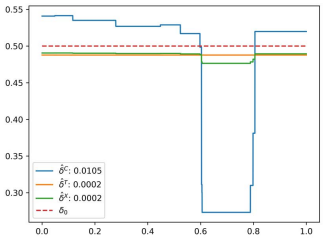


Figure 15.2: DGP 1. Different CATE estimates in the X-Learner. The legend displays the mean squared error of each estimate.

Example 15.1.1 (Imbalanced Dataset with Hard Baseline Response and Simple CATE) As a stark example, consider the case when $Z = X \sim U[0,1]$, treatment is very rare, i.e. $\mu(Z) = .05$, the treatment effect is constant, i.e. $\delta_0(Z) = .5$ and the baseline response is complex and contains a discontinuity:

$$\begin{aligned} Y &= .5\, D + .3\, \mathbb{1}\{Z \in [.6, .8]\} + N(0, \sigma = .05), \\ D &= \text{Bernoulli}(\mu(Z) = .05) \end{aligned} \tag{DGP 1}$$

In this case, the data that we collect, for $n = 500$, are depicted in Figure 15.7. If we use gradient boosted forest regression to estimate the two response functions under treatment and under control, we find that the $\hat{g}^T$ response function is substantially more regularized and the discontinuity is not learned, due to the small sample size. On the other hand $\hat{g}^C$ is much more accurate and the discontinuity is learned due to the large sample size. Subsequently, we see in Figure 15.2 that the estimate based on the CATC identification strategy is much less accurate than the one based on the CATT identification strategy. Moreover, the X-Learner is putting almost all the weight on the CATT estimate $\hat{\delta}^T$ and is highly accurate compared to $\hat{\delta}^C$. However, in this setting, we also find that other strategies that also use propensity modelling (e.g the R- or DR-Learners) also manage to correct the error in the T-Learner regression models and achieve similar accuracy to the X-Learner.

On the other hand, if the inductive bias that the CATE is simpler than the response functions under either treatment or control does not hold, then the superiority of the X-Learner strategy as compared for instance to the T-learner strategy for outcome modelling vanishes. For instance, if we instead have an outcome model of:

$$\begin{aligned} Y &= .5\, \mathbb{1}\{Z \in [.6, .8]\}\, D + .1 + N(0, \sigma = .05), \\ D &= \text{Bernoulli}(\mu(Z) = .05) \end{aligned} \tag{DGP 2}$$

then all methods that only rely on outcome modelling fail and methods that also combine propensity based identification start to outperform (see Figure 15.4). Even more vivid is the flip in performance if we further make the treatment more prevalent than the baseline:

$$\begin{aligned} Y &= .5\, \mathbb{1}\{Z \in [.6, .8]\}\, D + .1 + N(0, \sigma = .05), \\ D &= \text{Bernoulli}(\mu(Z) = .95) \end{aligned} \tag{DGP 3}$$

In this case it is more important to use the large amount of treated data, since $\mu(Z) = .95$, not to learn the response function, but rather to learn the CATE function (see Figure 15.5). In this case a T-Learner outcome modelling strategy and a T-Learner based DR-Learner is a better option.

We conclude by noting that the reasoning in the cross learner strategy can actually be used as a sub-process to improve outcome modelling in all other learners. In particular, note that the key advance of the cross learner is to observe than when the treatment is very rare, then we should be estimating the response $\hat{g}^C$ under control and then estimating only the effect $\hat{\delta}^T$ using the treatment data. In this case, we can also use $\hat{g}^T = \hat{g}^C + \hat{\delta}^T$ as our estimate of the response under treatment. Similarly, if the control group is very rare, then we should be estimating the response $\hat{g}^T$ under treatment and then estimating only the effect $\hat{\delta}^C$ using the control data. In this case, again we can also use $\hat{g}^C = \hat{g}^T - \hat{\delta}^C$ as our estimate of the response under control. Moreover, we can locally blend these two estimation strategies by weighting both estimates of the two response functions using the propensity, i.e. putting a weight of $(1 - \hat{\mu}(Z))$ to the first estimation strategy and a weight of $\hat{\mu}(Z)$ to the second estimation strategy. This approach is an alterantive outcome modelling process that can be used instead of the S or T learner approaches for learning the response functions under the different treatments. In that respect, the X-Learner outcome modelling strategy can be used in conjunction with the DR- or the R-Learner approaches, if one wants to introduce some robustness with respect to outcome modelling by incorprorating identification by propensity approaches. For instance, in Figure 15.3, we also depict the CATE learned if we combine the X-learner approach to outcome modelling with the doubly robust correction (coined the DRX-Learner).

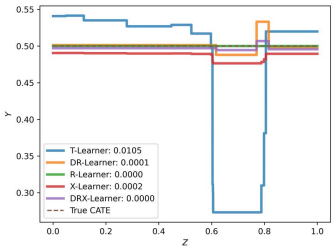


Figure 15.3: DGP 1. CATE estimates ($n = 500$) from other meta-learners. The legend displays the mean squared error of each learner.

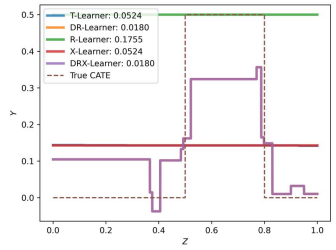


Figure 15.4: DGP 2. CATE estimates ($n = 500$) from meta-learners, when CATE is complex and baseline simple. The legend displays the mean squared error of each learner.

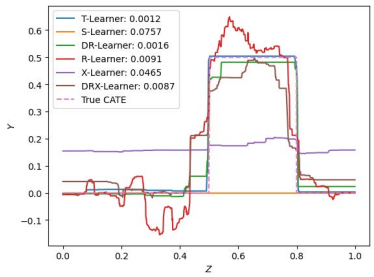


Figure 15.5: DGP 3. CATE estimates ($n = 500$) from meta-learners, when CATE is complex and baseline simple and treatment very prevalent. The legend displays the mean squared error of each learner.

Qualitative Comparison and Guidelines

We present here a set of bullet points that can guide the reader through the choosing among the different meta-learner strategies, dependent on inductive biases about their setting:

- S/T-Learner: they heavily rely on correct outcome modelling, trying to learn how the outcome relates to the control co-variates Z. If this estimation problem is hard to learn, then they will have poor performance, especially when the effect is a simple function and X is much lower dimensional than Z. However, they can have very low

variance and be more stable as they only depend on simple regression strategies. If one has to choose among the two, then the T-Learner should be preferable, if both treatments are sufficiently represented, since it avoids overly regularizing the treatment, which reduces the bias of the treatment effect estimate. If one of the two treatments is rare, then the X-Learner should be more preferable than the T-Learner.

- X-Learner: even if the X learner estimates a propensity, the propensity is primarily used to select locally, which outcome model is better and is not used to identify the effect. Thus the X learner is essentially also only performing outcome modelling. If we believe that the CATE function is simpler than the response functions under treatment or control, this outcome modelling estimation strategy should be preferred. Otherwise, if we believe that the CATE function is equally or more complex than the response functions, then a T-Learner approach can outperform an X-Learner approach. If we further believe that learning any of these outcome processes could potentially be a substantially harder task than learning the propensity, then this method can be heavily biased. In that case the DR-Learner or the R-Learner should be prefered. However, the X learner reasoning can still be useful in improving the outcome modelling part of the DR or R learners.
- DR-Learner: possesses doubly robust properties, in that the mean squared error of the CATE is small if either the outcome model is learned accurately or the propensity model. It is particularly useful in learning projections of the CATE on simpler function spaces or on small subsets X of the control variables Z. If X is very small compared to Z, then even the X-Learner needs to accurately learn the complex effect function $\delta_0(Z)$ accurately. However, the DR-Learner can learn the simpler CATE $\tau_0(X)$, if the propensity model is accurately learned. For instance, we saw in DGP 3 in Example 15.1.1, that when the CATE function was complex, then methods such as the DR-Learner that incorporate propensity modelling are more accurate. However, contrary to the S/T/X-learners, when the true data generating process has extreme propensities in parts of the covariate space (i.e. parts of the population are almost deterministically either treated or not treated), then the DR-Learner can have high variance and become unstable. On the other hand the R-Learner will be extrapolating the CATE from nearby regions where there

is more overlap and assuming the CATE model space is smooth enough, such model-based extrapolation can be quite accurate.

- R-Learner: posseses insensitivity properties related to Neyman orthogonality, in that the impact of errors in the propensity model or the outcome model impact the CATE model only in a second order manner. In particular, if the outcome model is wrong, but the propensity model is very accurate, the CATE will be highly accurate. However, it is more heavily relying on moderately accurate propensity modelling, unlike the DR-Learner. If for instance the outcome model is perfect, but the propensity model is very wrong, then the DR-Learner will be highly accurate but the R-Learner will not be. On the positive side, the R-Learner is much more stable than the DR-Learner in the presence of extreme propensities, as it does not divide by the propensity score when constructing the regression labels. The reason that it can bypass that is that it inherently estimates only an overlap-weighted projection of the CATE and not the true projection of the high-dimensional CATE δ_0, when the CATE model is either mis-specified or does not solely depend on X and not on the larger set of covariates Z. In both cases the DR-Learner converges to the true CATE, while the R-Learner can potentially ignore large parts of the population to reduce its variance; introducing bias and extrapolating the CATE from nearby highly overlapping regions. For instance, we show that in the case of DGP 1 in Example 15.1.1 the R-Learner was out-performing the DR-Learner, since the treatment effect was constant and the propensity very small. The R-Learner should be prefered to the DR-Learner when such overlap weighted projections are acceptable within the application context and when we believe we have a relatively accurate propensity model. In principle, a similar variance reduction can also be performed for the DR-Learner, by multiplying the DR-Learner loss with sample weights $W = \text{Var}(D \mid Z)^2$, which would then avoid dividing by the propensity and would converge to an overlap weighted projection of the CATE δ_0, with the aforementioned sample weights, while preserving the double robust nature of the estimation (i.e. that errors in propensity can be compensated by more accurate estimation in the outcome model and vice versa) (see e.g. [8]).

All-in-all, one should note that there is no clear winner among the X-, R- and DR-Learner methods and each can potentially be

the best performer in different contexts. The above discussion gives a high-level strategy of which method to use dependent on which types of phenomena one should be expecting to arise in their data. In the next section we give a more data-driven selection among these methods using out-of-sample scoring and ensembling.

Remark 15.1.1 (Guarding for Overfitting with Cross-fitting) To avoid having to worry about overfitted estimators, all the first stage nuisance models across all the meta-learners should preferably be estimated in a cross-fitting manner (i.e., the models $\hat{g}$, $\hat{g}^C$, $\hat{g}^T$, $\hat{\mu}$, $\hat{h}$) while the CATE models (i.e., $\hat{\tau}$, $\hat{\delta}^C$, $\hat{\delta}^T$) should be estimated using all the samples.

Remark 15.1.2 (Explainability and Interpretability) An important side benefit of the meta-learning approach to CATE estimation is that in the end we end up with an ML regression model that represents our estimate of the CATE function. Even though this regression model can be quite complicated (e.g. a random forest, a gradient boosted forest or a neural network), we can apply the multitude of interpretability approaches in machine learning to interpret the learned model. For instance, we can summarize how different features change the value of the CATE model via the widely used SHAP values [9] or approximate locally the CATE model with simpler linear models based on the widely used LIME framework. Finally, we can invoke *distillation methods* that fit simpler and easy to visualize models using the learned model predictions as labels. For instance, we can train a shallow binary tree regression model that approximates the CATE model predictions and then visualize the learned tree. For more elaborate treatment of interpretability methods in machine learning see [10].

Guarding for Covariate Shift

When machine learning models are evaluated on a different population of covariates than the one that they were trained on, then an important finite sample consideration is deterioration of their performance due to the covariate shift. Such population mis-match between training and evaluation typically arises when we employ ML algorithms within a CATE estimation. For instance, in the T-Learner we train an ML model on the treated datapoints and then we evaluate it on all the datapoints.

Similarly, in the X-Learner we train a CATE model on the treated points and then we evaluate it on all the datapoints.

In such settings, we should only expect the oracle ML model to have small mean squared error with respect to the distribution of its training data and with respect to the best approximation of the CEF, where the approximation error is calculated with respect to the training data. For instance, suppose that we estimate a regression model $\hat{h}$ that takes as input random variables X and predicts a variable Y, with sample weights W, by invoking an ML regression oracle as defined in Equation (15.1.2). Assuming that the CEF $h_* := \mathrm{E}(Y \mid X)$ does not change between train and evaluation data and letting D_t denote the distribution of X in the training data and D_e in the evaluation data, then our regression estimate satisfies that:

$$\mathrm{E}_{X \sim D_t} W(\hat{h}(X) - h_0(X))^2 \leq r_n$$

where $h_0 = \arg\min_{h \in H} \mathrm{E}_{X \sim D_t} W(h_*(X) - h(X))^2$. Since we evaluate this regression model on a different population, we would typically care about the following mean squared error:

$$\mathrm{E}_{X \sim D_e} W_e(\hat{h}(X) - h_*(X))^2$$

with some set of weights W_e that depend on some downstream use of the model.

Example 15.1.2 (Covariate Shift in X-Learner) In the context of the X-Learner, we train a model $\hat{\delta}^T$ on the treated data and then we use it to calculate $\hat{\delta}(Z) = \hat{\delta}^T(Z)\,(1 - \mu(Z)) + \hat{\delta}^C(Z)\,\mu(Z)$ on all the data points. Thus in this case, when measuring the quality of the downstream CATE estimate $\hat{\tau}$ in the final step of the X-Learner, we care about the quality of $\hat{\delta}^T$ as measured by the metric:

$$\mathrm{E}_Z (1 - \mu(Z))^2 (\hat{\delta}^T(Z) - \delta_0(Z))^2$$

On the contrary, the oracle for $\hat{\delta}^T$ would be guaranteeing:

$$\mathrm{E}_{Z|D=1} (\hat{\delta}^T(Z) - \tilde{\delta}_0(Z))^2$$

where $\tilde{\delta}_0 = \arg\min_{\delta \in \Delta} \mathrm{E}_{Z|D=1} (\delta_0(Z) - \delta(Z))^2$.

There are two sources of discrepancy: *first* the approximation error can be substantially different if we use the best approximation with respect to a different distribution and *second* the mean squared error is measured with respect to the wrong distribution. If the true CEF h_0 lies in the function space H, then

the first problem vanishes (though in finite samples and with some growing sieve space, we should always expect some finite sample approximation bias). Simiarly, if we denote with p_t the density of X under D_t and p_e under D_e, then if the density ratio $p_e(X)/p_t(X)$ is upper and lower bounded by some constants $[c, C]$, then we always have that:

$$\begin{aligned}\mathrm{E}_{X\sim D_e} W_e(h(X) - h_0(X))^2 &= \mathrm{E}_{X\sim D_t} \frac{p_e(X)}{p_t(X)} W_e(h(X) - h_0(X))^2 \\ &\in [c, C] \cdot \mathrm{E}_{X\sim D_t} W_e(h(X) - h_0(X))^2\end{aligned}$$

Thus even if we don't take any measures to address the co-variate shift, by minimizing the squared error under the training distribution, we are approximately minimizing the error under the evaluation distribution. However, these constants can be quite large in practice and the magnitude of the discrepancy can be comparable to the sample size.

For these reasons a large literature in machine learning has focused on addressing such co-variate shift problems by changing how we train the model, when we know what the target evaluation distribution or metric will be. In its simplest form, one can instead optimize for the density ratio weighted error, i.e.:

$$\mathrm{E}_{X\sim D_t} \frac{p_e(X)}{p_t(X)} W_e(h(X) - h_0(X))^2$$

Noting also that $\frac{p_e(X)}{p_t(X)} = \frac{p(X|e)}{p(X|t)} = \frac{p(e|X)p(t)}{p(t|X)p(t)}$, the above is equivalent to minimizing:

$$\mathrm{E}_{X\sim D_t} \frac{p(e \mid X)}{p(t \mid X)} W_e(h(X) - h_0(X))^2$$

which requires solving two classification problems (i.e. predicting the probability that a sample is in population e given X and predicting whether the sample is in population t given X, using the union of the populations).

Example 15.1.3 (Covariate Shift in X-Learner (continued)) Going back to our X-Learner example, we have $p(e \mid Z) = 1$ (since we evaluate on all the population) and $p(t \mid Z) = \mu_0(Z)$ (since we train only on the training population). Moreover, we care about evaluation weights $W_e = (1 - \hat{\mu}(Z))^2$. Thus it

would potentially be better in finite samples if one optimizes:

$$\mathrm{E}_{Z|D=1}\frac{1}{\hat{\mu}(Z)}(1-\hat{\mu}(Z))^2(\hat{\delta}^T(Z)-\tilde{\delta}_0(Z))^2$$

In other words, calling the ML oracle when training $\hat{\delta}^T$ with sample weights $W = \frac{1}{\hat{\mu}(Z)}(1-\hat{\mu}(Z))^2$.

For instance, if we employ such co-variate shift techniques in DGP 3 from Example 15.1.1, then we find that the performance of the Domain Adapted X-Learner (DAX-Learner) is restored (see Figure 15.6).

Analogous finite sample corrections can be taken throughout the meta-learner algorithms by first working out what is the target evaluation population and metric we care about and changing the training of the ML model appropriately.

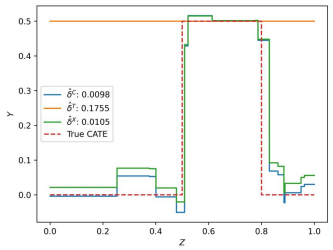


Figure 15.6: DGP 3. CATE estimates ($n = 500$) from meta-learners in adapted X-Learner, with co-variate shift corrections.

