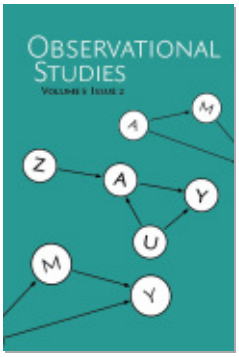
Machine Learning Analysis of Heterogeneity in the Effect of Student Mindset Interventions

Machine Learning Analysis of Heterogeneity in the Effect of Student Mindset Interventions



Fredrik D. Johansson

Observational Studies, Volume 5, Issue 2, 2019, pp. 71-82 (Article)

Published by University of Pennsylvania Press
DOI: *https://doi.org/10.1353/obs.2019.0003*

Observational Studies 5 (2019) 71-82 Submitted 7/19; Published 8/19

Machine Learning Analysis of Heterogeneity in the Effect of Student Mindset Interventions

Fredrik D. Johansson **fredrikj@mit.edu**
Institute for Medical Engineering & Science
Massachusetts Institute of Technology

Abstract

We study heterogeneity in the effect of a mindset intervention on student-level performance through an observational dataset from the National Study of Learning Mindsets (NSLM). Our analysis uses machine learning (ML) to address the following associated problems: assessing treatment group overlap and covariate balance, imputing conditional average treatment effects, and interpreting imputed effects. By comparing several different model families we illustrate the flexibility of both off-the-shelf and purpose-built estimators. We find that the mindset intervention has a positive average effect of 0.26, 95%-CI $[0.22, 0.30]$, and that heterogeneity in the range of $[0.1, 0.4]$ is moderated by school-level achievement level, poverty concentration, urbanicity, and student prior expectations.

Keywords: Machine learning, interpretability, counterfactual estimation

1. Methodology and Motivation

Machine learning (ML) has had widespread success in solving prediction problems in applications ranging from image and speech recognition (LeCun et al., 2015) to personalized medicine (Kononenko, 2001). This makes it an attractive tool also for studying heterogeneity in causal effects. In fact, ML excels at overcoming well-known limitations of traditional methods used to solve this task. For example, matching methods struggle to perform well when confounders and covariates are high-dimensional (Rubin and Thomas, 1996); generalized linear models are not flexible enough to discover variable interactions and non-linear trends; and propensity-based methods suffer from variance issues in estimation (Lee et al., 2011). In contrast, supervised machine learning has proven highly useful in discovering patterns in high-dimensional data (LeCun et al., 2015), approximating complex functions and trading off bias and variance (Swaminathan and Joachims, 2015).

In this observational study based on data from the National Study of Learning Mindsets (NSLM), we apply both off-the-shelf and purpose-built ML estimators to characterize heterogeneity in the effect of a student mindset intervention on future academic performance. In particular, we compare estimates of conditional average treatment effects based on linear models, random forests, gradient boosting and deep neural networks. Below, we introduce the problem of discovering treatment effect heterogeneity and describe our methodology.

1.1 Problem setup

We study the effect of a student mindset intervention based on observations of 10391 students in 76 schools from the NSLM study. The intervention, assigned at student level, is

represented by a binary variable $Z \in \{0, 1\}$ and the performance outcome by a real-valued variable $Y \in \mathbb{R}$. Students are observed through covariates S_3, C_1, C_2, C_3 and schools through covariates $X_1, ..., X_5$[1]. For convenience, we let $X = [S_3, C_1, C_2, C_3, X_1, ..., X_5]^\top$ represent the full set of covariates of a student-school pair. We let (x_{ij}, z_i, y_i) denote the observation corresponding to a student $i \in \{1, ..., m\}$ in a school $j \in \{1, ..., n\}$. As each student is enrolled in at most one school, we ommit the index j in the sequel. Observed treatment groups G_0 (control) and G_1 (treated) are defined by $G_z = \{i \in \{1, ..., m\} : z_i = z\}$. The full dataset of observations is denoted $\mathcal{D} = \{(x_1, z_1, y_1), \ldots, (x_m, z_m, y_m)\}$, and the density of all variables $p(X, Z, Y)$.

