Machine Learning - Final Project

Harsh Gupta, Jacob Light, Arjun Ramani, Andres Rodriguez

June 5, 2020

1 Introduction

One key advantage of using Machine Learning (ML) for causal inference is the ability to model flexible functional forms to estimate propensity scores, conditional means, and heterogeneous treatment effects. This flexibility reduces a researcher's dependence on parametric and functional form assumptions. However, this often comes at the cost of efficiency and interpretabiliy. Specifically, by using flexible ML methods, the researcher may not be efficiently employing outside-model information at his/her disposal. For example, the researcher may not be using pertinent information about the design of the experiment to discipline his/her analyze. The fundamental question we are interested in is whether, given enough data (both in terms of sample size and number of covariates), certain ML methods can produce results that are in line with results that are produced using more rigid specifications that are derived using outside-model information. We analyze this question in the context of one specific experiment.

Our experimental setting is the Partnership Schools for Liberia (PSL) program which was analyzed in "Outsourcing Education: Experimental Evidence from Liberia" by Mauricio Romero, Justin Sandefur, and Wayne Aaron Sandholtz (Romero, Sandefur, and Sandholtz (2020)). In the experiment, eight private providers took over the management of randomly chosen public schools. We describe the experiment in detail in section 2.

The experiment was conducted using a matched-pair design. Each provider submitted a list of facility requirements and based on those requirements, schools were assigned to providers. Once schools were assigned to providers, they were then paired by matching on the "school facilities" variable. Within each pair (also referred to as "group") one school was randomly chosen to be treated.

Even though the above experiment was an RCT, the design of the experiment provides us with critical information. First, different providers choose different pairs of schools to operate in and provided different types of services. Second, schools were not assigned to different groups or strata randomly but based on school facilities, which (presumably) affects outcomes. In this project, we analyze whether certain ML methods can provide accurate estimates if these two pieces of information (identifiers for providers and the fact that the experiment was a matched-pair design) are missing.

In Romero et al. (2020), the authors use a Weighted Ordinary Least Squares (OLS) regression with group fixed-effects (Wooldridge, 2016) to estimate the ATE for the entire sample. They use group-fixed effects to ensure that they take the matched-pair design of the experiment into account. We examine whether the non-parametric methods of Difference-in-means (ATE), Inverse Probability Weighting (IPW), and Augmented Inverse Probability Weighting (AIPW) can produce results that are in line with those of Romero et al. (2020) with and without group identifiers. First, we employ these ML methods with group identifiers. The primary purpose of this is to replicate the results from Romero et al. (2020) as a baseline sanity check. We expect that the non-parametric methods produces results similar to that of the OLS results and there should not be much difference in the estimates of the different non-parametric methods because we have the same information.

Second, we do not include group indicators as covariates but instead include "school facilities" in the covariates. "School facilities" was the variable that was used to create the groups (presumably because "school facilities" affects student outcomes). We assume that outcome varies across groups only due to differences in "school facilities" (there are no unobservable factors that affect outcomes across groups). Under this assumption, flexible-non-parametric ML methods should produce accurate results even if do not include group identifiers because

the ML methods should flexibly estimate how "school facilities" affects the outcome [1].

If we find that the ML methods produce the same results whether the covariates include group indicators or "school facilities", then it would highlight a potentially useful feature of ML methods. In some observational studies we may know outcomes vary across groups as a non-linear function of some variables, but we do not observe group identifiers (say due privacy restrictions) [2]. Our analysis provides evidence which suggests that the examined ML methods, in certain situations, can produces accurate estimates of ATEs even if group identifiers are unobservable.

The second part of our project relates to estimating Heterogeneous Treatment Effects (HTEs). Romero et al. (2020) posits that treatment effects are heterogeneous across providers. Interestingly, the providers are not identified in the public-use data. Consequently, we cannot directly estimate the heterogeneity across provider. Fortunately, we observe the variable - "school facilities" - which was used to assign schools to providers. We estimate Heterogeneous Treatment Effects (HTEs) using a variety of tree and forest methods (without using group or provider identifiers) in our sample. We use our trained (estimated) models to predict the Conditional Average Treatment Effect (CATE) for each observation, which we in turn use to calculate the predicted group ATE (i.e., the average of the CATE for each group) for all groups. We then calculate the ATE for each group separately (using the group identifiers). This provides us with a distribution of ATEs across groups. We refer to this distribution as the "true empirical ATE" distribution. We consider this to be the true distribution because it utilizes information about the experiment design. We then compare the true and predicted empirical distribution visually and analytically using the Kolmogorov–Smirnov test (Massey Jr, 1951) to evaluate the performance of the ML methods.