Covariate shift techniques when overlap fails. Beyond this simple approach of density weighting, many other ML methods have been developed in the literature to guard against covariate shift. One advantage of many of these alternative methods, is that they are applicable even when there is lack of overlap (i.e. when the density ratio can be unbounded or zero). For instance, one large class of covariate shift approaches within the context of neural network training, makes the assumption that overlap holds on some latent representation space $\phi(X)$ and not on the observed covariate space X and that the conditional expectation function can be written as a function of these latent variables, i.e. $\mathrm{E}(Y \mid X) \approx \mathrm{E}(Y \mid \phi(X))$. In this case, one can train a neural network architecture where the first few layers of the neural network are used to construct the mapping $\phi(X)$ and the subsequent layers are used to construct $\mathrm{E}(Y \mid \phi(X))$. Subsequently a distribution distance measure is introduced as a regularizer, that measures the distribution distance of $\phi(X)$ between samples that stem from the training and evaluation population. A popular metric is a variant of the Wasserstein distance. In this manner, we are trying to construct a latent representation that has approximately the same distribution under the two populations and which predicts well the target Y.

Shared representation learning with neural networks. In the context of CATE estimation, the latter approach was utilized by [11, 12] within the T-Learner framework for outcome modelling. In particular, the first few layers of the network are used to represent $\phi(Z)$, which then is used to represent both g_0^T and

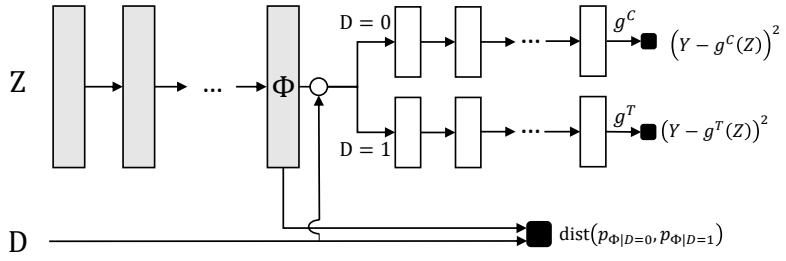


Figure 15.7: Counterfactual regret network of [11, 12], to guard against covariate shift in the T-Learner.

g_0^C. Moreover, a wasserstein penalty is introduced so that the distribution of $\phi(Z)$ is similar between the treated and control population. The resulting method is typically referred to as the CFR-Net. If one believes that their setting satisfies this inductive bias, i.e. that there exists a latent representation that is sufficient for the CEF of the outcome and in which overlap holds, then this approach can be used for better outcome modelling within the context of any meta-learner. For instance, one can use the CFR-Net together with the DR- or R- learners for estimating g_0^T, g_0^C and hence also g_0 and $\hat{h}$ (potentially using the same shared representation, when estimating the propensity; to avoid extreme propensities). See also [13] for the empirical evaluation of variants of such neural network approaches, combined with doubly robust learning.

Example 15.1.4 (Meta-Learners in the 401(k) Example) We applied each of the meta-learner models to estimate the CATE in the 401(k) example. We estimated a CATE model that uses all the available variables for heterogeneity (i.e. $X = Z$) and used gradient boosted forests (based on the xgboost library) as oracle regression models for each step of each meta-learner. We depict below the CATE predictions of each of the meta-learner models as a function of income (x-axis), when all other features are fixed to their overall median value.

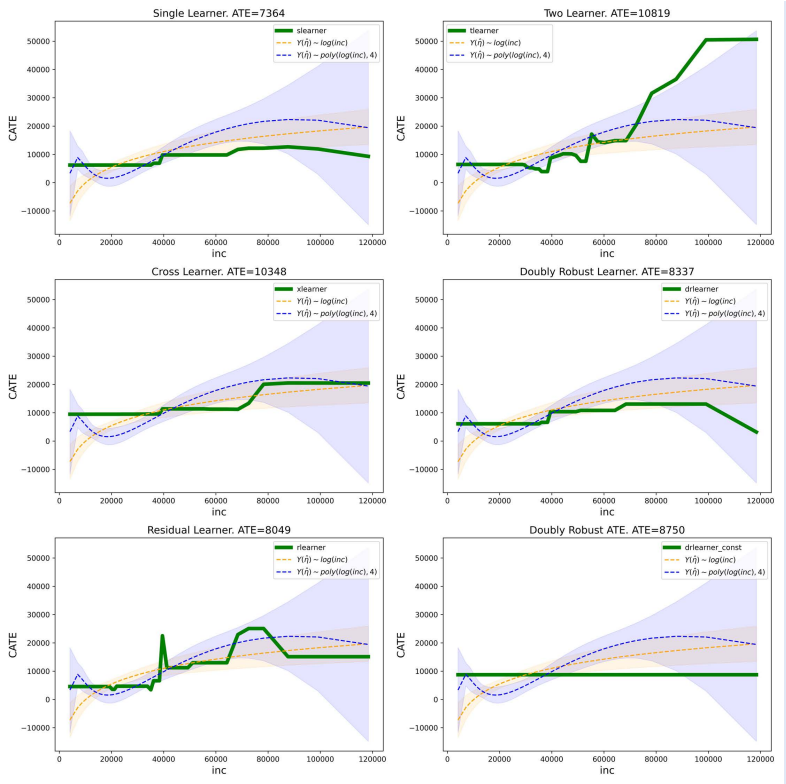
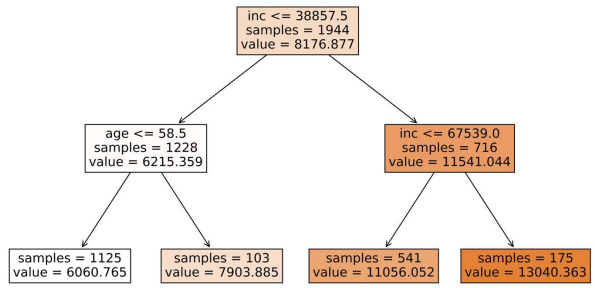
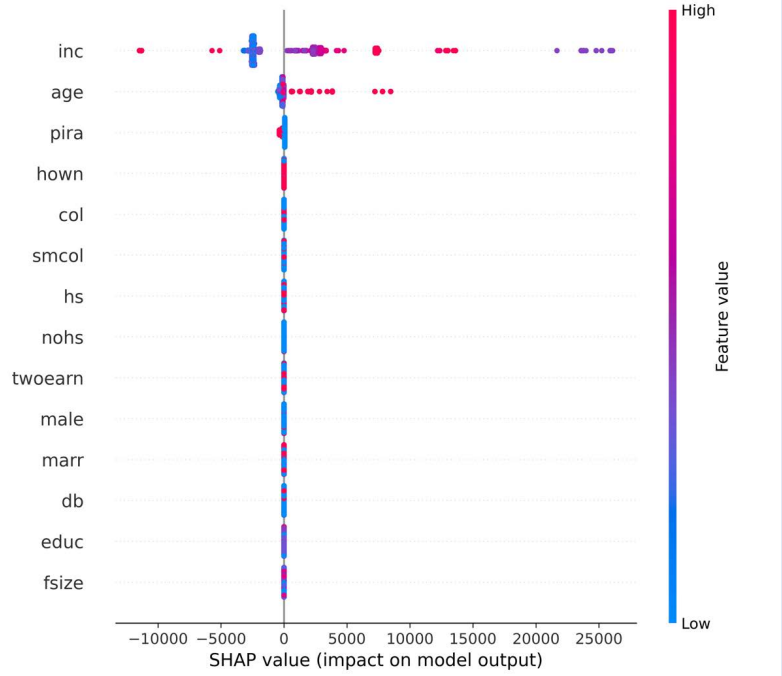


Figure 15.8: CATE predictions of different meta-learners in the 401k example. Gradient boosted forests (via the xgboost library) were used as ML oracles for regression and classification. The CATE is predicted on a grid of income points, corresponding to equally spaced income quantiles. All other covariates were imputed at their median values. For comparison, each plot also displays the doubly robust best linear predictor of the CATE with 5-95% confidence intervals on a simple linear form of engineered features of the income.

Subsequently, we investigate for interpretability reasons, the main factors that are driving the predictions of the DR-Learner model. We do this by fitting a simple shallow binary regression tree on the predictions of the model. We find that the model's CATE predictions are primarily driven by income and age factors. In particular, the model finds that 401(k) eligibility has the lowest effect ($\approx \$6k$) in net financial assets for low income ($\lesssim \$39k$) and younger people ($< 59$ years), while it has the highest effect ($\approx \$13k$) for high income people ($\gtrsim \$68k$).



We also depict the SHAP values for each feature in the CATE model, that identifies the directionality and magnitude of the change that each feature drives in the model's output. We again identify that the main factors that drive variation in the output of the DR-Learner CATE model are income and age.



We find that even though we did not hard-code income or age as factors of effect heterogeneity, the generic ML approach identified these two factors as the key drivers; a conclusion that is inline with domain knowledge.

15.2 Scoring for CATE Model Selection and Ensembling

The previous section gave an overview of how to qualitatively select among the different meta-learning strategies. Here we discuss how one can automate the process of selection using out-of-sample scoring and moreover how to potentially ensemble the models that come out of different estimation strategies into a single CATE model. In this section, we envision that the user has split their data into a training and scoring set and based on the training set they have fitted a set of candidate CATE models $T := \{\tau_1, \ldots, \tau_M\}$. These models could correspond to the result of the different meta-learning strategies and with different regression style oracles. For instance, τ_1 could be the result of an X-Learner with random forest regression oracles,

τ_2 the result of an X-Learner with gradient boosted regression oracles, τ_3 the result of the DR-Learner with linear/logistic model oracles.

Let n denote the size of the scoring set. Our goal is to be able to use the scoring set in order to evaluate which of these M models is more accurate (with confidence) and to device approaches to select a single model τ_* that could either correspond to one of the models T or to an ensemble of these models with weights $(w_1, \ldots, w_M)$, such that $\tau_*(X) = \sum_{i=1}^{M} w_i \tau_i(X)$. Such a model τ_* should ideally be competing with the best model in T, i.e., with high probability:

$$\mathrm{E}(\tau_*(X) - \tau_0(X))^2 \leq \min_{j=1}^{M} \mathrm{E}(\tau_j(X) - \tau_0(X))^2 + \epsilon(n, M) \tag{15.2.1}$$

for some error function $\epsilon(n, M)$ that should decay fast to zero as a function of n and should grow slowly with the number of candidate models M. We can use again the doubly robust outcome approach, viewing the problem as a regression problem with the doubly robust proxy outcomes $Y_i(\hat{\eta})$ as the labels and utilize techniques from model scoring, ensembling, model selection and stacking for regression problems.

Comparing Models with Confidence

We can use the doubly robust loss:

$$\hat{L}_{DR}(\tau; \hat{\eta}) := \mathbb{E}_n(Y(\hat{\eta}) - \tau(X))^2 \tag{15.2.2}$$

as a quality score for each of the candidate models. Since we care about selecting among the models in T, we primarily care about choosing a score function that orders the models accurately. Hence, we primarily care that *differences* in the score between two models, i.e.:

$$\hat{\delta}_{i,j}(\hat{\eta}) = \hat{L}_{DR}(\tau_i; \hat{\eta}) - \hat{L}_{DR}(\tau_j; \hat{\eta}), \tag{15.2.3}$$

approximate well *differences* in mean squared error, i.e.:

$$\delta^*_{i,j} = \mathrm{E}(\tau_i(X) - \tau_0(X))^2 - \mathrm{E}(\tau_j(X) - \tau_0(X))^2 \tag{15.2.4}$$

Consider the population analogues of the score and differences in the score, i.e.:

$$L_{DR}(\tau;\eta) := \mathrm{E}(Y(\eta) - \tau(X))^2 \tag{15.2.5}$$
$$\delta_{i,j}(\eta) := L_{DR}(\tau_i;\eta) - L_{DR}(\tau_j;\eta) \tag{15.2.6}$$

Since $\mathrm{E}[Y(\eta_0) \mid X] = \tau_0(X)$, we have that:[5]

5: Prove this as an exercise.

$$\delta_{i,j}(\eta_0) = \delta^*_{i,j} \tag{15.2.7}$$

Moreover, note that the population difference at the estimate $\hat{\eta}$ satisfies (by simply expanding the squares):

$$\begin{aligned}\delta_{i,j}(\hat{\eta}) &= \mathrm{E}[\tau_i(X)^2 - \tau_j(X)^2 - 2\,Y(\hat{\eta})\,(\tau_i(X) - \tau_j(X))] \\ &= \mathrm{E}[\tau_i(X)^2 - \tau_j(X)^2 - 2\,\mathrm{E}[Y(\hat{\eta}) \mid X]\,(\tau_i(X) - \tau_j(X))]\end{aligned}$$

Thus the error in the model comparison due to the error in the estimate $\hat{\eta}$ is:

$$\delta_{i,j}(\hat{\eta}) - \delta_{i,j}(\eta_0) = 2\,\mathrm{E}[\mathrm{E}[Y(\eta_0) - Y(\hat{\eta}) \mid X]\,(\tau_i(X) - \tau_j(X))]$$

We see that a good scoring rule, should be using proxy labels that have small bias, i.e.:

$$\mathrm{bias}(X;\hat{\eta}) := \mathrm{E}[Y(\eta_0) - Y(\hat{\eta}) \mid X] \tag{15.2.8}$$

The doubly robust proxy labels exactly achieve this property. In particular, we can show based on results in prior sections:[6]

6: Prove this as an exercise.

$$\mathrm{bias}(X;\hat{\eta}) = (H(\mu_0) - H(\hat{\mu}))\,(g_0(D,Z) - \hat{g}(D,Z)) \tag{15.2.9}$$

Thus we derived that the error in the comparison between model τ_i and model τ_j, due to the estimation error in $\hat{\eta}$ is:

$$2\,\mathrm{E}[(H(\mu_0) - H(\hat{\mu}))\,(g_0(D,Z) - \hat{g}(D,Z))\,(\tau_i(X) - \tau_j(X))]$$

This has doubly robust properties, i.e. if either $H(\hat{\mu})$ is accurate or $\hat{g}$ is accurate, then the comparison between the two models will be accurate. Moreover, the difference in scores also satisfies the Neyman orthogonality property:[7]

7: Prove this as an exercise.

$$\partial_\eta \delta_{i,j}(\eta_0) = 0 \tag{15.2.10}$$

and since $\hat{\delta}_{i,j}(\eta)$ is the empirical analogue of $\delta_{i,j}(\eta)$, we can apply the general framework of Neyman orthogonality to deduce that the score difference estimate $\hat{\delta}_{i,j}(\hat{\eta})$ is root-n asymptotically normal:[8]

8: A similar theorem also holds for the case of cross-fitted estimates. In practice, one can either use the nuisance estimates that were constructed on the training set, which was also used to construct the functions $\{\tau_1, \ldots, \tau_M\}$ or perform cross-fitting within the scoring set.