We adopt the Neyman-Rubin causal model (Rubin, 2005), and denote by $Y(0), Y(1)$ the potential outcomes corresponding to interventions $Z = 0$ and $Z = 1$ respectively. The goal of this study is to characterize heterogeneity in the treatment effect $Y(1) - Y(0)$ across students and schools. As is well known, this effect is not identifiable without additional assumptions as each student is observed in only one treatment group. Instead, we estimate the *conditional average treatment effect* (CATE) with respect to observed covariates X.

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

CATE is identifiable from observational data under the standard assumptions of ignorability

$$Y(1), Y(0) \perp\!\!\!\perp Z \mid X \ ,$$

consistency, $Y = ZY(1) + (1 - Z)Y(0)$, and overlap (positivity)

$$\forall x : p(Z = 0 \mid X = x) > 0 \Leftrightarrow p(Z = 1 \mid X = x) > 0 \ .$$

The CATE conditioned on the full set of covariates X is the closest we get to estimating the treatment effect for an individual student. However, to design policies, it is rarely necessary to have this level of resolution. In later sections, we estimate conditional effects also with respect to subsets or functions of X, such as the average effect stratified by school achievement level. By first identifying $\tau(x)$ and then marginalizing it with respect to such functions, we adjust for confounding to the best of our ability.

1.2 Methodology overview

The flexibility of ML estimators creates both opportunities and challenges. For example, it is typical for the number of parameters of the best performing model on a task to exceed the number of available samples. This is made possible by mitigating overfitting (high variance) through appropriate regularization. Indeed, many models achieve the best fit only after having carefully set several tuning parameters that control the bias and variance trade-off. It is standard practice in ML to use sample splitting for this purpose. Here, we apply such a pipeline to CATE estimation, proceeding through the following steps.

1. Split the observed data $\mathcal{D}$ into two partitions, a training set $\mathcal{D}_t$ and a validation set $\mathcal{D}_v$, for parameter fitting and model selection respectively.

1. The meanings of the different covariates are described in later sections of the manuscript.

2. Fit estimators f_0, f_1 of potential outcomes $Y(0)$ and $Y(1)$ to $\mathcal{D}_t$ and select tuning parameters based on held-out error on $\mathcal{D}_v$

3. Impute CATE, $\hat{\tau}_i := f_1(x_i) - f_0(x_i)$, for every student in $\mathcal{D}$ and fit an *interpretable* model $h(x) \sim \hat{\tau}$ to characterize treatment effect heterogeneity

This pipeline allows us to find the best fitting (black-box) estimators possible in Step 2 without regard for the interpretability of their predictions. By fitting a simpler, more interpretable model to the imputed effects in Step 3, we may explain the predictions of the more complex model in terms of known quantities. This procedure is particularly well suited when the effect is a simpler function than the response and it also allows us to control the granularity at which we study heterogeneity.

The data from NSLM has a multi-level nature; students (level 1) are grouped into schools (level 2) and each level is associated with its own set of covariates. The literature is rich with studies of causal effects in multi-level settings, see for example Gelman and Hill (2006). However, this is primarily targeted towards studying the effects of high-level (e.g. school-level) interventions on lower-level subjects (e.g. students), and the increased uncertainty that comes with such an analysis. While interventions are assigned at student-level, it is important to note that only 76 values of school-level variables are observed, which introduces the risk of overfitting to these covariates specifically. We adjust for the multi-level nature of the data in sample splitting, bootstrapping and the analysis of imputed effects.

In the following sections we describe each step of our methodology in detail.