Last, we use the predicted group ATEs to extrapolate which groups were assigned to which

[1] However, the same cannot be said about OLS because OLS imposes linearity. We need that "school facilities" affect the outcomes linearly for us to have the same results when we use "school facilities" instead of group fixed effects.

[2] A hypothetical example is if we observe that healthcare outcomes are different for people living in different neighborhoods due to differences in income. However, we may not observe the addresses of individuals due privacy concern but may be able to observe the income of individuals.

providers. Under the assumption that the treatment effect is constant for a given provider but heterogeneous across providers, we should observe clusters of group treatment effects with each cluster corresponding to a given provider. We check to see if the distribution of average of treatment effects across clusters is qualitatively similar to the distribution of ATEs across providers. For example, we check to see if the cluster with lowest average of group treatment effects is centered around -0.19 which was the lowest ATE across providers. We consider two clustering methods: k-means and clustering by quantiles.

In summary, we employ ML methods to recover treatment effects by groups in a variety of ways. The exercise helps us to understand the applicability and limitations of the different methods. Moreover, we test the waters for a novel applications of these methods to observational datasets in which group identifiers are unobserved. The rest of the project is organized as follows. In Section 2, we describe the experiment in detail. In Section 3, we argue that the assumptions of unconfoundedness and overlap hold in our study. In Section 4, we present and discuss our results. In the appendix, we provide a description of our implementation procedure and provide the code.

2 The Experiment

2.1 Design

2.1.1 Overview

We analyze The Partnership Schools for Liberia (PSL) program which was analyzed in "Outsourcing Education: Experimental Evidence from Liberia" by Mauricio Romero, Justin Sandefur, and Wayne Aaron Sandholtz (Romero et al. (2020)). The PSL program delegated management of government schools and employees to private providers. The primary motivation behind the program was to improve access and quality of public education in Liberia.

The Government selected eight providers to manage public schools under the PSL program.

The providers followed the national curriculum but intervened in different ways such as new teaching materials, teacher training, and managerial oversight of the schools' day- to-day operations. The type and intensity of the activities was quite heterogeneous across providers. Additionally, the Government provided each private provider with additional funding (on top of existing funding) of $50 per pupil.

The Government chose 299 schools that satisfied certain eligibility criteria for the PSL program. Schools in the RCT generally have better facilities and infrastructure than most schools in the country. PSL schools remain public schools and are required to be free of charge and non-selective. PSL school buildings remain under the ownership of the government. Teachers in PSL schools are civil servants, drawn from the existing pool of government teachers.

2.1.2 Assignment Mechanism and Outcomes

The assignment mechanism for the PSL program followed a matched-pair design. The providers submitted a list of the regions in which they were willing to work to the government. Based on preferences and requirements of the providers, the list of eligible schools was partitioned across providers. The schools were then paired based on a principal component analysis (PCA) index of school resources. (We refer to this variable as "school facilities".) The pairing stratified treatment by school resources for *each provider but not across providers*. That is, the schools treated by different providers were heterogeneous in terms of resources. Then within each pair (also refereed to as "group") a school was randomly chosen to be treated [3].

The outcomes were measured in multiple ways at multiple levels. For the present project, we only use a composite item response theory (IRT) model as the outcome variable. The IRT is essentially a test score that measures student learning. We only examine the outcome at the end of the experiment. The IRT scores are normalized with respect to the control group.

[3]Some providers did not treat a few assigned schools and some students moved across school during the experiment. Strictly speaking, the estimates are Intent-to-Treat Estimates rather than Treatment Effects. In order to keep the terminology simple, we just discuss the effects as if they were Treatment Effects.

2.2 Data

We download the data from the public use repository. We extract four key sets of variables. First, we extract the IRT scores which is our outcome of interest. Second, we extract an indicator which tells us whether a school was treated.

Third, we extract student characteristics as controls. In line with the original paper, we use four variables: age, gender, wealth index, and grade in 2015. Wealth index is an index constructed by the authors using several measures of a student's resources. The school-level covariates we study are: enrollment, rural, school facilities, time to bank, and NGO donations. Enrollment measures a school's enrollment in 2015; rural is indicator of whether the school is rural; school facilities is an index created by PCA that measures the quality of a facilities at a given school; time to bank is how far the school is from a bank and measures how urban the school is; and NGO donations measures the donation received by a school form various NGOs in 2015.

It is critical that we observe school facilities because school facilities is used for stratification in treatment assignment. This is important for two reasons. First, we know different providers were willing to manage schools which had certain facilities; consequently, school facilities variable is likely to be correlated with providers. Second, we expect to find that the treatment effect is heterogeneous by school facilities because this is what (presumably) motivated the stratification in the first place.