Theorem 15.2.1 *Let $v_{ij}(X) = \tau_i(X) - \tau_j(X)$ and suppose that $\mathbb{E}_n v_{ij}(X)^2 \geq c$ for some constant $c > 0$ and let n grow to infinity. As long as $\hat{\mu}$ and $\hat{g}$ are estimated on a separate sample and satisfy that:*

$$\sqrt{n}\mathrm{E}[(H(\mu_0) - H(\hat{\mu}))\,(g_0(D,Z) - \hat{g}(D,Z))\,v_{ij}(X)] \approx 0$$

and both nuisance functions are consistent, i.e.:

$$\|\hat{\mu} - \mu_0\|_{L^2} + \|\hat{g} - g_0\|_{L^2} \approx 0 \tag{15.2.11}$$

Then the estimation error in the nuisance functions $\hat{\mu}, \hat{\eta}$, does not have a first order effect in the estimation error of the score difference between two models:

$$\sqrt{n}(\hat{\delta}_{i,j}(\hat{\eta}) - \delta^*_{i,j}) \approx \sqrt{n}\mathbb{E}_n(Y(\eta_0) - \tau_i(X))^2 - (Y(\eta_0) - \tau_j(X))^2$$

Consequently, the estimate $\hat{\delta}_{i,j}$ concentrates in a $1/\sqrt{n}$ neigborhood of $\delta^*_{i,j}$ with deviations controlled by the Gaussian law:

$$\sqrt{n}(\hat{\delta}_{i,j}(\hat{\eta}) - \delta^*_{i,j}) \overset{a}{\sim} N(0, \mathsf{V}) \tag{15.2.12}$$

where:

$$\mathsf{V} := \mathrm{E}\left((Y(\eta_0) - \tau_i(X))^2 - (Y(\eta_0) - \tau_j(X))^2 - \delta^*_{i,j}\right)^2$$

Moreover, confidence intervals on the performance difference between two models can be constructed as:

$$\mathrm{P}\left(\delta^*_{i,j} \in \left[\hat{\delta}_{i,j}(\hat{\eta}) \pm c\sqrt{\hat{\mathsf{V}}/n}\right]\right) \approx 1 - \alpha \tag{15.2.13}$$

where c is the $(1-\alpha/2)$-quantile of the standard normal distribution and

$$\hat{\mathsf{V}} := \mathbb{E}_n\left((Y(\hat{\eta}) - \tau_i(X))^2 - (Y(\hat{\eta}) - \tau_j(X))^2 - \hat{\delta}_{i,j}(\hat{\eta})\right)^2$$

The above theorem can also directly be used to construct

Remark 15.2.1 (Sample-dependent base models) The assumption that $\mathbb{E}_n v_{ij}(X)^2$ is some non-zero constant c independent of n required so that the variance V is non-zero. In practice, the candidate models will also be changing with n as we will be growing the sample size of the training set together with the scoring set. As the size of the training set converges to infinity it is highly probable that c will be converging to zero,

in which case V will also be converging to zero. The above theorem allows one to compare models that are distinct in their average predictions by at least some constant. If we want to be comparing models whose distinctness (i.e. $\mathrm{E}v_{ij}(X)^2$) shrinks with the sample size, then we need to be more careful in the asymptotic normal approximation. In this case, it is more appropriate to consider the asymptotic properties of the self-normalized quanity:

$$\sqrt{\frac{n}{V_n}}(\hat{\delta}_{i,j}(\hat{\eta}) - \delta^*_{i,j}),$$

where V_n is now allowed to depend on n, since τ_i, τ_j are allowed to depend on n. In this case, to ignore the error due to $\hat{\eta}$ we would need that:

$$\sqrt{\frac{n}{V_n}}\mathrm{E}[(H(\mu_0) - H(\hat{\mu}))\,(g_0(D,Z) - \hat{g}(D,Z))\,v_{ij}(X)] \approx 0$$

As we show in Appendix 15.A:

$$V_n \geq 4\mathrm{E}v_{ij}(X)^2\,\mathrm{Var}(Y(\eta_0) \mid X)$$

Thus if we assume that $\mathrm{Var}(Y(\eta_0) \mid X) \geq c > 0$, then $V_n \geq 4c\|v_{ij}\|^2_{L^2}$ and it suffices that:

$$\sqrt{n}\mathrm{E}\left[(H(\mu_0) - H(\hat{\mu}))\,(g_0(D,Z) - \hat{g}(D,Z))\,\frac{v_{ij}(X)}{\|v_{ij}\|_{L^2}}\right] \approx 0$$

If $\left|v_{ij}(X)\right| \leq C\|v_{ij}\|_{L^2}$ almost surely, then the above would hold under the standard condition that:

$$\sqrt{n}\|H(\mu_0) - H(\hat{\mu})\|_{L^2}\,\|g_0 - \hat{g}\|_{L^2} \approx 0$$

Even when this condition does not hold, by an application of the Cauchy-Schwarz inequality it suffices that:

$$\sqrt{n}\sqrt{\mathrm{E}\left[(H(\mu_0) - H(\hat{\mu}))^2\,(g_0(D,Z) - \hat{g}(D,Z))^2\right]} \approx 0$$

which would hold if:

$$\sqrt{n}\|H(\mu_0) - H(\hat{\mu})\|_{L^4}\,\|g_0 - \hat{g}\|_{L^4} \approx 0 \tag{15.2.14}$$

Moreover, for the confidence interval to be valid, we would also need that:

$$\frac{|\mathsf{V} - \hat{\mathsf{V}}|}{\hat{\mathsf{V}}} \approx 0 \tag{15.2.15}$$

If the denominator is lower bounded by a constant, the above holds under benign regularity conditions. However, if we consider models whose separation shrinks at some rate σ_n^2, we can enforce the same rate of shrinkage on our estimate, of the variance, in which case we would need the estimation error in the variance to shrink faster than σ_n^2. Thus we can only consider comparison of models that are separated by at least some amount that dominates the error we expect in our variance estimate. This separation would always be of larger order than $1/n$. However, how small we can take this rate also depends on rates of convergence of our nuisance estimates individually and not just their product.

Remark 15.2.2 (Normalized Interpretable DR-Score) The loss $\hat{L}_{DR}(\tau; \hat{\eta})$ might not be very interpretable in practice as the result depends on the unobserved heterogeneity of the outcome and on the units of the outcome. As a more interpretable performance metric we can consider comparing the loss of any candidate model as compared to the loss of the best constant effect model fitted on the training sample. Let $\hat{\tau}_c$ denote the constant effect model that always outputs the estimate $\hat{\theta}$ of the average treatment effect, estimated on the training data. Then we can define the normalized score:

$$\hat{S}(\tau; \hat{\eta}) = \frac{\hat{L}_{DR}(\hat{\tau}_c; \hat{\eta}) - \hat{L}_{DR}(\tau; \hat{\eta})}{\hat{L}_{DR}(\hat{\tau}_c; \hat{\eta})} \tag{15.2.16}$$

This can be interpreted as a relative improvement in performance over a constant model and is a number in $[-\infty, 1]$. A larger score hints at a better CATE model. Moreover, for any reasonable model model this score will be a non-negative number in $[0, 1]$.

Competing with the Best Model

The doubly robust loss can also be used for constructin an ensemble τ_* that competes with the best model in T. The simplest approach would be to choose the model with the best score, i.e.:

$$\tau_* = \arg\min_{\tau \in T} \hat{L}_{DR}(\tau; \hat{\eta}) \tag{15.2.17}$$

Such a model satisfies the oracle performance guarantee in Equation (15.2.1) with (see e.g. [4])

$$\epsilon(n, M) \lesssim \sqrt{\frac{\log(M)}{n}} + \|H(\mu_0) - H(\hat{\mu})\|_{L^2} \|g_0 - \hat{g}\|_{L^2}$$

The leading term in this result is unfortunate, since it does not decay fast with the sample size n, i.e. as $1/n$. For instance, for parametric base models, we would expect the base models to have RMSE performance of $\lesssim 1/n$, in which case the above $1/\sqrt{n}$ rate becomes a dominant term.

One problem with this approach is the non-convexity of the space of models over which we are optimizing (i.e. optimizing over singleton models). This non-convexity can be alleviate by stacking approaches that convexify the optimization space over which we optimize and minimize the doubly robust loss over linear combinations of the base cate models, i.e.:

$$\tau_* := \sum_{i=1}^{M} w_i^* \tau_i, \quad w^* := \arg\min_{w \in W} \hat{L}_{DR}\left(\sum_{i=1}^{M} w_i \tau_i; \hat{\eta}\right) \tag{15.2.18}$$

where W could either be $\mathbb{R}^M$, in which case this is simply OLS regression with covariates $\tau_1(X), \ldots, \tau_M(X)$ and target outcome $Y(\hat{\eta})$,[9] or W could be the M-dimensional simplex, i.e.

$$W := \left\{ w \in \mathbb{R}^M : w_i \geq 0, \sum_{i=1}^{M} w_i = 1 \right\},$$

in which case this corresponds to a convex regression with the same covariates and outcome as in the OLS case. In the absence of further assumptions on the quality of the base models τ_i, the above yield a model τ_* that satisfies the oracle performance guarantee in Equation (15.2.1) with (see e.g. [4, 14])

$$\epsilon(n, M) \lesssim \min \left\{ \frac{M}{n} + \|H(\mu_0) - H(\hat{\mu})\|_{L^4} \|g_0 - \hat{g}\|_{L^4}, \right.$$
$$\left. \sqrt{\frac{\log(M)}{n}} + \|H(\mu_0) - H(\hat{\mu})\|_{L^2} \|g_0 - \hat{g}\|_{L^2} \right\}$$

The above approach yields a fast rate guarantee with respect to the sample size, but suffers from a large set of base models M. The reason being that the convexification of the optimization space introduced M parameters that correspond to the weights for each model and no penalty to encourage sparsity of the solution.

One can achieve the ideal leading rate of $\log(M)/n$, that is both fast with respect to the sample size n and grows only logarith-

9: This is mathematically equivalent to the BLP approach we described in the first section of this chapter, albeit using the predictions of the base CATE models as the engineered features.

mically with the number of base models M, by a penalized stacking approach called Q-aggregation [15], which penalizes different models based on their individual performance:

$$w = \arg\min_{w \in W} \hat{L}_{DR}\left(\sum_{i=1}^{M} w_i \tau_i; \hat{\eta}\right) + \sum_{i=1}^{M} w_i \hat{L}_{DR}(\tau_i; \hat{\eta}) \quad (15.2.19)$$

where W is the M-dimensional simplex. This is an M-dimensional convex optimization program that can be solved very fast with modern convex optimization solvers. The resulting ensemble model competes with the best model at the statistically optimal leading rate of (see [16]):

$$\epsilon(n, M) \lesssim \frac{\log(M)}{n} + \|H(\mu_0) - H(\hat{\mu})\|_{L^4} \, \|g_0 - \hat{g}\|_{L^4}$$

Remark 15.2.3 (ATE and Intercept of Stacked Model) In practice, one might also include an intercept in the stacking model, i.e.

$$w = \arg\min_{w \in W, c \in \mathbb{R}} \hat{L}_{DR}\left(c + \sum_{i=1}^{M} w_i \tilde{\tau}_i; \hat{\eta}\right) \quad (15.2.20)$$

where $\tilde{\tau}_i$ are de-meaned versions of the CATE models, i.e. $\tilde{\tau}_i(X) = \tau_i(X) - \mathbb{E}_n \tau_i(X)$. In this case, the constant c can be interpreted as the final estimate of the Average Treatment Effect. Given that typically the scoring dataset will be smaller than the training dataset, it might be more advisable to use an estimate of the ATE that comes from the trainin dataset. For instance, we can use as c the doubly robust estimate of the ATE from the training dataset, denoted as $\hat{\theta}_{DR}^{\text{train}}$ and not optimize over it in the scoring dataset, i.e.

$$w = \arg\min_{w \in W} \hat{L}_{DR}\left(\hat{\theta}_{DR}^{\text{train}} + \sum_{i=1}^{M} w_i \tilde{\tau}_i; \hat{\eta}\right) \quad (15.2.21)$$

Remark 15.2.4 (Generic Stacking for CATE) Another approach that is typically employed in regression stacking is using more flexible stacking regressors. In the case of stacking for the CATE we can treat the ensemble problem as yet another regression problem of predicting $Y(\hat{\eta})$ from the covariates

$$\tau \circ X := (\tau_1(X), \ldots, \tau_M(X)),$$

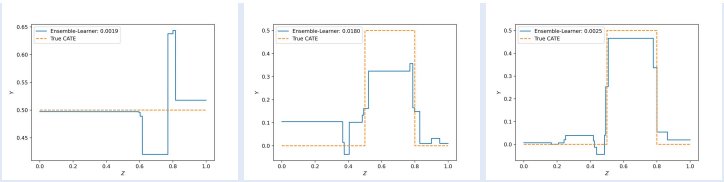
using the scoring dataset. Thus we can call a generic ML regression oracle to get τ_*:

$$\tau_* := O_{T^*}\left(\{\tau \circ X_i, Y_i(\hat{\eta})\}_{i=1}^n\right)$$

However, in this case one should worry about variance and overfitting, as typically the scoring dataset will be of smaller size than the training dataset. Thus, very flexible models can deteriorate the performance. The benefits of more general stacking regressors are not clear if one wants to solely compete with the best base model. However, more general regressors can potentially find models that perform better than the best base model. In practice, the most commonly used oracle models are penalized linear models, such as Lasso or Ridge regression, potentially with postivity constraints on the coefficients.

Remark 15.2.5 (Stability) The final CATE ensemble model that we selected based on the aforementioned process (i.e. training generic meta-learner models on a training set and scoring and stacking on a test set) unfortunately does not come with confidence intervals. Even though this is a process that can lead to a model with small mean squared error, the exact model can be quite sensitive to small variants of the data analysis pipeline (e.g. the randomness in the train/score split, the randomness in the estimators, the removal of a few samples). Even though we cannot construct valid confidence intervals for the predictions of the model or the findings in its structure (i.e. which are the important features), it is still advisable to perform some form of sensitivity or stability check of the model to such variants. For instance, one can run the same pipeline with different random seeds or remove random small fractions of the data and see how stable the different aspects of the model are. In the next section, we will also discuss more formal statistical tests that one can perform on a separate validation set (i.e. if one splits their data into train/scoring/validation sets), which validate aspects of the chosen CATE model.

Example 15.2.1 (Model Selection in Simple DGPs) We revisit the three data generating processes from Example 15.1.1. We depict below the performance of the Q-aggregation ensemble in a random sample of each of the data generating processes ($n = 500$).



(a) DGP 1 **(b)** DGP 1 **(c)** DGP 1

Moreover, in Table 15.10, we depict the average performance of each meta-learning method and of three variants of the ensemble methods (based on Q-aggregation, convex regression and simply choosing the single best score model) in terms of CATE RMSE over 100 experiments. We find that even though different learners are optimal in each of the DGPs, the ensemble learners are consistently close to the best performer across the board, while each of the other learners fails by a large margin in at least one DGP.

	DGP 1	DGP 2	DGP 3
DR	[0.016 ± 0.012] 0.015 (0.036)	[0.193 ± 0.036] 0.195 (0.243)	[0.049 ± 0.011] 0.047 (0.070)
DRX	[0.012 ± 0.009] 0.010 (0.030)	[0.193 ± 0.037] 0.195 (0.243)	[0.178 ± 0.034] 0.181 (0.230)
R	[0.002 ± 0.007] 0.000 (0.020)	[0.374 ± 0.080] 0.418 (0.421)	[0.374 ± 0.079] 0.418 (0.421)
T	[0.038 ± 0.005] 0.037 (0.047)	[0.235 ± 0.007] 0.232 (0.245)	[0.036 ± 0.011] 0.035 (0.056)
X	[0.010 ± 0.008] 0.008 (0.028)	[0.235 ± 0.007] 0.232 (0.245)	[0.223 ± 0.006] 0.221 (0.235)
DAX	[0.013 ± 0.008] 0.010 (0.030)	[0.156 ± 0.053] 0.151 (0.243)	[0.155 ± 0.045] 0.149 (0.232)
Q	[0.017 ± 0.014] 0.014 (0.038)	[0.165 ± 0.049] 0.161 (0.243)	[0.037 ± 0.012] 0.035 (0.056)
Convex	[0.019 ± 0.013] 0.018 (0.037)	[0.163 ± 0.042] 0.164 (0.236)	[0.038 ± 0.011] 0.037 (0.056)
Best	[0.017 ± 0.015] 0.011 (0.042)	[0.171 ± 0.055] 0.164 (0.269)	[0.037 ± 0.013] 0.036 (0.061)

Figure 15.10: CATE RMSE performance of each of the meta-learning and ensemble methods in the three simple DGPs across 100 experiments. Each cell displays [mean ± standard deviation], median (95%) of the RMSE across the 100 experiments.

Example 15.2.2 (Model Selection in the 401(k) Example) We also revisit the 401(k) example from the perspective of model comparison and ensembling. We first investigate the comparison of each of the performance of each of the meta-learning models compared to the constant model in Figure 15.11. We find that we cannot statistically conclude that the RMSE performance of any of these models is better than the constant effect model.

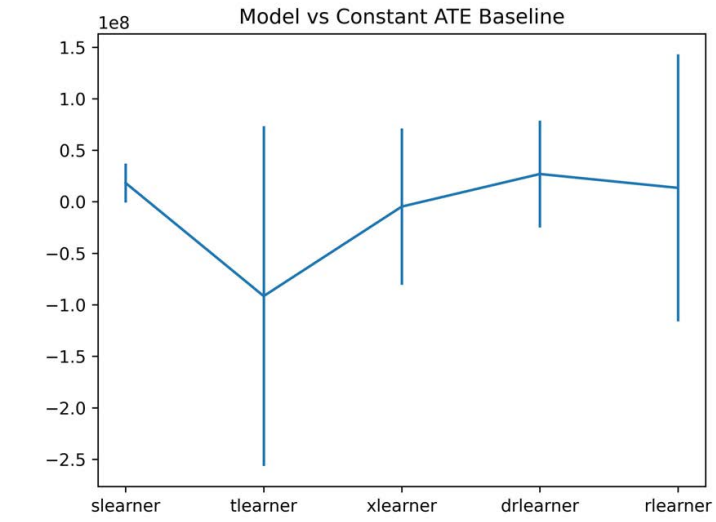


Figure 15.11: Score difference and confidence interval for each of the meta-learner models as compared to a constant treatment effect baseline. We find that among all models, only for the s-learner we can barely find that it has better CATE evaluation accuracy as compared to a constant effect model with statistical significance.

Subsequently, we investigate the ensemble models that are selected by each stacking method. We find that very flexible methods such as OLS, or Ridge can be quite un-stable, while methods that either constrain the weights to be positive or lie in the simplex, or induce sparse ensembles are qualitatively more reasonable.

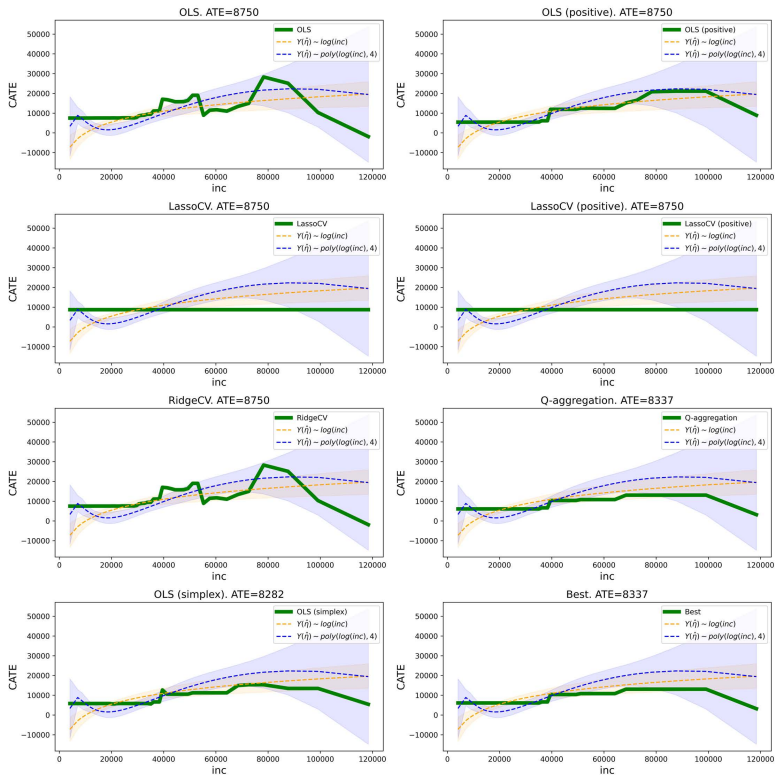


Figure 15.12: CATE predictions of different stacked ensemble models in the 401k example. Gradient boosted forests (via the xgboost library) were used as ML oracles for regression and classification. The CATE is predicted on a grid of income points, corresponding to equally spaced income quantiles. All other covariates were imputed at their median values. For comparison, each plot also displays the doubly robust best linear predictor of the CATE with 5-95% confidence intervals on a simple linear form of engineered features of the income.