Step 1. Sample splitting

To enable unbiased estimation of prediction error and select tuning parameters, we divide the dataset $\mathcal{D}$ into two parts with 80% of the data used for the training set $\mathcal{D}_t$ and 20% for a validation set $\mathcal{D}_v$. We partition the set of schools, rather than students, making sure that the entire student body of any one school appears only in either $\mathcal{D}_t$ or $\mathcal{D}_v$. This is to mitigate overfitting to school-level covariates. As there are only 76 schools, random sampling may create sets that have very different characteristics. To mitigate this, we balance $\mathcal{D}_t$ and $\mathcal{D}_v$ by constructing a large number of splits uniformly at random and selecting the one that minimizes the Euclidean distance between summary statistics of the two sets. In particular, we compare the first and second order moments of all covariates. We increase the influence of the treatment variable Z by a factor 10 in this comparison to ensure that treatment groups are split evenly across $\mathcal{D}_t$ and $\mathcal{D}_v$.

Step 2. Estimation of potential outcomes

The conditional average treatment effect is the difference between expected potential outcomes, here denoted μ_0 and μ_1. Under ignorability w.r.t. X (see above), we have that

$$\mu_z(x) := \mathbb{E}[Y(z) \mid X = x] = \mathbb{E}[Y \mid X = x, Z = z] \quad \text{for} \quad z \in \{0, 1\},$$

and thus, $\tau(x) = \mu_1(x) - \mu_0(x)$. A straight-forward route to estimating τ is to independently fit the conditionals $\mathbb{E}[Y \mid X = x, Z = z]$ for each value of $z \in \{0, 1\}$ and compute their difference. This has recently been dubbed the *T-learner* approach to distinguish it from

other learning paradigms (Künzel et al., 2017). Below, we briefly cover theory that motivates this method and point out some of its shortcomings. To study heterogeneity, we consider several T-learners as well as two alternative approaches described below.

We approximate μ_0, μ_1 using hypotheses f_0, f_1 and measure their quality by the mean squared error. The group-conditional expected and empirical risks are defined as follows

$$\underbrace{\mathcal{R}_z(f_z) := \mathbb{E}[(\mu_z(x) - f_z(x))^2 \mid Z = z]}_{\text{Expected group-conditional risk}} \quad \text{and} \quad \underbrace{\hat{\mathcal{R}}_z(f_z) := \frac{1}{|G_z|} \sum_{i \in G_z} (f(x_i; \theta) - y_i)^2}_{\text{Empirical group-conditional risk}} . \quad (2)$$

We never observe μ_z directly, but learn from noisy observations y. Statistical learning theory helps resolve this issue by bounding the expected risk in terms of its empirical counterpart and a measure of function complexity (Vapnik, 1999). For hypotheses in a class $\mathcal{F}$ with a particular complexity measure $\mathcal{C}_{\mathcal{F}}(\delta, n)$ with logarithmic dependence on n (e.g. a function of the covering number), it holds with probability greater than $1 - \delta$ that

$$\forall f_z \in \mathcal{F} : \quad \mathcal{R}_z(f_z) \leq \hat{\mathcal{R}}_z(f_z) + \frac{\mathcal{C}_{\mathcal{F}}(\delta, n)}{\sqrt{n}} - \sigma_Y^2 , \quad (3)$$

where σ_Y^2 is a bound on the expected variance in Y (see Johansson et al. (2018) for a full derivation). This class of bounds illustrate the bias-variance trade-off that is typical for machine learning and motivates the use of regularization to control model complexity. In our experiments, we consider several T-learner models that estimate each potential outcome independently using regularized empirical risk minimization, solving the following problem.

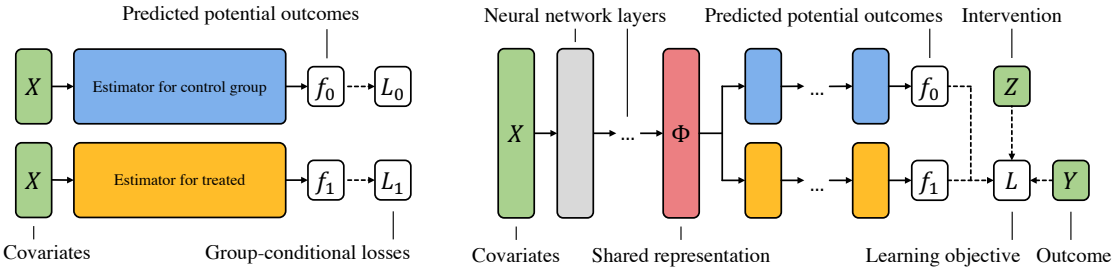
$$f_z = \underset{f(\cdot;\theta) \in \mathcal{F}}{\arg\min} \quad \hat{R}_z(f(x; \theta)) + \lambda r(\theta) \quad (4)$$

Here, $f(x; \theta)$ is a function parameterized by θ and $r(\theta)$ a regularizer of model parameters such as the ℓ_1-norm (LASSO) or ℓ_2-norm (Ridge) penalties. In our analysis, we compare four commonly used off-the-shelf machine learning estimators: ridge regression, random forests, gradient boosting and deep neural networks.

Sharing power between treatment groups A drawback of T-learners is that no information is shared between estimators of different potential outcomes. In problems where the baseline response $Y(0)$ is a more complex function than the effect τ itself, the T-learner is wasteful in terms of statistical power (Künzel et al., 2017; Nie and Wager, 2017). As an alternative, we apply the Treatment-Agnostic Representation (TARNet) neural network architecture of Shalit et al. (2017). TARNet estimates all potential outcomes $\{Y(z)\}$ jointly as compositions $f_z(x) := h_z(\Phi(x))$ of treatment-specific hypotheses $f_z(\Phi)$ and treatment-agnostic representations $\Phi(x)$. Trained by minimizing the overall empirical risk, as described in (4), this choice of architecture encourages sharing of information across treatment groups in learning the average response, while capturing heterogeneity in treatment effects. For an illustration comparing T-learners and TARNet, see Figure 1.

Generalizing across treatment groups

The careful reader may have noticed that the population and empirical risk in equations (2)–(4) are defined with respect to the observed treatment assignments. To estimate CATE, we



(a) T-learner estimator

(b) TARNet architecture (Shalit et al., 2017).

Figure 1: Illustration of T-learner estimator (left) and Treatment-Agnostic Representation (TARNet) architecture (Shalit et al., 2017). TARNet estimators learn representations of covariates $\Phi(x)$ that are *shared* between treatment groups and model different potential outcomes $Y(z)$ as functions of Φ. Counterfactual Regression (CFR) (Shalit et al., 2017) extends TARNet by regularizing Φ to encourage *balance* across different treatment groups.

want our estimates of potential outcomes to be accurate for the counterfactual assignment as well. In other words, we want for the risk on the full cohort,

$$R(f_z) := \mathbb{E}[(\mu_z(x) - f_z(x))^2]$$

to be small. When treatment groups $p(X \mid Z = 0)$ and $p(X \mid Z = 1)$ are highly imbalanced, the expected risk within one treatment group may not be representative of the risk on the full population. This is another drawback of T-learner estimators, which do not adjust for this discrepancy.

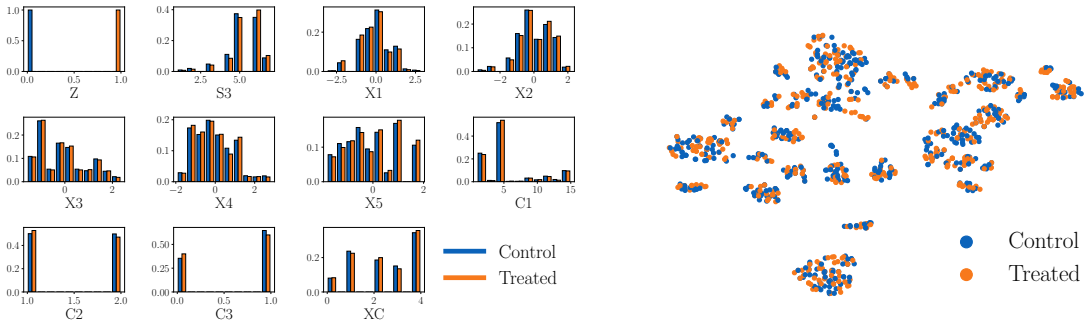
In recent work, Shalit et al. (2017) characterize the difference between $R(f_z)$ and $R_z(f_z)$ and bound the error in CATE estimates using a distance metric between treatment groups. In particular, they considered the *integral probability metric* (IPM) family of distances, defined with respect to a function family $\mathcal{G}$ and densities p, q as