Subsequently, we investigate for interpretability reasons, the main factors that are driving the predictions of the ensemble model chosen by Q-aggregation. We do this by fitting a simple shallow binary regression tree on the predictions of the model. Given that the ensemble chooses to put weight primarily on the DR-Learner model, the insights are similar to those in Example 15.1.4.

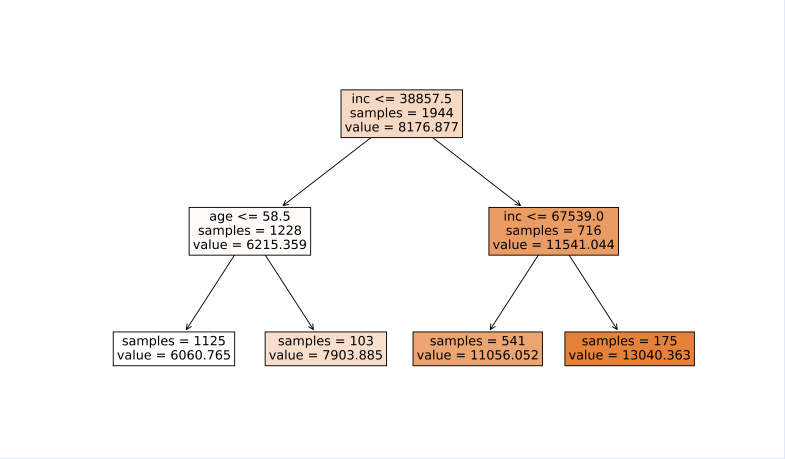


Figure 15.13: Single binary regression tree distillation of the Q-aggregation based stacked ensemble.

15.3 CATE Model Validation

Now that we have a selected a winning CATE model or ensemble (e.g., the ensemble that comes out of Q-aggregation on the scoring data), we want to run formal statistical tests that validate whether the model that we chose contains any signal of treatment effect heterogeneity, or whether it is a confident model on average, or whether it is a useful model to drive personalized policy decisions as compared to simple benchmarks. We will refer to all these methodologies as CATE model validation and all the techniques can be thought as diagnostics that one should run on their CATE model before deployment or before using it to drive personalized decisions. As a side benefit, many of these diagnostics can also be used as a formal statistical test of the presence of treatment effect heterogeneity.

Throughout this section we assume that one has held out yet another dataset, called the test set (e.g., by splitting their data into train, validation and test) and that one has selected a CATE model τ_* without using the test set (e.g., by running some ensemble pipeline on on the train and validation set).

Heterogeneity Test Based on Doubly Robust BLP

If we calculate the doubly robust pseudo outcomes $Y^{DR}(\hat{\eta})$ on the test set (using cross-fitting within the test set to estimate the models $\hat{\eta}$ or using the union of the training and scoring data to estimate $\hat{\eta}$). Then we know that if the model of the CATE τ_* is good, the best linear predictor of the true CATE using $(1, \tau_*(X))$ as features should yield a statistically significant coefficient on the feature associated with the CATE model. In fact, in an ideal world this coefficient should be 1.

Thus we can run such a significance test to measure whether the CATE model τ_* has picked up any signal that is correlated with the true CATE. Note that if $\tau_0(X)$ is the true CATE $\mathrm{E}[Y(1)-Y(0) \mid X]$, then the coefficient associated with τ_* in the OLS regression $Y(\hat{\eta}) \sim (1, \tau_*(X))$ is converging in the population limit to the quantity:

$$\beta_1 := \frac{\mathrm{Cov}(\tau_0(X), \tau_*(X))}{\mathrm{Var}(\tau_*(X))} = \frac{\mathrm{Cov}(Y(1) - Y(0), \tau_*(X))}{\mathrm{Var}(\tau_*(X))} \quad (15.3.1)$$

Thus, the statistical test of whether β_1 is non-zero is a statistical test on the correlation of the individual treatment effect $Y(1) - Y(0)$ and the learned model $\tau_*(X)$. Note that if this test comes up as significant, then this also implies that there exists treatment effect heterogeneity, as a function of the observed features X. Moreover, the theory from the first section of this chapter applies here to show that the statistical test based on OLS regression is a valid test, as long as the product of the regression and propensity estimation errors, converges faster than $n^{-1/2}$.

Example 15.3.1 (Heterogeneity Test in 401(k) Example) Returning to our 401k example, we can split the data in three sets (train, score, test) and employ the heterogeneity test on the score set using the ensemble model chosen based on the Q-aggregation method of stacking. Out of the 9716 samples used in this analysis, 60% were used for training and 20% for scoring and 20% for testing. For the training of the nuisance parameters $\hat{\eta}$ that are used in the doubly robust proxy labels in the test set, we used the union of the (train, score) datasets, due to the relatively small size of the test set. Moreover, for better interpretability of the results we ran OLS on the centered CATE predictions, i.e. on features $(1, \tau_*(X) - \mathbb{E}_n \tau_*(X))$. This does not change the intepretation of β_1, but it changes the interpretation of the constant term as the ATE (which wouldn't have been the case without centering).

The results are depicted in Figure 15.14. We find that the ensemble model does indeed have a statistically significant correlation with the individual treatment effect $Y(1) - Y(0)$ and that the confidence interval on that coefficient includes the ideal coefficient of 1. We did find, however, large variation of the specific numbers reported in the table, dependent on the random split that was chosen in (train, score, test) sets from the original data.

	coef	**std err**	**P> |z|**	**[0.025**	**0.975]**
const	7634.4018	1805.185	0.000	4096.303	1.12e+04
$\tau_*(X)$	2.2406	0.718	0.002	0.833	3.648

Figure 15.14: OLS statistical test regression $Y^{DR}(\hat{\eta})$ on the features $(1, \tau_*(X) - \mathbb{E}_n\tau_*(X))$ in the 401(k) example. Standard Errors are heteroscedasticity robust (HC1). τ_* corresponds to the stacked ensemble based on Q-aggregation.

Validation Based on Calibration

A good CATE model should also be well calibrated. In the context of regression, a regression model $g(X)$ that predicts some outcome Y, is calibrated if the expected value of the outcome, conditional on the model returning a value of γ, should be equal to γ, i.e.

$$\mathrm{E}[Y \mid g(X) = \gamma] = \gamma. \tag{15.3.2}$$

Similarly, we can say that a CATE model τ_* is calibrated if the expected value of the treatment effect, conditional on the model returning a value of t, should be equal to t, i.e.

$$\gamma(\tau_*, t) := \mathrm{E}[Y(1) - Y(0) \mid \tau_*(X) = t] = t \tag{15.3.3}$$

Moreover, if we let P_{τ_*} denote the distribution of treatment effects returned by model τ_*, then we can define average calibration scores across different values of t. Some popular measures defined in the literature [17–19] are either the ℓ_2 or the ℓ_1-expected calibration error:

$$\mathrm{CAL}_1(\tau_*) := \int |\gamma(\tau_*, t) - t| \, d\mathrm{P}_{\tau_*}(t) \tag{15.3.4}$$

$$\mathrm{CAL}_2(\tau_*) := \int (\gamma(\tau_*, t) - t)^2 d\mathrm{P}_{\tau_*}(t) \tag{15.3.5}$$

In fact, an interesting property of the ℓ_2-calibration error, is that the MSE of a CATE model τ_* satisfies a calibration-distortion decomposition (analogous to the bias-variance decomposition):

$$\|\tau_* - \tau_0\|_{L^2} = \mathrm{CAL}_2(\tau_*) + \mathrm{DIS}(\tau_*) \tag{15.3.6}$$

where $\text{DIS}(\tau_*) = \text{E}\left[\text{Var}(\tau_0(X) \mid \tau_*(X))\right]$. Thus any consistent τ_* model will eventually also be calibrated. However, calibration is a self-consistency guarantee that should be desireable for many models and should not account for the majority of the MSE. Moreover, even if a model is far from τ_0, it is still desirable from a "steakholder experience" perspective that it should be calibrated.

The aforementioned desiderata can be taken to data by invoking again the proxy outcome regression approach. In particular, note that by the properties of the doubly robust proxy labels:

$$\gamma(\tau_*, t) = \text{E}[Y(\eta_0) \mid \tau_*(X) = t] \tag{15.3.7}$$

we can use observed data and out-of-sample estimates $\hat{\eta}$ of the nuisance functions η_0, to measure the calibration properties of a candidate CATE model.

To avoid having to run a non-parametric regression of $Y(\eta_0)$ on $\tau_*(X)$, in order to estimate the function $\gamma(\tau_*, t)$, a typical way that calbiration is evaluated is by looking at quantile bins of the distribution of CATE. For instance, if we let $q_1 \leq \ldots \leq q_K$ denote a set of K equally spaced quantiles of the distribution P_{τ_*}, then a well-calibrated model should satisfy that:

$$\text{E}[Y(\eta_0) \mid \tau_*(X) \in [q_t, q_{t+1}]] = \text{E}[\tau_*(X) \mid \tau_*(X) \in [q_t, q_{t+1}]]$$

In other words, consider any group G_t, defined defined by some quantile interval $[q_t, q_{t+1}]$ of the predictions of the model τ_*. Then the group average treatment effect (GATE) for the group G_t, should be the same, whether we calculate it by using the doubly robust GATE, i.e., $\text{E}[Y(\eta_0) \mid X \in G_t]$ or whether we calculate it by using the average CATE value of the model τ_* within that group, i.e., $\text{E}[\tau_*(X) \mid X \in G_t]$.

We can now easily take the latter approach to data. For some small K (e.g. $K = 4$), we can consider a set of thresholds $q_1 \leq \ldots \leq q_{K+1}$ that roughly approximate equally spaced quantiles of the CATE distribution P_{τ_*} and which are calculate without looking at the test sample (e.g. this can be calculated as the empirical quantiles of the empirical distribution of values of the model τ_* on the union of the training and scoring samples). These now define a set of K groups, $G_1, \ldots, G_K$ as described in the previous paragraph. Subsequently, we can estimate the GATE for each group, using the doubly robust approach on the

test data, i.e.

$$\hat{\theta}_k^{DR} = \frac{1}{|\{i \in [n] : X_i \in G_k\}|} \sum_{i \in [n]: X_i \in G_k} Y_i(\hat{\eta}) \tag{15.3.8}$$

where $\hat{\eta}$ is either estimated in a cross-fitting manner on the test set or using the union of training and scoring samples. Equivalently, we can simulaneoulsy estimated all these parameters by running OLS of $Y(\hat{eta})$ on the one-hot-encodings of the group membership indicator functions, as in the first section of the chapter. Moreover, confidence intervals can be directly obtained for these values (e.g. based on the OLS heteroskedasticity robust confidence intervals or based on the simple formula for the standard error of an average of i.i.d. observations; in this case we have the average of the $|\{i \in [n] : X_i \in G_k\}|$ observations $\{Y_i(\hat{\eta}) : i \in [n], X_i \in G_k\}$). These confidence intervals can also be used to test whether these different groups have statistically signficant different average treatment effects, i.e. whether the groups are separated statistically.

Moreover, these estimates can then also be used to construct approximate analogues of the ℓ_2 and ℓ_2-average calibration scores. For each group G_k, we can also calculate the average value of the model τ_*, i.e.,

$$\hat{\theta}_k^{*} = \frac{1}{|\{i \in [n] : X_i \in G_k\}|} \sum_{i \in [n]: X_i \in G_k} \tau_*(X_i) \tag{15.3.9}$$

Ideally, if the model was reasonable, $\hat{\theta}_k^{*}$ should be very close to $\hat{\theta}_k^{DR}$. The average difference can be considered as a quality metric of τ_*, i.e.,

$$\widetilde{\mathrm{CAL}}_1(\tau_*) := \sum_{k=1}^{K} \left|\hat{\theta}_k^{DR} - \hat{\theta}_k^{*}\right| \cdot |\{i \in [n] : X_i \in G_k\}| \tag{15.3.10}$$

$$\widetilde{\mathrm{CAL}}_2(\tau_*) := \sum_{k=1}^{K} \left(\hat{\theta}_k^{DR} - \hat{\theta}_k^{*}\right)^2 \cdot |\{i \in [n] : X_i \in G_k\}| \tag{15.3.11}$$

These can be viewed as binning approximations to the ℓ_1- and ℓ_2-average calibration scores (the first one was recommended as a calibration score in the context of randomized trials by [20]).

Example 15.3.2 (Calibration in the 401(k) example) We revisit the 401(k) example from the perspective of calibration. Following the same data analysis pipeline as in Example 15.3.1, we now also calculate the doubly robust GATEs for groups

defined by quartiles of the CATE distribution of the ensemble τ_* constructed based on Q-aggregation stacking. The bottom group corresponds to the bottom 25% of predicted CATEs, the next group to the 25%-50% of predicted CATEs, etc. In Figure 15.15, we depict on the x-axis the average CATEs, as calculated based on τ_*, within each group and on the y-axis and the doubly robust estimate and 5-95% confidence interval for the GATE as calculated based on the doubly robust proxy labels $Y(\hat{\eta})$ on the test set.

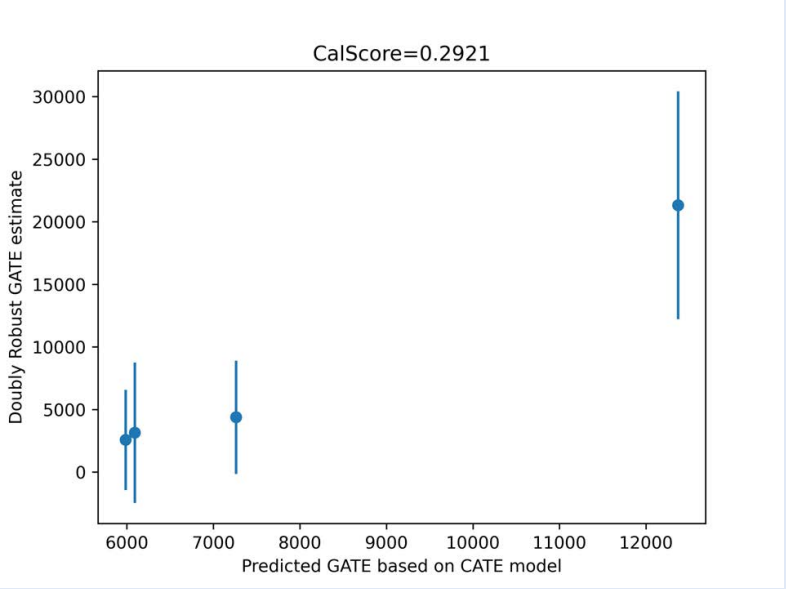


Figure 15.15: Calibration check for chosen ensemble model τ_* in the 401(k) example. Test samples are splitted in four groups based on CATE predictions and CATE quantiles (e.g. bottom group contains samples whose CATE predictions lie in the bottoms 25% of predictions). The x-axis depicts the average predicted CATE within each group based on τ_*, while the y-axis depicts the GATE as calculated based on the doubly robust pseudo-outcomes calculated on the test set.

Interpretation via Distillation and Group Differences. We can also try to interpret what are the differences of characteristics between the top and bottom CATE groups; if we find that they have statistically significantly different GATEs. We can do that by either reporting the mean values of the covariates in the two groups or building some interpretable classification model that distinguishes between the two groups.

	group1 mean ± s.e.	group2 mean ± s.e.	group1 - group2 mean ± s.e.
age	40.01 ± 0.27	42.56 ± 0.45	-2.56 ± 0.72
inc	26898 ± 346	65771 ± 760	-38873 ± 1106
fsize	2.82 ± 0.04	3.12 ± 0.07	-0.30 ± 0.11
educ	12.77 ± 0.07	14.74 ± 0.11	-1.97 ± 0.18
db	0.24 ± 0.01	0.37 ± 0.02	-0.14 ± 0.03
marr	0.52 ± 0.01	0.84 ± 0.02	-0.32 ± 0.03
male	0.22 ± 0.01	0.20 ± 0.02	0.02 ± 0.03
twoearn	0.29 ± 0.01	0.66 ± 0.02	-0.37 ± 0.03
pira	0.16 ± 0.01	0.42 ± 0.02	-0.26 ± 0.03
nohs	0.16 ± 0.01	0.02 ± 0.01	0.13 ± 0.02
hs	0.41 ± 0.01	0.26 ± 0.02	0.15 ± 0.03
smcol	0.24 ± 0.01	0.26 ± 0.02	-0.02 ± 0.03
col	0.19 ± 0.01	0.46 ± 0.02	-0.27 ± 0.03
hown	0.57 ± 0.01	0.84 ± 0.02	-0.27 ± 0.03

Figure 15.16: Group differences between the top 25% predicted CATE group (group2) and the bottom 75% predicted CATE group (group1) in the 401k example.

For instance, we can train a shallow binary classification tree that tries to predict whether a sample comes from the bottom or the top group, based on X, using the union of samples from the two groups.

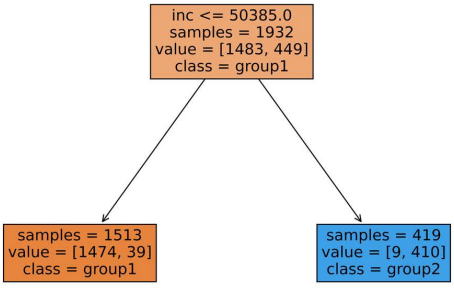


Figure 15.17: Decision tree that distills the main differences between group1 and group2 as defined in Figure 15.16.

Validation Based on Uplift Curves

Another way that we can judge the quality of a CATE model is by testing its ability to help us prioritize or stratify which part of the population we should be treating. In Section 14.3, we studied the value of the optimal policy subject to treating exactly a q-fraction of the overall population. In a sense, how much this value varies or how different it is from the ATE is a measure of the amount of uplift offered by personalizing

optimally based on X using the true CATE τ_0. We can study the same question from the lens of τ_*, which can give a lesser or equal level of uplift, the higher the uplift the better the model. Unlike Section 14.3, where τ_0 was a nuisance to be estimated and plugged into the optimal constrained or unconstrained policy, here τ_* is fixed and given (as it is based on a separate data set). In particular, our evaluation procedures would work even if τ_0 were hard to learn or had discontinuities in its distribution, since we focus on a fixed τ_* instead.

Let $\mu(\tau_*, q)$ denote an estimate based on non-test data of the $1-q$ quantile of the distribution P_{τ_*} of CATEs produced by the model τ_*. Then the group $\{X : \tau_*(X) \geq \mu(\tau_*, q)\}$ is fixed in terms of the test data. The corresponding GATE is

$$\mathrm{GATE}(q) := \mathrm{E}[Y(1) - Y(0) \mid \tau_*(X) \geq \mu(\tau_*, q)] \tag{15.3.12}$$

The improvement in the average effect of the treated, induced by the prioritization rule based on τ_*, as compared to treating a random q fraction of the population, would be:

$$\mathrm{TOC}(q) := \mathrm{GATE}(q) - \mathrm{ATE} \tag{15.3.13}$$

and the improvement in the total effect would be:

$$\mathrm{QINI}(q) := \mathrm{TOC}(q)\,\mathrm{P}(\tau_*(X) \geq \mu(\tau_*, q)). \tag{15.3.14}$$

For any fixed CATE model τ_*, the function $\mathrm{TOC}(\tau_*, \cdot)$ is referred to in the literature as the *Treatment Operating Characteristic* curve, while the function $\mathrm{QINI}(\tau_*, \cdot)$ is referred to as the *QINI* curve (analogous to the Gini curve for classification models).[10]

10: These terminologies primarily stem from the uplift modelling literature in Computer Science [21–23].

These curves also have interesting interpretations as covariances of the individual treatment effect $Y(1) - Y(0)$ with non-linear functions of the CATE model τ_* (see proofs in Appendix 15.B):

$$\begin{aligned}
\mathrm{TOC}(q) &= \mathrm{Cov}\left(Y(1) - Y(0), \frac{1\{\tau_*(X) \geq \mu(\tau_*, q)\}}{\mathrm{P}(\tau_*(X) \geq \mu(\tau_*, q))}\right) \\
\mathrm{QINI}(q) &= \mathrm{Cov}\left(Y(1) - Y(0), 1\{\tau_*(X) \geq \mu(\tau_*, q)\}\right)
\end{aligned}$$

Since the second term in each covariance is a function of X alone and $\mathrm{E}[Y(1) - Y(0) \mid X] = \mathrm{E}[Y(\eta_0) \mid X]$, these quantities are identified by replacing the individual effects with the doubly robust pseudo-outcomes:

$$\begin{aligned}
\mathrm{TOC}(q) &= \mathrm{Cov}\left(Y(\eta_0), \frac{1\{\tau_*(X) \geq \mu(\tau_*, q)\}}{\mathrm{P}(\tau_*(X) \geq \mu(\tau_*, q))}\right) \\
\mathrm{QINI}(q) &= \mathrm{Cov}\left(Y(\eta_0), 1\{\tau_*(X) \geq \mu(\tau_*, q)\}\right)
\end{aligned}$$

Area Under the Curve (AUC) Viewing the above two quantities as functions of the target fraction q, we can calculate the areas under these two curves as scalar measures of quality of the model τ_* in its ability to correctly target sub-parts of the population at different levels of treatment population size targets, i.e.

$$AUTOC := \int_0^1 \text{TOC}(q)dq \tag{15.3.15}$$

$$AUQC := \int_0^1 \text{QINI}(q)dq \tag{15.3.16}$$

The larger the Area Under the Curve, the better the CATE model is at treatment prioritization or stratification.

Moreover, these measures are signals of treatment effect heterogeneity. If any of the two measures are statistically non-zero, then treatment effect heterogeneity was detected with statistical significance. In fact, if we detect that any of these curves lies above zero at any point q, with statistical significance, then that also serves as a test for treatment effect heterogeneity. For this reason, we will now develop confidence intervals and simultaneous confidence bands for these two curves, when they are estimated from samples.

Remark 15.3.1 (Tie-Breaking) If our CATE model returns a constant effect for a large segment of the population, then the quantile estimate function $\mu(\tau_*, q)$ will contain many flat regions and ties cannot be ignored. In this case, if we want to approximately target a q-fraction of the population, then we should be treating deterministically everyone with $\tau_*(X) > \mu(\tau_*, q)$ and units with $\tau_*(X) = \mu(\tau_*, q)$, we should be treating with probability:

$$\frac{q - \text{P}(\tau_*(X) > \mu(\tau_*, q))}{\text{P}(\tau_*(X) = \mu(\tau_*, q))}$$

which corresponds to the probability mass that remains after treating everyone above $\mu(\tau_*, q)$ (i.e. $q - \text{P}(\tau_*(X) > \mu(\tau_*, q))$), divided by the probability mass in the group of units that have predicted CATE equal to $\mu(\tau_*, q)$. We can then use an estimate of this quantity using the training and scoring datasets, e.g.

$$\lambda = \frac{q - \text{P}_n(\tau_*(X) > \mu(\tau_*, q))}{\text{P}_n(\tau_*(X) = \mu(\tau_*, q))}$$

and consider the policy that treats deterministically for units

with $\tau_*(X) > \mu(\tau_*, q)$ and treats with probability λ for units with $\tau_*(X) = \mu(\tau_*, q)$. In this case, the TOC and QINI curves will take a slightly more complex form:

$$\text{Cov}\left(Y(\eta_0), \frac{1\{\tau_*(X) > \mu(\tau_*, q)\} + \lambda 1\{\tau_*(X) = \mu(\tau_*, q)\}}{\text{P}(\tau_*(X) \geq \mu(\tau_*, q)) + \lambda \text{P}(\tau_*(X) = \mu(\tau_*, q))}\right)$$
$$\text{Cov}\left(Y(\eta_0), 1\{\tau_*(X) \geq \mu(\tau_*, q)\} + \lambda 1\{\tau_*(X) = \mu(\tau_*, q)\}\right)$$

All the conclusions and intution in what follows, directly extends to account for tie-breaking, so we will omit tie-breaking for simplicity of exposition.

Estimation and inference. To estimate the TOC and the QINI curves, we will use the doubly robust proxy outcome approach. We will train nuisance models $\hat{\eta}$ without using the test sample and then construct estimates of the TOC and QINI curves as:

$$\widehat{\text{TOC}}(q) = \text{Cov}_n\left(Y(\hat{\eta}), \frac{1\{\tau_*(X) \geq \mu(\tau_*, q)\}}{\hat{\pi}(q)}\right) \tag{15.3.17}$$

$$\widehat{\text{QINI}}(q) = \text{Cov}_n\left(Y(\hat{\eta}), 1\{\tau_*(X) \geq \mu(\tau_*, q)\}\right) \tag{15.3.18}$$

where $\hat{\pi}(q) = \mathbb{E}_n 1\{\tau_*(X) \geq \mu(\tau_*, q)\}$ and we used the shorthand notation:

$$\text{Cov}_n(A, B) = \mathbb{E}_n\left[(A - \mathbb{E}_n(A))(B - \mathbb{E}_n(B))\right]$$

Both of these estimates are of the general estimation form that can be handled by the Neyman orthogonality framework. For each q, we can view each of the estimates as an estimate of the form:

$$\hat{\theta}(q; \nu) = \mathbb{E}_n[\psi(W; \nu)]$$

for some appropriate defined function ψ and with ν being a vector of nuisance quantities, which contain η, π and $\theta_0 = \text{E}[Y(\eta_0)]$ and which satisfies Neyman orthogonality with respect to all of these nuisance quantities. Thus these estimates will be asymptotically Gaussian with the effect of the nuisances being negligible. Moreover, even if we evaluate these curves at many points, as long as the number of points q that we use does not grow exponentially with the sample size, then these estimates will be jointly Gaussian and we can construct simultaneous confidence bands as in Section 4.4.

Theorem 15.3.1 *Let $q \in Q := \{q_1, \ldots, q_p\}$ denote a grid of quantiles. Let $\alpha = (\alpha_1, \ldots, \alpha_p)$ denote the p-dimensional vector*

whose t-th coordinate is $TOC(q)$ and $\hat{\alpha}$ the corresponding vector of estimates $\widehat{TOC}(q)$. Let $\mathbb{I}(q) := 1\{\tau_*(X) \geq \mu(\tau_*, q)\}$ and:

$$\psi_\ell(W) = (Y(\eta_0) - \theta_0)\left(\frac{\mathbb{I}(q_\ell)}{\pi_0(q_\ell)} - 1\right) - \alpha_\ell$$

with $\theta_0 = \mathrm{E}Y(\eta_0)$ and $\pi_0(q) = \mathrm{E}\mathbb{I}(q)$. Suppose that the nuisance estimates $\hat{\eta}$ are trained on a separate sample and satisfy,

$$\sqrt{n}\|H(\hat{\mu}) - H(\mu_0)\|_{L^2}\|\hat{g} - g_0\|_{L^2} \approx 0$$
$$\|H(\hat{\mu}) - H(\mu_0)\|_{L^2} + \|\hat{g} - g_0\|_{L^2} \approx 0$$

Provided that $\log(p)^5/n$ is small and the estimates satisfy the adaptivity property:

$$\sqrt{\log(p)}\max_{\ell=1}^{p}\left|\sqrt{n}(\hat{\alpha}_\ell - \alpha_\ell) - \mathbb{E}_n\phi_\ell(W)\right| \approx 0$$

the following Gaussian approximation holds:

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \overset{a}{\sim} N(0, V),$$

where

$$V_{\ell k} = \mathrm{E}\psi_\ell(W)\,\psi_k(W)$$

Analogous theorem also applies to the QINI curve estimates.

This result can be used to construct simultaneous confidence bands for the value of the TOC curve at many quantiles q as described in Remark 4.4.1. We can consider the estimate of the variance:

$$\hat{V}_{\ell k} = \mathbb{E}_n\hat{\psi}_\ell(W)\hat{\psi}_k(W) \quad \hat{\psi}_\ell(W) = (Y(\hat{\eta}) - \hat{\theta})\left(\frac{\mathbb{I}(q_\ell)}{\hat{\pi}(q_\ell)} - 1\right) - \hat{\alpha}_\ell$$

and construct a confidence band at confidence level α:

$$CR = \times_{\ell=1}^{p}[\hat{\alpha}_\ell \pm c\sqrt{\hat{V}_{\ell\ell}/n}]$$

where c is the $1-\alpha$ quantile of the distribution of $\|Z\|_\infty$ for a random variable $Z \sim N(0, \hat{D}^{-1/2}\hat{V}\hat{D}^{-1/2})$, where $\hat{D} = \mathrm{diag}(\hat{V})$ is the matrix with diagonal entries $\hat{V}_{\ell\ell}$ and zero off-diagonal entries.

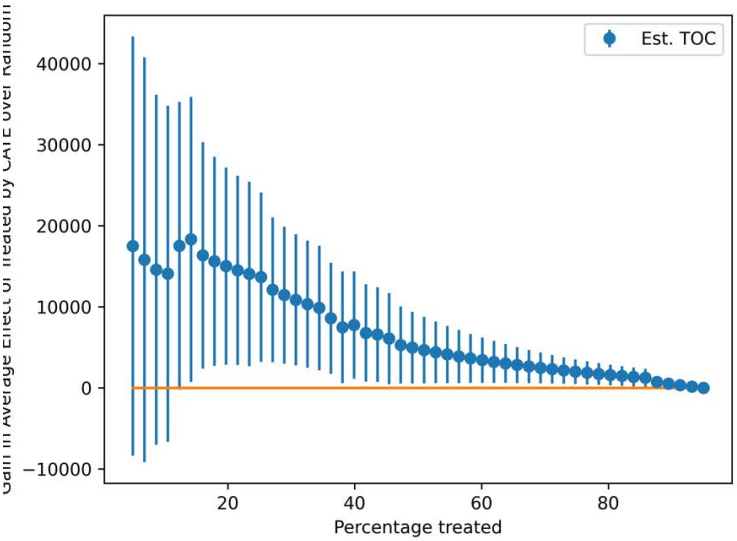


Figure 15.18: Point estimates and uniform confidence band of the TOC curve for the Q-aggregation ensemble τ_* in the 401(k) example.

Note that if there is any point that is above the zero line, with confidence, in this curve, then the CATE model $\hat{\tau}$ has identified heterogeneity in the effect in a statistically significant manner. For such a test we can calculate a one-sided confidence interval, as we only care that the quantities are larger than some value with high confidence. Using the Gaussian approximation, a one-sided confidence band, at confidence level α, can be calculated as:

$$CR = \times_{\ell=1}^{p} \left[\hat{\alpha}_\ell - c\sqrt{\hat{V}_{\ell\ell}/n}, \infty \right)$$

where c is the $1 - \alpha/2$ quantile of the distribution of $\|Z\|_\infty$ for a random variable Z as defined in the previous paragraph.

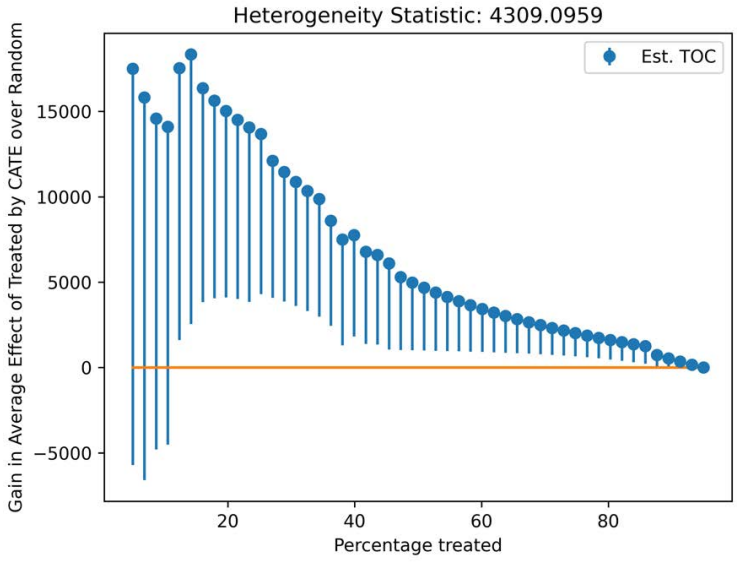


Figure 15.19: Point estimates and one-sided uniform confidence band of the TOC curve for the Q-aggregation ensemble τ_* in the 401(k) example. The heterogeneity statistic depicted in the title corresponds to the largest lower bound of the confidence band across all quantile points and is a statistical signal for the presence treatment effect heterogeneity.

We can also calculate the area under the curve using the discrete difference approximation:

$$\widehat{AUTOC} = \sum_{\ell=1}^{p} \widehat{\text{TOC}}(q_\ell)\,(q_{\ell+1} - q_\ell) \tag{15.3.19}$$

Note that since this is a linear combination of the estimates at each q_ℓ, under the assumptions of Theorem 15.3.1, the estimate of the area under the curve will be asympotically normal and centered around the quantity:

$$AUTOC = \sum_{\ell=1}^{p} \text{TOC}(q_\ell)\,(q_{\ell+1} - q_\ell)$$

and we can calculate a one-sided confidence interval as:

$$AUTOC \in \left[\widehat{AUTOC} - \sqrt{\hat{V}/n}, \infty\right)$$

where the estimate of the variance is:

$$\hat{V} = \mathbb{E}_n \hat{\psi}(W)^2 \qquad \hat{\psi}(W) = \sum_{\ell=1}^{p} \hat{\psi}_\ell(W)\,(q_{\ell+1} - q_\ell)$$

If the confidence interval does not contain zero, then we have again detected heterogeneity.

AUTOC	s.e.	One-Sided 95% CI
5228.9137	1471.8731	[2807.8980, Infty]

Figure 15.20: AUTOC point estimate and one-sided confidence interval for the Q-aggregation ensemble τ_* in the 401(k) example.