$$\text{IPM}_{\mathcal{G}}(p, q) := \sup_{g \in \mathcal{G}} \left| \int_x g(x)(p(x) - q(x))dx \right| ,$$

resulting in the following relation between population and treatment group risk

$$\underbrace{R(f_z)}_{\text{Population risk}} \leq \underbrace{R_z(f_z)}_{\text{Treatment group risk}} + \underbrace{\text{IPM}_{\mathcal{G}}(p(X \mid Z = 0), p(X \mid Z = 1))}_{\text{Treatment group imbalance}} , \tag{5}$$

under appropriate assumptions. This bound inspired the estimator Counterfactual Regression (CFR) in which the TARNet architecture (see above) is trained to minimize the upper bound in (5) applied to the learned representation ϕ, instead of the empirical risk. This encourages *balance* between treatment groups in the learned representation space. In our analysis, we apply CFR with $\mathcal{G}$ the family of functions in the reproducing-kernel Hilbert space defined by the Gaussian RBF-kernel; the resulting IPM is known as the Maximum Mean Discrepancy (Gretton et al., 2012) and may be estimated efficiently from samples.



(a) Marginals of covariates X (b) t-SNE projection of covariates X

Figure 2: Examination of overlap through covariate marginal distributions and a low-dimensional t-SNE projection of covariates (Maaten and Hinton, 2008). Marker color corresponds to treatment assignment Z. Best viewed in color.

Step 3. Characterization of heterogeneity in CATE

After fitting models f_0, f_1 for each potential outcome, the conditional average treatment effect is imputed for each student by $\hat{\tau}_i = f_1(x_i) - f_0(x_i)$. Unlike with linear regressors, the predictions of most ML estimators are difficult to interpret directly through model parameters. For this reason, ML models are often considered *black-box* methods (Lipton, 2016). However, in the study of heterogeneity, it is crucial to characterize for which subjects the effect of an intervention is low and for which it is high. To accomplish this, we adopt the common practice of post-hoc interpretation—fitting a simpler, more interpretable model $h \in \mathcal{H}$ to the imputed effects $\{\hat{\tau}_i\}$.

In its very simplest form $h(x_i)$ may be a function of a single attribute, such as the school size, effectively averaging over other attributes. This is usually a good way of discovering global trends in the data but will neglect meaningful interactions between variables, much like a linear model. As a more flexible alternative, we also fit decision tree models and inspect the learned trees. Trees of only two variables may be visualized directly in the covariate space, and larger trees in terms of their decision rules.

2. Workshop results

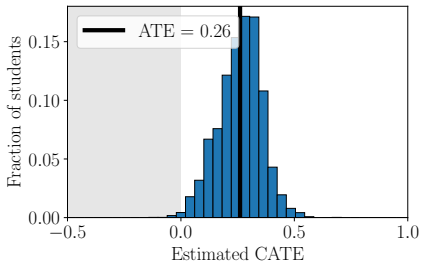
We present the first results of our analysis as shown in the workshop Empirical Investigation of Methods for Heterogeneity at the Atlantic Causal Inference Conference, 2018.

2.1 Covariate balance

First, we investigate the extent to which the overlap assumption holds true by comparing the covariate statistics of the treatment and control groups. In Figure 2, we visualize the marginal distributions of each covariate, as well as a 2D t-SNE projection of the entire covariate set (Maaten and Hinton, 2008). The observed difference between the marginal covariate distributions of the two treatment groups is very small. Also the non-linear t-SNE projection reveals little difference between treatment groups. The less imbalance between

Estimator	$\hat{\text{ATE}}$		R^2	
Naïve	0.30		—	
RR	0.26	$[0.22, 0.29]$	0.26	$[0.17, 0.29]$
RF	0.27	$[0.23, 0.30]$	0.25	$[0.22, 0.29]$
GB	0.26	$[0.21, 0.30]$	0.25	$[0.20, 0.30]$
NN	0.27	$[0.17, 0.38]$	0.14	$[-0.08, 0.23]$
TARNet	0.26	$[0.23, 0.30]$	0.27	$[0.21, 0.31]$
CFR	0.26	$[0.22, 0.30]$	0.27	$[0.22, 0.31]$

(a) ATE and held-out R^2 score with 95% school-level cluster bootstrap confidence intervals.