The exact same analysis can be conducted for the QINI curve, constructing doubly robust point estimates and a simultaneous one-sided confidence band, as well as a one-sided confidence interval for the discretized quantile approximation of the AUQC, i.e.

$$AUQC = \sum_{\ell=1}^{p} \text{QINI}(q_\ell)\,(q_{\ell+1} - q_\ell)$$

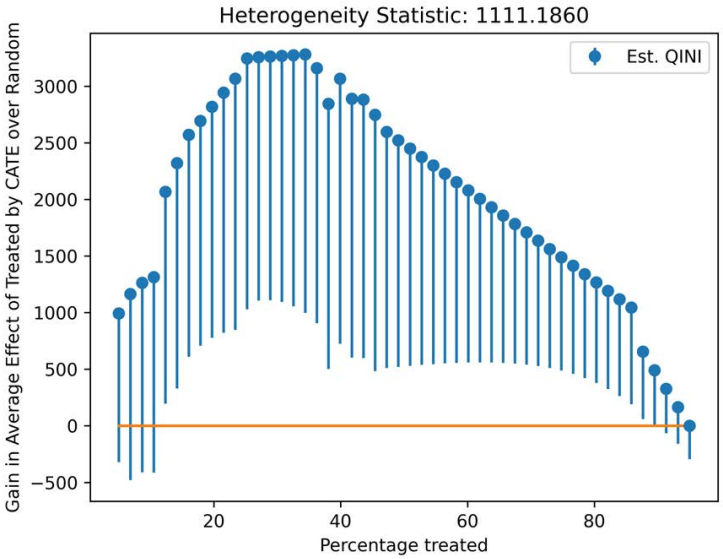


Figure 15.21: Point estimates and one-sided uniform confidence band of the QINI curve for the Q-aggregation ensemble τ_* in the 401(k) example. The heterogeneity statistic depicted in the title corresponds to the largest lower bound of the confidence band across all quantile points and is a statistical signal for the presence treatment effect heterogeneity.

AUQC	s.e.	One-Sided 95% CI
1542.4581	385.2292	[908.8125, Infty]

Figure 15.22: AUQC point estimate and one-sided confidence interval for the Q-aggregation ensemble τ_* in the 401(k) example.

Remark 15.3.2 (Rank-Average Weighted Treatment Effects) The analysis in this section viewed the quantile function $\mu(\tau_*, q)$ as fixed and considered a variant of the uplift curves based on the targeting policy $\pi_q(X) := 1\{\tau_*(X) \geq \mu(\tau_*, q)\}$. If this was the policy that was deployed in the population, then we define the TOC curve at value q, as the average effect of the treated population of policy π_q and we perform inference on this quantity. An alternative view of the TOC curve is to consider the ranking viewpoint that at deployment time will rank the population based on the prediction models predictions and will treat exactly the top q fraction of the population. From this viewpoint, the accuracy of the quantile estimate matters a lot and should be incorporated into the uncertainty estimates. Quantifying the uncertainty that stems from estimation errors in the quantiles $\mu(\tau_*, q)$ is a more involved topic. The recent work of [24] takes this view and performs inference on the ranking interpretation of the TOC and QINI curves, correctly accounting for the uncertainty in the estimation of the quantiles of the CATE distribution and offers procedures for asymptotically correct confidence intervals (albeit not confidence bands).

15.4 Personalized Policy Learning

In Section 14.3 we studied evaluation of personalized policies, in particular optimal ones. However, we did not delve into the learning of optimal policies, just as we discussed inference on CATE inChapter 14 but did not delve into learning it using flexible non-parametric methods. While any CATE model learned as in the present chapter can be used to prioritize treatment, a CATE model would only be a means to an end and not the object of interest itself, which may be learned more directly. The primary object of interest is a good personalized treatment policy π that given any instance of the variable X returns a treatment assignment $\pi(X) \in \{0, 1\}$.

Note that learning a good policy is an inherently different statistical task than learning a good CATE model. For a good unconstrained policy, it suffices that we learn whether the CATE $\tau(X) = \mathrm{E}[Y(1) - Y(0) \mid X]$ is positive or negative. As seen in Section 14.3, the optimal policy is given by looking at the sign of CATE: $\pi^*(X) = \mathbb{1}\{\tau_0(X) \geq 0\}$. Thus policy learning is more akin to a classification problem that tries to predict the sign of the CATE as opposed to a regression problem that tries to learn the magnitude of CATE too. Of course, mistakes in predicting the sign are more detrimental when the magnitude of the CATE is larger and therefore should be weighed differently. Thus policy learning is more accurately described as a classification problem with sample dependent mis-classification costs, known in the machine learning literature as *cost-sensitive classification*.

Recall from Section 14.3 that we define in Eq. 14.3.1 the gains of policy over no treatment as $V(\pi) = \mathrm{E}[\pi(X)\, Y(1) + (1 - \pi(X))\, Y(0)] - \mathrm{E}[Y(0)] = \mathrm{E}[\pi(X)\, Y(\eta_0)]$. Optimizing $V(\pi)$ over π is equivalent to a sample-weighted classification problem, where the goal of π is to match the sign of $Y(\eta_0)$, with sample weights $|Y(\eta_0)|$. More formally, note that:

$$\mathrm{argmax}_\pi V(\pi) = \mathrm{argmax}_\pi \mathrm{E}\left[(2\pi(X) - 1)\, Y(\eta_0)\right]$$

and we can simplify the latter as:

$$\begin{aligned}\mathrm{E}\left[(2\pi(X) - 1)\, Y(\eta_0)\right] &= \mathrm{E}\left[(2\pi(X) - 1)\ \mathrm{sign}\left(Y(\eta_0)\right)\ |Y(\eta_0)|\right] \\ &= \mathrm{E}\left[\mathbb{1}\left\{2\pi(X) - 1 = \mathrm{sign}\left(Y(\eta_0)\right)\right\} |Y(\eta_0)|\right]\end{aligned}$$

Thus we can treat the sign of $Y(\eta_0)$ as the "label" of the sample in a classification problem and $|Y(\eta_0)|$ as the weight of the sample, and our centered treatment policy $2\pi(Z) - 1$ is trying to predict the label. We can therefore invoke any machine learning classification approach in a meta-learning manner, so

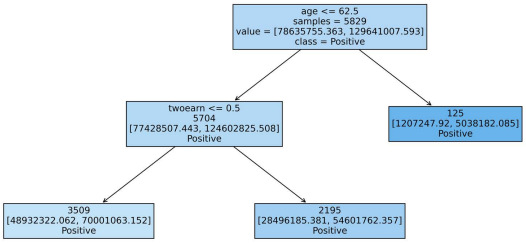


Figure 15.23: The details that are displayed on each node are also useful in understanding the group average treatment effect for each node. In particular, the information 'samples=N', gives us the size of each node N, and the information 'value=[A, B]', then 'A' is the sum of the $|Y(\hat{\eta})|$ for the samples where $Y(\hat{\eta}) < 0$ and similarly, 'B' is the sum of $|Y(\hat{\eta})|$ for the samples where $Y(\hat{\eta}) > 0$. Thus to get the GATE for each node, we simply do '(B-A)/N', which would correspond to $\frac{1}{N}\sum_{i \in \text{node}} Y(\hat{\eta})$, which is the doubly robust estimate of the GATE for the node.

as to solve this weighted classification problem. One popular approach is to use a decision tree classifier, since it will lead to an interpretable policy that is easy to visualize.

In finite samples, we would also need to construct estimates $\hat{\eta}$ of the nuisance parameters η_0 in a cross-fitting manner using arbitrary ML regression methods, as discussed in prior sections and then solve a sample weighted classification problem with samples $\{(X_i, \text{sign}(Y_i(\hat{\eta})), W_i = |Y_i(\hat{\eta})|\}_{i=1}^n$. The results in [25] show that the regret of the returned policy $\hat{\pi}$, as compared to the optimal policy within some policy space Π, i.e.:

$$R(\hat{\pi}) = \max_{\pi_* \in \Pi} V(\pi_*) - V(\hat{\pi}) \tag{15.4.1}$$

inherit the double robustness property and decay at the order of

$$\approx \sqrt{\frac{V_* \, VC(\Pi)}{n}} + \|H(\hat{\mu}) - H(\mu_0))\|_{L^2} \, \|\hat{g} - g_0\|_{L^2}$$

where $VC(\Pi)$ is a measure of statistical complexity of the policy space Π (e.g. a small constant for shallow binary decision trees) and V_* is a constant that in many practical scenarios can be thought as some constant multiple of the variance of the value of the optimal policy in the class $\pi_* = \text{argmax}_{\pi \in \Pi} V(\pi)$, i.e.

$$V_* \approx \text{Var}(\pi_*(X) Y(\eta_0))$$

See also [4, 26] for generalizations and variations of this result.

Remark 15.4.1 (Probabilistic Policies) The aforementioned analysis also applies if we allow our policy space to output

probabilistic choices, i.e. $\pi(X) \in [0, 1]$ denotes the probability of treatment. In this case, our objective can equivalently be thought as optimizing a weighted classiciation problem of the form:

$$\mathrm{E}\left[\mathrm{P}_{D\sim\pi}[(2D-1) = \mathrm{sign}(Y(\eta_0))]\,|Y(\eta_0)|\right] \tag{15.4.2}$$

Remark 15.4.2 (Variance Penalization Methods) One caveat of treating the policy optimization problem as a weighted classification problem and calling a classification oracle is that we might be artificially favoring policies that have high variance. In particular, suppose that some policy π assigns a large probability to a treatment at some region of X in which the observed data that has a very low probability. In this case, the variance of this policy is very large, due to the fact that we are dividing by the propensity in the observed data. In this case, one would expect that V_* in the aforementioned regret rate will be a very large multiple of the variance of the optimal policy $\mathrm{Var}(\pi_*(X)\,Y(\eta_0))$. To avoid dependence on the worst case such overlap ratio between any policy in Π and the observed policy, i.e. $\sup_{x\in X} \frac{\pi(x)}{\mu_0(x)}$, one needs to amend the objective function that we are optimizing to penalize policies that are expected to have large variance (equiv. small overlap with the policy that was deployed in the observational data). In it's simplest form, one can invoke explicit variance penalization in the empirical objective:

$$\max_{\pi\in\Pi} \mathbb{E}_n\left[\pi(X)\,Y(\hat{\eta})\right] - \lambda\sqrt{\mathrm{Var}_n(\pi(X)\,Y(\hat{\eta}))} \tag{15.4.3}$$

where λ is a hyper-parameter that for policy spaces with bounded VC dimension should be set to some constant multiple of $\sqrt{\frac{VC(\Pi)\log(n)}{n}}$.

The one caveat of this approach is that the optimization problem is no longer a simple classification problem and one cannot invoke an out-of-the-box ML classification oracle. Several other approaches have been proposed in the literature that have benefits either on the computational side or on the statistical side, such as distributionally robust optimization [27] (i.e. optimizing the worst case policy value over a ball of distributions that are close to the empirical distribuiton), pessimism [28] (i.e. optimizing a proxy of the lower bound of a confidence interval for the value of a policy), out-of-sample regularization [26] (i.e. optimizing policies that also achieve small error on a held-out sample). This is an active area of

research, especially in the area of reinforcement learning and is referred to in the literature as offline policy learning or offline reinforcement learning. Even more complex is the problem of adaptively collecting data and randomizing in an adaptive manner, so as to find an optimal policy, which is referred to in the literature as online policy learning or online reinforcement learning. See [29] for a recent survey.

15.5 Empirical Example: The "Welfare" Experiment

We revisit the welfare experiment dataset that we analzed in Chapter 14 and deploy all the methods described in this section. We remind that this dataset corresponds to an experiment that was run as part of the General Social Survey (GSS)[11] , where some respondents received a questionaire about their willingness to support a "Welfare Program" (which will be viewed as the treatment, i.e. $D = 1$, in our analysis), while others received the same questionaire but the program was referred to as "Assistance to the Poor"(which will be viewed as the control, i.e. $D = 0$, in our analysis).

11: See e.g. https://gssdataexplorer.norc.org/variables/vfilter for a full description of the variables in the survey.

After some preprocessing, the dataset contains 12907 individuals and 42 covariates. Instead of simply estimating the projection of the CATE onto a simple model that is linear in the political views variable or its one-hot-encoding, we instead train generic ML models based on all the methods outlined in this chapter. We then score each of the models and construct an ensemble CATE model using Q-aggregation.

In Figure 15.24 we depict the predictions of the Q-aggregation ensemble, as a function of the political views variable, fixing all other covariates to their median values. We find that the fully data-driven model did pick up political views as a relevant variable, but the degree of variation is much smaller than the one that is identified using the doubly robust method for the projection of the CATE on political views. Potentially, this demonstrates that other variables are also relevant and some of the variation picked up by the CATE projection models should have been attributed to other covariates that co-vary with political views.

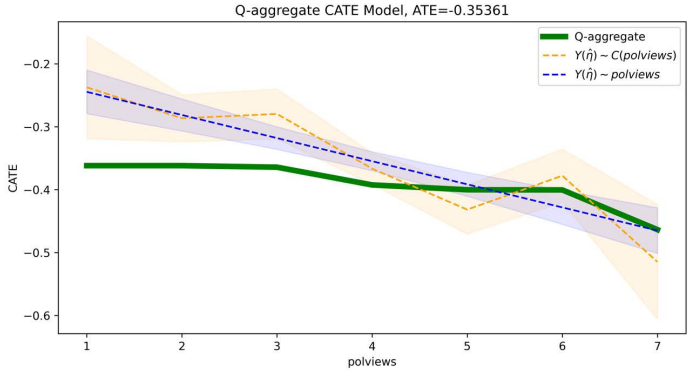


Figure 15.24: CATE predictions of the Q-aggregation stacked ensemble. Gradient boosted forests (via the xgboost library) were used as ML oracles for regression and classification. The CATE is predicted on a grid of income points, corresponding to equally spaced income quantiles. All other covariates were imputed at their median values. For comparison, each plot also displays the doubly robust best linear predictor of the CATE with 5-95% confidence intervals as a linear function of the covariate 'polviews' and as a linear function of the one-hot-encoding of the covariate 'polviews'.

To understand the heterogeneity patterns that were identified by the ensemble CATE model, we fit a single shallow binary decision tree to the predictions of the CATE ensemble model. We depict the learned tree in Figure 15.25. We see that political views is indeed the single most important factor that discriminates the predictions of the learned model, however we also see that the ensemble model also learned that education and race also creates heterogeneity in the reaction to programs labeled as "welfare" (as opposed to "assistance to the poor"). In particular, the model identified that people with more left-wing political views and more than 15 years of education (i.e., 4-year college educated individuals) have the least adverse reaction to the word "welfare", while more right wing individuals who did not identify as black (i.e., race2=0) have the most adverse reaction to the term "welfare". Moreover, we see that political views alone does not create a large variation, but it is the combination of political views and college education that creates the largest degree of heterogeneity in the effect.

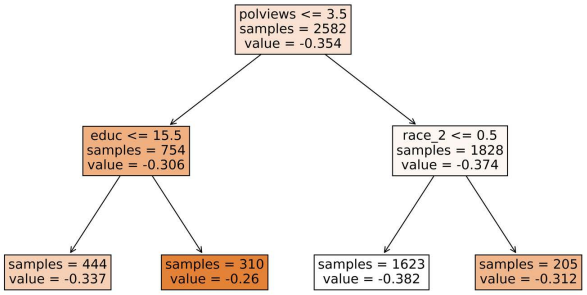


Figure 15.25: Single binary regression tree distillation of the Q-aggregation based stacked ensemble.

An alternative way to visualize the importance of the different variables in changing the output of the CATE ensemble is by using the SHAP values. In Figure 15.26. These values identify how each individual variable contributes to changes in the output of the ensemble model. We again identify that political views and education create the largest variation in the output, though here we see that other variables can also be attributed changes in the prediction, such as the number of hours worked last week (hrs1). We see here that having worked less hours last week increases the output of the model, i.e., leads to less adverse reaction to the word "welfare". So people that worked more hours were less eager to contribute to a program termed "welfare".

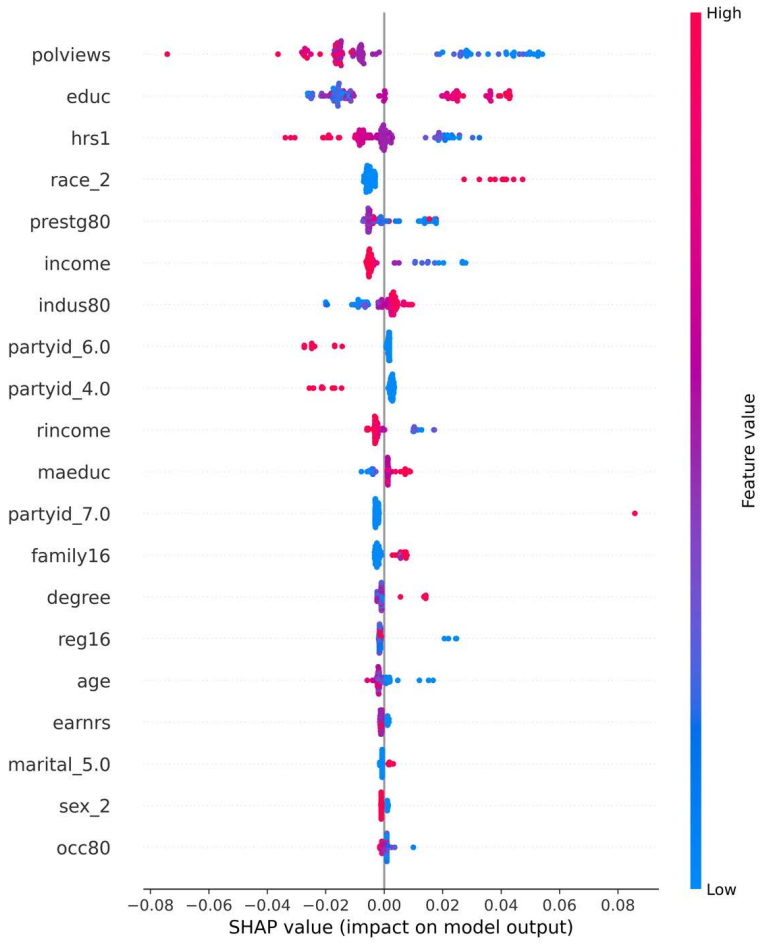


Figure 15.26: SHAP values for the Q-aggreagation based stacked ensemble in the welfare experiment dataset.

We can also validate the learned model by running several statistical tests on a held-out sample. For instance, in Figure 15.27 we run an OLS regression of the doubly robust outcome $Y(\hat{\eta})$ on $(1, \tau_*(X))$. We find that the coefficient associated with the stacked ensemble was statistically significant and the confidence interval included the value 1. Hence, this validates that the model carries significant information on the heterogeneity of

the effect.

	coef	std err	P> |z|	[0.025	0.975]
const	-0.3839	0.016	0.000	-0.416	-0.352
$\tau_*(X)$	1.4655	0.267	0.000	0.943	1.988

Figure 15.27: OLS statistical test regression $Y(\hat{\eta})$ on $(1, \tau_*(X))$ in the Criteo example. Standard Errors are heteroscedasticity robust (HC1). τ_* corresponds to the stacked ensemble based on Q-aggregation.

We also evaluate how calibrated the model is by depicting the group average treatment effects for each quartile of the predicted CATE distribution. The GATE was estimated using the doubly robust approach on the held-out sample. We see that the bottom and top quartiles are separated in a statistically significant manner, while also the calibration score of the model is quite high (0.4461).

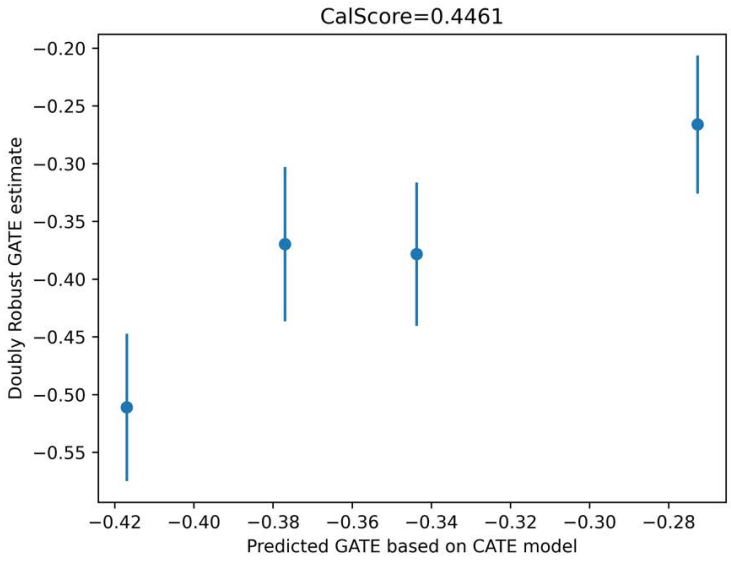


Figure 15.28: Calibration check for chosen ensemble model τ_* in the welfare example. Test samples are splitted in four groups based on CATE predictions and CATE quantiles (e.g. bottom group contains samples whose CATE predictions lie in the bottoms 25% of predictions). The x-axis depicts the average predicted CATE within each group based on τ_*, while the y-axis depicts the GATE as calculated based on the doubly robust pseudo-outcomes calculated on the test set.

Given that we identified that the bottom and top quartile are different in a statistically significant manner, we can also visualize the differences of these two groups, by simply depicting the difference in means of each of the covariates in the two groups. We see for instance, that hours worked last week was significantly different in the two groups, as well as income, age, political views and education, reinforcing our prior findings.

	group1 mean ± s.e.	group2 mean ± s.e.	group1 - group2 mean ± s.e.
hrs1	44.16 ± 0.30	36.74 ± 0.58	7.42 ± 0.88
income	11.52 ± 0.03	10.61 ± 0.10	0.92 ± 0.12
rincome	10.58 ± 0.05	9.23 ± 0.14	1.35 ± 0.19
age	41.75 ± 0.27	37.52 ± 0.49	4.23 ± 0.76
polviews	4.39 ± 0.03	3.07 ± 0.05	1.32 ± 0.08
educ	13.58 ± 0.06	15.41 ± 0.11	-1.82 ± 0.17
earnrs	1.81 ± 0.02	1.60 ± 0.03	0.21 ± 0.05
sibs	3.40 ± 0.06	3.42 ± 0.14	-0.02 ± 0.20
childs	1.68 ± 0.03	1.24 ± 0.06	0.44 ± 0.09
occ80	351.52 ± 5.66	280.24 ± 8.53	71.28 ± 14.19

Figure 15.29: Group differences between the top 25% predicted CATE group (group2) and the bottom 25% predicted CATE group (group1) in the welfare example.

We can also visualize the differences between the two groups by fitting a shallow binary classification tree to predict membership in the top quartile vs bottom quartile groups. We see again that political views, education and race are the most important distinguishing factors for membership in the two groups. For instance, as we see in Figure 15.30, in the held-out dataset, among the 309 college-educated and left-wing individuals, only 5 were in the bottom quartile group (which had a statistically signficant more adverse reaction to welfare), while 304 were in the top quartile group. Similarly, among the 1622 right-wing and not black individuals, 1541 were in the bottom quartile group vs. 81 in the top quartile group.

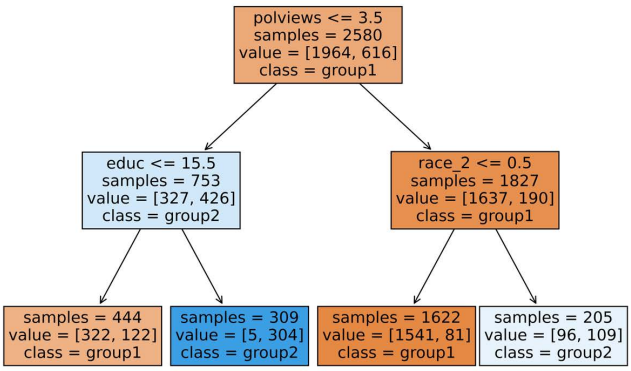


Figure 15.30: Decision tree that distills the main differences between group1 and group2 as defined in Figure 15.29.

Finally, we can verify that we detected a statistically significant heterogeneity of effect by looking at the uplift curves, i.e. the TOC (c.f. Figure 15.31) and QINI (c.f. Figure 15.32) curves. We find that both curves lie above the zero line, even when we

incorporate one-sided confidence bands. The largest lower point of this confidence band is depicted as a heterogeneity statistic in the title. For instance, we see that the largest lower point is 0.1039 in the TOC curve, which occurs at roughly 5%. This means that, with 95% confidence level, if we look at the group that corresponds to the top 5% of the CATE predictions, then we expect to see an average effect within that group that is at least 0.1039 larger than the average effect in the overall population.

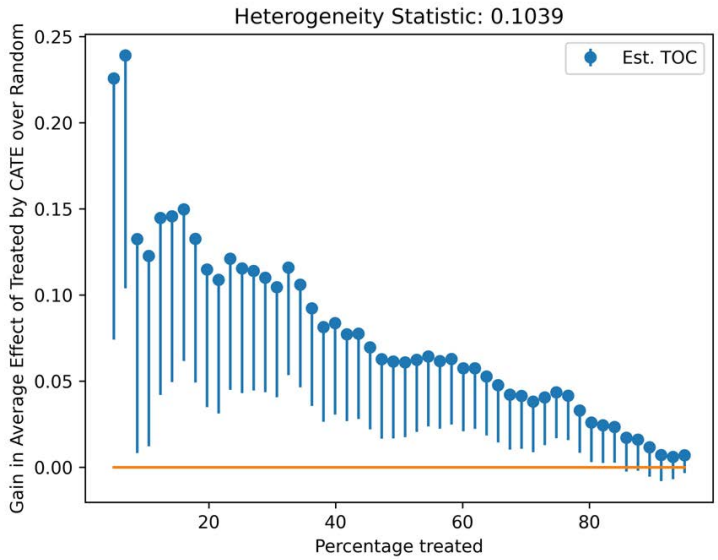


Figure 15.31: TOC curve with a 95% one-sided confidence band for the welfare experiment dataset.

Similarly, in the QINI curve we find that this heterogeneity statistic is 0.0164 and occurs at roughly 30%, which means that if were to treat the group of people that corresponds to the top 30% of CATE predictions, then we would expect to get a total effect in the population that is 0.0164 larger than if we were to treat a random 30% fraction of the population.

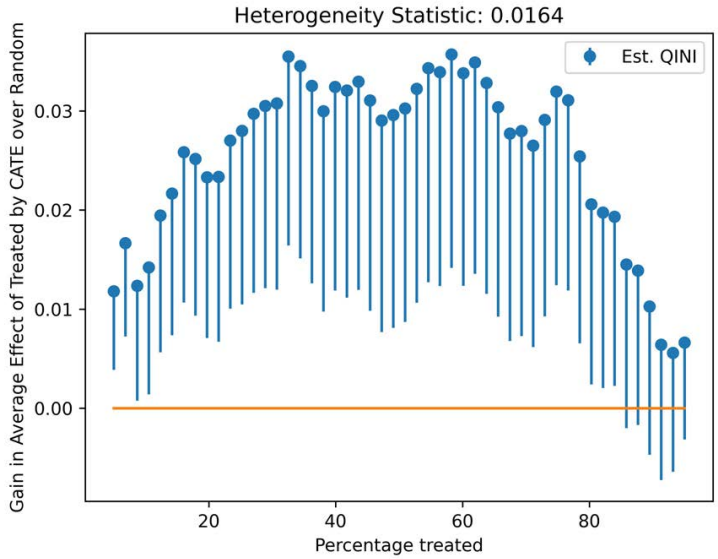


Figure 15.32: QINI curve with a 95% one-sided confidence band for the welfare experiment dataset.

We can also calcualte the area under these curves and the confidence interval for that area. If the confidence interval does not contain zero, then we have again detected heterogeneity with statistical significance.

AUTOC	s.e.	One-Sided 95% CI
0.0667	0.0128	[0.0457, Infty]

AUT Qini	s.e.	One-Sided 95% CI
0.0232	0.0046	[0.0156, Infty]

15.6 Empirical Example: Digital Advertising A/B Test

We now consider an application of CATE estimation in the context of estimating the effects of digital advertising. We will be using a publicly available dataset released by the digital advertising company Criteo®[30]. The dataset consists of approximately 14 million samples (each corresponding to an online visitor) and 10 (anonymized) features that describe the user and the context of the visit. This dataset is the combination of several incrementality tests ran by the company. In each incrementality test a random subset of the population is prevented from being targeted by digital advertising. Subsequently the company tracks whether the user visited or not the webpage

that corresponded to the ad that was being shown in some period after the advertising campaign. The latter will be the outcome of interest. Thus we will be measuring the effects of digital advertising on drawing traffic to a particular webpage (the dataset also contains other relevant outcomes such as "conversion", i.e. whether the visitor purchased an associated product).

We applied the CATE estimation pipeline outlined in this section. In Figure 15.33 we depict the performance of each of the meta-learners, compared to the performance of a baseline model that fits a constant treatment effect, as measured by the doubly robust score (see Theorem 15.2.1). Note that since here we are in a randomized trial, the propensity is known and hence the rate requirements in that theorem are satisfied. We see that all meta-learners perform better than a constant effect, indicating statistically significant heterogeneity. Moreover, we see that all learners except the S-learner have comparable performance.

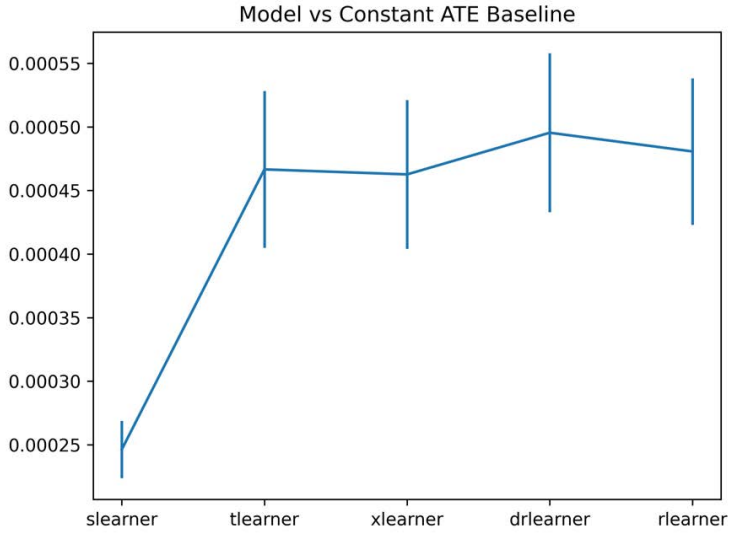


Figure 15.33: Performance (and 95% confidence intervals) of meta-learner models in the Criteo example compared to a constant effect model, as measured by the Doubly Robust score.

We also compared the meta-learning approach to the Best-Linear-Predictor approach presented in the prior chapter on heterogeneous treatment effects. Instead of learning a CATE model using all the features, we fitted the best linear CATE when using only feature 'f3' or a second degree polynomial of that feature. We find that this BLP approach in this setting is quite un-stable due to poor extrapolation behavior. In particular, the feature 'f3' has very heavy negative tails (see Figure 15.35). The different parametric models overfit the parametric curve to the region of high density and extrapolate very poorly in the heavy negative tail. On the contrary the Q-aggregation ensemble

of the meta-learning models is more stable and regularizes appropriately in this regime.

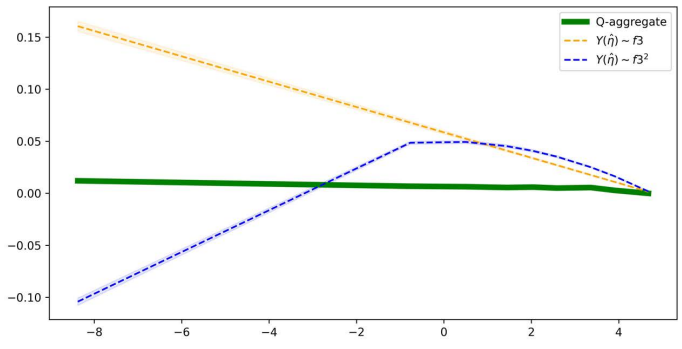


Figure 15.34: Predictions of the Q-aggregation stacked ensemble and of the Doubly Robust BLP of CATE as a linear or quadratic function of feature 'f3' in Criteo example.

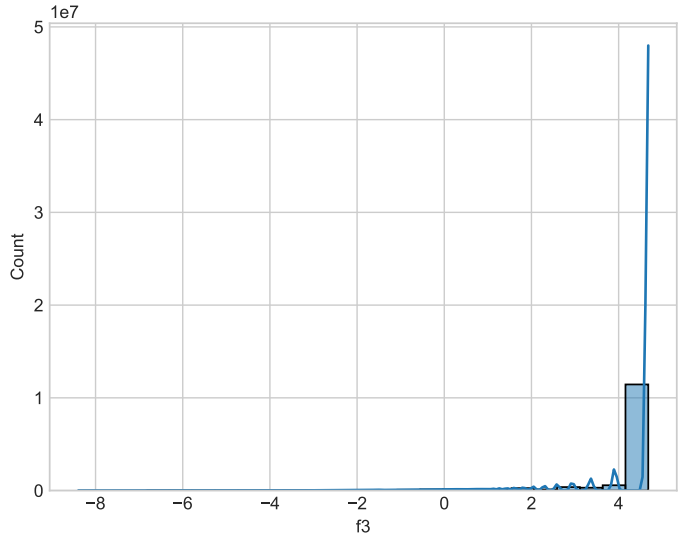


Figure 15.35: Histogram of distribution of feature 'f3'

We then validate the Q-aggregation ensemble using all the validation methods presented in this chapter. In Table 15.36 we run an OLS regression of the doubly robust pseudo-outcome on the CATE predictions and an intercept. We find that the coefficient of the CATE predictor is very accurately estimated to be 1.

	coef	**std err**	**P> |z|**	**[0.025**	**0.975]**
const	0.0074	0.000	0.000	0.007	0.008
$\tau_*(X)$	1.0096	0.036	0.000	0.940	1.079

Figure 15.36: OLS statistical test regression $Y(\hat{\eta})$ on $(1, \tau_*(X))$ in the Criteo example. Standard Errors are heteroscedasticity robust (HC1). τ_* corresponds to the stacked ensemble based on Q-aggregation.

We also see that the predictions of the CATE model are very well calibrated and the doubly robust GATE estimates for each

quartile of CATE prediction groups lies almost on the 45 degree line, with a calibration score that is very close to 1.