(b) Histogram of CATE (CFR).

Figure 3: Comparison between the naïve estimator, T-learners, and representation learning methods (left). Heterogeneity in treatment effect estimated by CFR (right).

treatment groups, the closer our problem is to standard supervised learning. Said differently, the density ratio $p(Z = 1 \mid X)/p(Z = 0 \mid X)$ is close to 1.0 and the IPM distance between conditional distributions, see (5), is small. Hence, we expect neither propensity re-weighting nor balanced representations (e.g. CFR) to have a large effect on the results.

2.2 Estimation of potential outcomes

In our analysis, we compare four T-learners based on ridge regression (RR), random forests (RF), gradient boosting (GB), and neural networks (NN). In addition, we compare the representation-learning algorithms TARNet and Counterfactual Regression (CFR) (Shalit et al., 2017). For each estimator family, we fit models of both potential outcomes on the training set $\mathcal{D}_t$ and select tuning parameters based on the held-out R^2 score on the validation set $\mathcal{D}_v$. To estimate uncertainty in model predictions, we perform school-level bootstrapping of the training set (Cameron et al., 2008), fitting each model to each bootstrap sample[2]. In Table 3, we give the estimate of the average treatment effect (ATE) from each model, the held-out R^2 score of the fit of factual outcomes, and 95% confidence intervals based on the empirical bootstrap. In addition, we give the naïve estimate of the ATE—the difference between observed average outcomes in the two treatment groups.

We see that all methods produce very similar estimates of ATE and perform comparably in terms of R^2. As expected, based on the small covariate imbalance shown in the previous section, the regression adjusted estimates are close to the naïve estimate of the ATE. This likely also explains the small difference between TARNet and CFR, as even for moderate to large imbalance regularization, the empirical risk dominates the objective function. The performance of the neural network T-learner would likely be improved with a different choice of architecture or tuning parameters. This is consistent with Shalit et al. (2017) in which TARNet architecture achieved half of the error of the T-learner on the IHDP benchmark.

2. The bootstrap analysis was added after the workshop results, but is presented here for completeness.

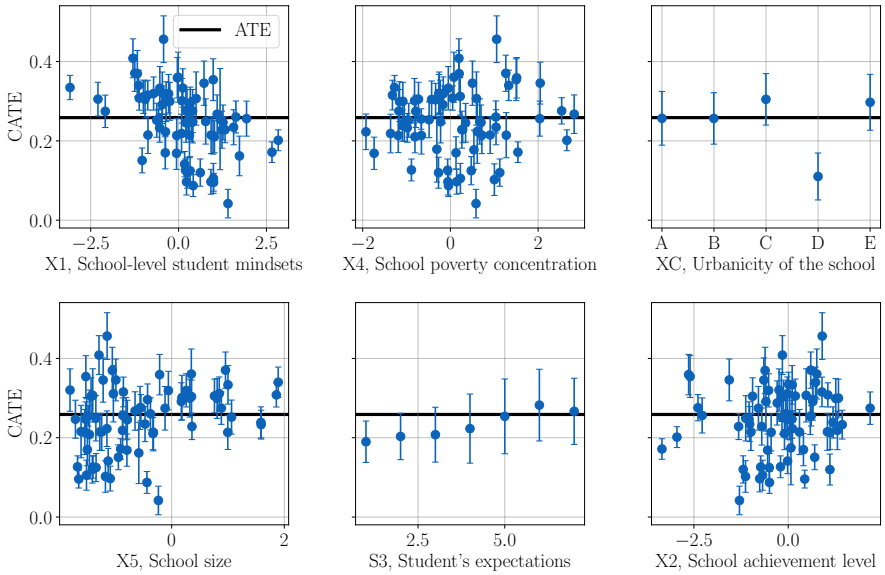
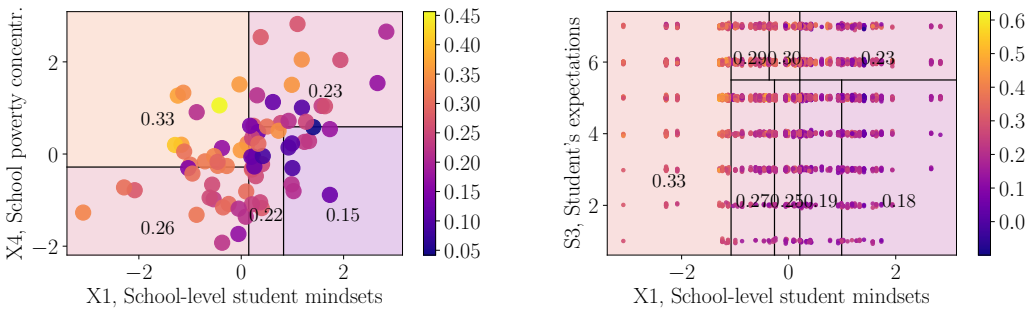


Figure 4: Heterogeneity in causal effect estimated using counterfactual regression (CFR) stratified by different covariates. Bars indicate variation in point estimates across subjects.

2.3 Heterogeneity in causal effect

We examine further the CATE for each student imputed by the best fitting model. As CFR had a slight edge in R^2 over T-learning estimators (although confidence intervals overlap) and has stronger theoretical justification, we analyze the effects imputed by CFR below. In Figure 3b, we visualize the distribution of imputed CATEs. We see that for almost all students, the effect is estimated to be positive, indicating an improvement in performance as an effect of the mindset intervention. Recall that the average treatment effect was estimated to be 0.26. Around 95% of students were estimated to have an effect in the range $[0.05, 0.45]$. For reference, the mean of the observed outcome was 0.10 and the standard deviation 0.64.

To discover drivers of heterogeneity we fit a random forest estimator to imputed effects and inspect the feature importance of each variable—the frequency with which it is used to split the nodes of a tree. The five most important variables of the random forest were X_1, X_4, X_C, X_5 and S_3. In Figure 4 we stratify imputed CATE with respect to these variables, as well as X_2 which is of interest to the study organizers. We see a strong trend that the effect of the intervention decreases with prior school-level student mindset, X_1. The urbanicity of the school, X_C, is a categorical variable for which Category D appears to be associated with substantially lower effect. In contrast, the effect of the intervention appears to increase with students' prior expectations S_3. One of the questions of the original study was whether there exists a "Goldilocks effect" for school-achievement level X_2, meaning that the intervention only has an effect for schools that are neither achieving too poorly nor too well. These results cannot confirm this hypothesis, nor reject it.



(a) CATE vs. X1 and X4 (school-level) (b) CATE vs. X1 and S3 (student-level)

Figure 5: Intrepretation of CATE estimates using regression trees fit to pairs of covariates. Each dot represents a single school (left) or student (right). The color represents the predicted CATE. Black lines correspond to leaf boundaries. Background color and numbers in boxes correspond to the average predicted CATE in that box. Best viewed in color.

3. Post-workshop analysis

Heterogeneity in treatment effect may be a non-linear or non-additive function of observed covariates. Such patterns remain hidden when analyzing CATE as a function of a single variable at a time or using linear regression. To reveal richer patterns of heterogeneity, we fit highly regularized regression tree models and inspect their decision rules. First, we consider combinations only of pairs of variables at a time. We note that for school-level variables, only 76 unique values exist, one for each school. To prevent overfitting to these variables, we require that each leaf in the regression tree contains samples from at least 10 schools. When student-level covariates are included, we require leaves to have samples of at least 1000 students.