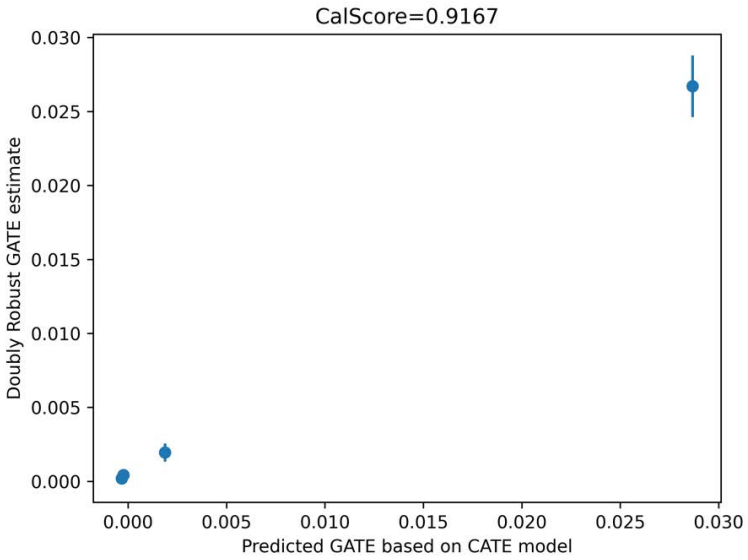


Figure 15.37: Calibration check for chosen ensemble model τ_* in the welfare example. Test samples are splitted in four groups based on CATE predictions and CATE quantiles (e.g. bottom group contains samples whose CATE predictions lie in the bottoms 25% of predictions). The x-axis depicts the average predicted CATE within each group based on τ_*, while the y-axis depicts the GATE as calculated based on the doubly robust pseudo-outcomes calculated on the test set.

The TOC and Qini Curves are depicted in Figure 15.38 and Figure 15.39, with one-sided 95% uniform confidence bands. We see that there is statistically significant heterogeneity as these curves lie well above the zero line. For instance, the TOC curve tells as that if we treat roughly the group that corresponds to roughly the top 5% of CATE predictions then we should expect that the average treatment effect of that group to be approximately 0.07 larger than the average treatment effect. Moreover, the Qini curve tells us that if we treat approximately the group that corresponds to the top 15% of CATE predictions, then we should expect the total effect of such a treatment policy to be ≈ 0.005 larger than the total effect if we were to treat a random 15% subset of the population. Thus our CATE model carries significant information that is valuable for better ad targeting.

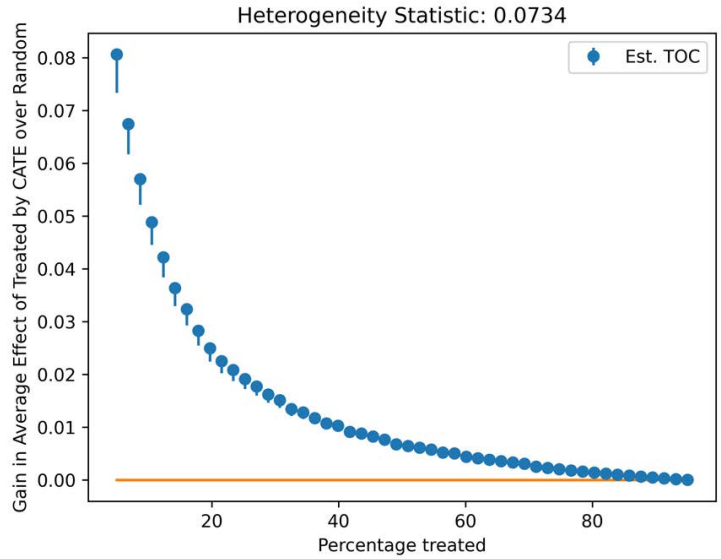


Figure 15.38: TOC curve with a 95% one-sided confidence band for the digital advertising dataset.

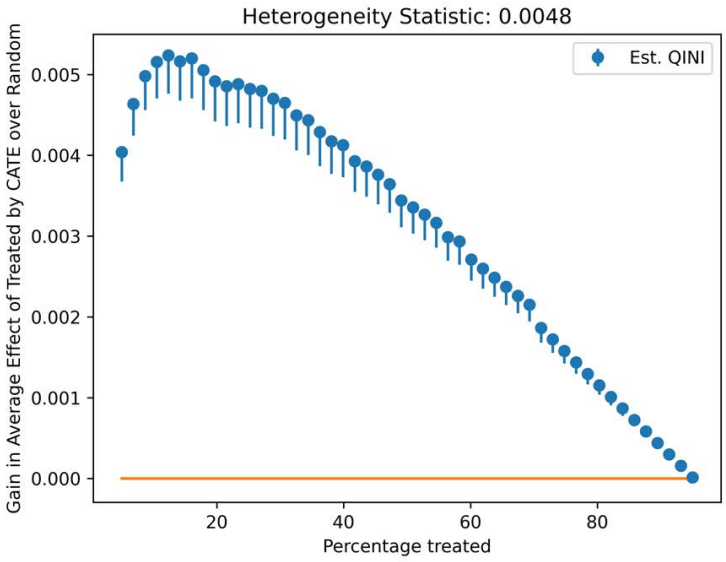


Figure 15.39: QINI curve with a 95% one-sided confidence band for the digital advertising dataset.

Given that in this setting we primarily care about personalized ad targeting, we can also apply the direct policy learning methodologies and learn an interpretable decision tree policy of who to target. We assume here that the cost of treatment is 0.04 (which can be interpreted as the cost of ad display divided by the average value of a webpage visit) and learn a binary decision tree that dictates who should be treated. The learned policy tree is depicted in Figure 15.40. We see that the method learned that we should not be treating visitors with a high value of the feature 'f8' and we should definitely be treating visitors with a low value for both the features 'f8' and 'f3'. For visitors with a small 'f8' and large 'f3' the model is rather indifferent

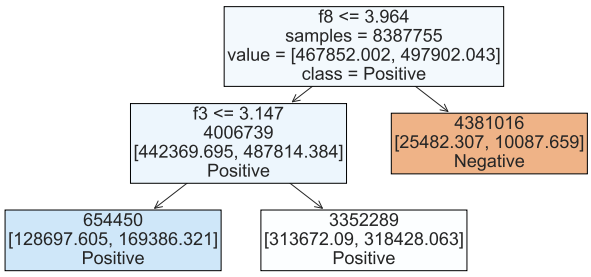


Figure 15.40: The details that are displayed on each node are also useful in understanding the group average treatment effect for each node. In particular, the information 'samples=N', gives us the size of each node N, and the information 'value=[A, B]', then 'A' is the sum of the $|Y^{DR}(g,p)|$ for the samples where $Y^{DR}(g,p) < 0$ and similarly, 'B' is the sum of $|Y^{DR}(g,p)|$ for the samples where $Y^{DR}(g,p) > 0$. Thus to get the GATE for each node, we simply do '(B-A)/N', which would correspond to $\frac{1}{N}\sum_{i\in\text{node}} Y^{DR}(g,p)$, which is the doubly robust estimate of the GATE for the node.

and weakly recommends treatment.

We can evaluate the performance of the learned policy out of sample using the policy evaluation method presented in Section 14.3. We find that the value of the learned policy is 0.00669 with standard error 0.00029 and 95% confidence interval $[0.00613, 0.00725]$. On the contrary if we were to treat everyone in the population, i.e. display an ad to everyone, then we would get a policy value of 0.00304 with standard error 0.00029 and 95% confidence interval $[0.00248, 0.00360]$. This showcases again the large benefits of personalized policy learning, which yields almost a double net profit.

Notebooks

- Python Notebook for CATE analyzes CATE of welfare experiment and criteo experiment and for the 401k dataset with generic machine learning.

15.A Appendix: Lower Bound on Variance in Model Comparison

First we observe that:

$$\begin{aligned}\Delta_{i,j} &:= (Y(\eta_0) - \tau_i(X))^2 - (Y(\eta_0) - \tau_j(X))^2 \\ &= \tau_i(X)^2 - \tau_j(X)^2 - 2Y(\eta_0)(\tau_i(X) - \tau_j(X)) \\ &= (\tau_i(X) - \tau_j(X))\,(\tau_i(X) + \tau_j(X) - 2Y(\eta_0))\end{aligned}$$

and note that:

$$\mathsf{V}_n = \mathrm{E}\Delta_{i,j}^2 - \left(\mathrm{E}\Delta_{i,j}\right)^2$$

Moreover,

$$\mathrm{E}\Delta_{i,j}^2 = \mathrm{E}\left[(\tau_i(X) - \tau_j(X))^2\, \mathrm{E}\left[(\tau_i(X) + \tau_j(X) - 2Y(\eta_0))^2 \mid X\right]\right]$$

By a variance decomposition argument and since $\tau_0(X) = \mathrm{E}(Y(\eta_0) \mid X)$:

$$\begin{aligned}&\mathrm{E}\left[(\tau_i(X) + \tau_j(X) - 2Y(\eta_0))^2 \mid X\right] \\ &\quad = 4\,\mathrm{Var}(Y(\eta_0) \mid X) + \mathrm{E}\left[(\tau_i(X) + \tau_j(X) - 2\tau_0(X))^2 \mid X\right]\end{aligned}$$

Thus we have derived that:

$$\begin{aligned}\mathrm{E}\Delta_{i,j}^2 = 4\mathrm{E}&\left[(\tau_i(X) - \tau_j(X))^2\ \mathrm{Var}(Y(\eta_0) \mid X)\right] \\ &+ \mathrm{E}\left[(\tau_i(X) - \tau_j(X))^2\,(\tau_i(X) + \tau_j(X) - 2\tau_0(X))^2\right]\end{aligned}$$

Moreover, note that by Jensen's inequality:

$$\begin{aligned}\left(\mathrm{E}\Delta_{i,j}\right)^2 &= \left(\mathrm{E}(\tau_i(X) - \tau_j(X))\,(\tau_i(X) + \tau_j(X) - 2\tau_0(X))\right)^2 \\ &\leq \mathrm{E}(\tau_i(X) - \tau_j(X))^2\,(\tau_i(X) + \tau_j(X) - 2\tau_0(X))^2\end{aligned}$$

Thus we can conclude that:

$$\mathsf{V}_n \geq 4\mathrm{E}\left[(\tau_i(X) - \tau_j(X))^2\ \mathrm{Var}(Y(\eta_0) \mid X)\right] \tag{15.A.1}$$

15.B Appendix: Interpretation of Uplift curves

We derive first the covariance interpretation of the TOC uplift curve.

$$\begin{aligned}\mathrm{TOC}(q) &= \mathrm{E}[Y(1) - Y(0) \mid \hat{\tau}(X) \geq \mu(q)] - \mathrm{E}[Y(1) - Y(0)] \\ &= \mathrm{E}\left[(Y(1) - Y(0))\frac{1\{\hat{\tau}(X) \geq \mu(q)\}}{\mathrm{P}(\hat{\tau}(X) \geq \mu(q))}\right] - \mathrm{E}[Y(1) - Y(0)]\end{aligned}$$

Let $A = Y(1) - Y(0)$ and $B = \frac{1\{\hat{\tau}(X) \geq \mu(q)\}}{\mathrm{P}(\hat{\tau}(X) \geq \mu(q))}$ and note that $\mathrm{E}[B] = 1$. Thus we have:

$$\mathrm{TOC}(q) = \mathrm{E}\left[A\,B\right] - \mathrm{E}[A] = \mathrm{E}\left[A\,B\right] - \mathrm{E}[A]\,\mathrm{E}[B] = \mathrm{Cov}(A, B)$$

Next, we derive the covariance interpretation of the QINI uplift curve. Let $A = Y(1) - Y(0)$ and $B = \hat{\tau}(Z) \geq \mu(q)$. Then by the

definition of the QINI curve:

$$
\begin{aligned}
\tau_{\text{QINI}}(q) &:= \tau(q)\,\mathrm{P}(\hat{\tau}(Z) \geq \hat{\mu}(q)) \\
&= \left(\mathrm{E}[A \mid B] - \mathrm{E}[A]\right)\,\mathrm{P}(B) \\
&= \left(\mathrm{E}[A\frac{\mathbb{1}\{B\}}{\mathrm{P}(B)}] - \mathrm{E}[A]\right)\,\mathrm{P}(B) \\
&= \mathrm{E}[A\mathbb{1}\{B\}] - \mathrm{E}[A]\mathrm{E}[\mathbb{1}\{B\}] \;=\; \mathrm{Cov}\left(A, \mathbb{1}\{B\}\right)
\end{aligned}
$$

Bibliography

[1] Arthur Conan Doyle. *The sign of four*. Spencer Blackett, 1890 (cited on page 386).

[2] Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. 'Regularised orthogonal machine learning for nonlinear semiparametric models'. In: *The Econometrics Journal* 25.1 (2022), pp. 233–255 (cited on page 390).

[3] Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. 'Orthogonal random forest for causal inference'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4932–4941 (cited on page 390).

[4] Dylan J Foster and Vasilis Syrgkanis. 'Orthogonal statistical learning'. In: *The Annals of Statistics* 51.3 (2023), pp. 879–908 (cited on pages 390, 392, 412, 432).

[5] Edward H Kennedy. 'Towards optimal doubly robust estimation of heterogeneous causal effects'. In: *arXiv preprint arXiv:2004.14497* (2020) (cited on page 390).

[6] Greg Lewis and Vasilis Syrgkanis. 'Double/debiased machine learning for dynamic treatment effects via g-estimation'. In: *arXiv preprint arXiv:2002.07285* (2020) (cited on pages 390, 392).

[7] Xinkun Nie and Stefan Wager. 'Quasi-oracle estimation of heterogeneous treatment effects'. In: *Biometrika* 108.2 (2021), pp. 299–319 (cited on page 392).

[8] Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. 'Machine learning estimation of heterogeneous treatment effects with instruments'. In: *Advances in Neural Information Processing Systems* 32 (2019) (cited on page 399).

[9] Scott M Lundberg and Su-In Lee. 'A Unified Approach to Interpreting Model Predictions'. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 4765–4774 (cited on page 400).

[10] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020 (cited on page 400).

[11] Uri Shalit, Fredrik D Johansson, and David Sontag. 'Estimating individual treatment effect: generalization bounds and algorithms'. In: *International conference on machine learning*. PMLR. 2017, pp. 3076–3085 (cited on pages 403, 404).

[12] Fredrik Johansson, Uri Shalit, and David Sontag. 'Learning representations for counterfactual inference'. In: *International conference on machine learning*. PMLR. 2016, pp. 3020–3029 (cited on pages 403, 404).

[13] Alicia Curth and Mihaela van der Schaar. 'On inductive biases for heterogeneous treatment effect estimation'. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15883–15894 (cited on page 404).

[14] Kevin Wu Han and Han Wu. 'Ensemble Method for Estimating Individualized Treatment Effects'. In: *arXiv preprint arXiv:2202.12445* (2022) (cited on page 412).

[15] Guillaume Lecué and Philippe Rigollet. 'Optimal learning with Q-aggregation'. In: *The Annals of Statistics* 42.1 (2014), pp. 211–224 (cited on page 413).

[16] Hui Lan and Vasilis Syrgkanis. 'Causal Q-Aggregation for CATE Model Selection'. In: *arXiv preprint arXiv:2310.16945* (2023) (cited on page 413).

[17] Lars van der Laan, Ernesto Ulloa-Pérez, Marco Carone, and Alex Luedtke. 'Causal isotonic calibration for heterogeneous treatment effects'. In: *arXiv preprint arXiv:2302.14011* (2023) (cited on page 419).

[18] Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. 'Distribution-free binary classification: prediction sets, confidence intervals and calibration'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3711–3723 (cited on page 419).

[19] Chirag Gupta and Aaditya Ramdas. 'Distribution-free calibration guarantees for histogram binning without sample splitting'. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 3942–3952 (cited on page 419).

[20] Raaz Dwivedi, Yan Shuo Tan, Briton Park, Mian Wei, Kevin Horgan, David Madigan, and Bin Yu. 'Stable discovery of interpretable subgroups via calibration in causal studies'. In: *International Statistical Review* 88 (2020), S135–S178 (cited on page 421).

[21] Nicholas Radcliffe. 'Using control groups to target on predicted lift: Building and assessing uplift model'. In: *Direct Marketing Analytics Journal* (2007), pp. 14–21 (cited on page 424).

[22] Nicholas J Radcliffe and Patrick D Surry. 'Real-world uplift modelling with significance-based uplift trees'. In: *White Paper TR-2011-1, Stochastic Solutions* (2011), pp. 1–33 (cited on page 424).

[23] Patrick D Surry and Nicholas J Radcliffe. 'Quality measures for uplift models'. In: *submitted to KDD2011* (2011) (cited on page 424).

[24] Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. 'Evaluating treatment prioritization rules via rank-weighted average treatment effects'. In: *arXiv preprint arXiv:2111.07966* (2021) (cited on page 430).

[25] Susan Athey and Stefan Wager. 'Policy learning with observational data'. In: *Econometrica* 89.1 (2021), pp. 133–161 (cited on page 432).

[26] Victor Chernozhukov, Mert Demirer, Greg Lewis, and Vasilis Syrgkanis. 'Semi-parametric efficient policy learning with continuous actions'. In: *Advances in Neural Information Processing Systems* 32 (2019) (cited on pages 432, 433).

[27] John C Duchi, Peter W Glynn, and Hongseok Namkoong. 'Statistics of robust optimization: A generalized empirical likelihood approach'. In: *Mathematics of Operations Research* 46.3 (2021), pp. 946–969 (cited on page 433).

[28] Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. 'Policy learning" without"overlap: Pessimism and generalized empirical Bernstein's inequality'. In: *arXiv preprint arXiv:2212.09900* (2022) (cited on page 433).

[29] Dylan J Foster and Alexander Rakhlin. 'Foundations of Reinforcement Learning and Interactive Decision Making'. In: *arXiv preprint arXiv:2312.16730* (2023) (cited on page 434).

[30] Eustache Diemert, Artem Betlei, Christophe Renaudin, Massih-Reza Amini, Théophane Gregoir, and Thibaud Rahier. 'A large scale benchmark for individual treatment effect prediction and uplift modeling'. In: *arXiv preprint arXiv:2111.10106* (2021) (cited on page 440).