In Figure 5, we visualize trees fit to two distinct variable pairs. We note a very slight non-linear pattern in heterogeneity as a function of X_1 (school-level student mindset) and X_4 (school poverty concentration), and that X_1 explains a lot of the variance observed at moderate values of X_4 in Figure 4. We emphasize, however, that the sample size at the school-level is small, and that observed patterns have high variance. In the right-hand figure, S_3 (student's expectations) appears associated with a larger effect only if the average mindset of the school is sufficiently high. This pattern disappears when using a linear model. In the Appendix, we show a regression tree fit to the entire covariate set.

4. Discussion

Machine learning offers a broad range of tools for flexible function approximation and provides theoretical guarantees for statistical estimation under model misspecification. This makes it a suitable framework for estimation of causal effects from non-linear, imbalanced or high-dimensional observational data. The flexibility of machine learning comes at a price however: many methods come with tuning parameters that are challenging to set for causal estimation; models are often difficult to optimize globally; and interpretability of

models suffers. While progress has been made independently on each of these problems, a standardized set of tools has yet to emerge.

In the analysis of the NLSM data, machine learning appears well-suited to study overlap, potential outcomes and heterogeneity in imputed effects. However, the analysis also opens some methodological questions. The multi-level nature of covariates is not accounted for in most off-the-shelf ML models and regularization of models applied to multi-level data has been comparatively less studied than for single-level data. In addition, as pointed out by several authors (Künzel et al., 2017; Nie and Wager, 2017), the T-learner approach to causal effect estimation may suffer from compounding bias and from wasting statistical power. This may be one of the reasons we observe a slight advantage of representation learning methods such as TARNet and CFR.

References

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Johansson, F. D., Kallus, N., Shalit, U., and Sontag, D. (2018). Learning weighted representations for generalization across designs. *arXiv:1802.08598*.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2017). Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv:1706.03461*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436.

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174.

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv:1606.03490*.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Nie, X. and Wager, S. (2017). Learning objectives for treatment effect estimation. *arXiv:1712.04912*.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085.

Swaminathan, A. and Joachims, T. (2015). Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999.

Appendix A. Regression tree explanation of CATE

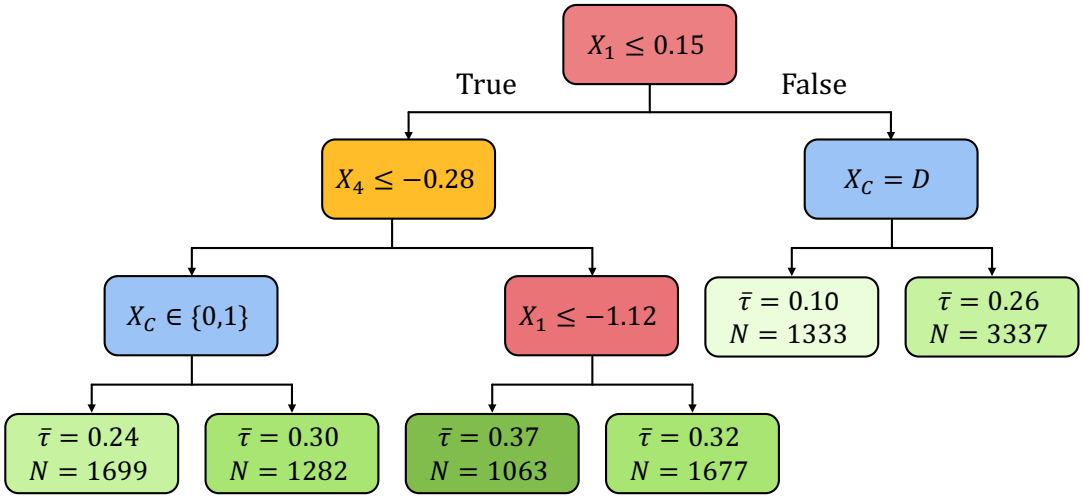


Figure 6: Visualization of a regression tree fit to the imputed CATE values based on the full covariate set.