We have identifiers that allow us to uniquely identify students, schools, and groups. Critically, we do not have identifiers for providers. The identity of providers was not available in the public use dataset because of privacy restrictions.

3 Identification of Causal Effects

We briefly argue that two key assumptions necessary for causal inference - unconfoundedness and overlap - are met in our study. If our application was a simple RCT then these two

assumptions are met by construction. The key feature of the experiment is that is there was a match-pair design.

In standard notation, unconfoundedness requires that

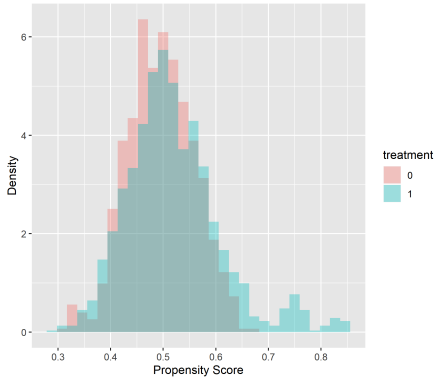
$$(Y_i(1), Y_i(0)) \perp W_i | X_i \, .$$

Overlap holds if $\epsilon < e(X_1) < 1 - \epsilon \forall i$ and some small positive ϵ.

For causal inference using a given match-pair treatment is assigned randomly. Consequently, unconfoundedness and overlap are satisfied (Lu and Deng (2017)). For analysis across match pairs, the treatment assignment depends on how the matching and stratification is done. In particular, we need to observe the variables which were used to match. In our experiment, the matching was done based on school facilities which we fortunately observe.Consequently, as long as X_i includes school facilities in our analysis, unconfoundedness and overlap are satisfied (Imai (2008)).

We also empirically check overlap in Figure 1 and find that - as expected in RCTs - overlap is not an issue.

Figure 1: Overlap in Propensity Scores



4 Results

We now proceed to analyze the effects of the PSL intervention using ML methods. Our analysis first focuses on replicating the general ITT estimates found by Romero et al. (2020) and then explores possible sources of treatment heterogeneity using the publicly available data. It is worth noting that throughout the whole analysis we only estimate ITT effects. However, we will use ATE and ITT interchangeably for the purpose of clarity as the former was the terminology used in class.

4.1 Replication of ITT Effects

The PSL program evaluated by Romero et al. (2020) relied on a randomized matched-pair design that allowed them to identify estimates for the ITT effects of the program using a simple OLS specification with group fixed-effects. Our first step is to replicate their results using a set of ATE estimators which use ML methods to correct for confounding by observables.

We first use the simple regression method used by the authors and compare it to an IPW estimator and an AIPW estimator. Our analysis focuses on the main outcomes of academic achievement that the authors use to asses the effects of the program (a composite score of English, Abstract thinking, and Math calculated using item response theory for proper comparisons). We only analyze the composite score in the present project. Since the PSL intervention was randomized, we expect that all the methods we employ will yield very similar results to each other and to what the authors find using only the stratification variable (in our case the group fixed effects). This is indeed what we find in the Table 1.

If we do not include group identifiers, then the design does not allow a simple comparison between treated and control units because of confounding that we know exists because providers selected groups before the randomization of within group treatment assignment. Because these assignment of groups to provider was done based on observables, our ML methods should correct for the confounding even in the absence of group identifiers. Particularly, we expect that a regular unweighted estimation and an IPW estimation do not solve

the problem because the propensity score is not the issue at hand; all schools technically had a 50% chance of being treated. The issue is that the potential outcomes are potentially correlated with observables that were used to assign groups to providers; that is, better providers could have selected schools with better or worse facilities. Our predictions were are in line with the results as shown in Table 1. AIPW is the best estimator to retrieve the real ITT effect when the group identifier is not observed because it adjusts for correlation between potential outcomes and observables.

Table 1: Compare Treatment Effect Estimates Under Different Weighting

	Include Group FE		Omit Group FE	
Type	estimate	se	estimate	se
Unweighted	0.193	0.032	0.210	0.052
IPW	0.195	0.032	0.210	0.051
AIPW	0.194	0.029	0.194	0.029

4.2 Treatment Heterogeneity

In addition to using ML methods to get estimates for average treatment effects, we believe these methods can contribute to the analysis of the PSL intervention by providing better insights about the heterogeneity of treatment effects (HTE). Romero et al. (2020) provided some insights about possible sources of heterogeneity in treatment effects. The authors claim that student characteristics do not appear to mediate the effects of PSL. Rather, they show some evidence suggesting that the main source of heterogeneity comes from differences in providers implementing PSL.

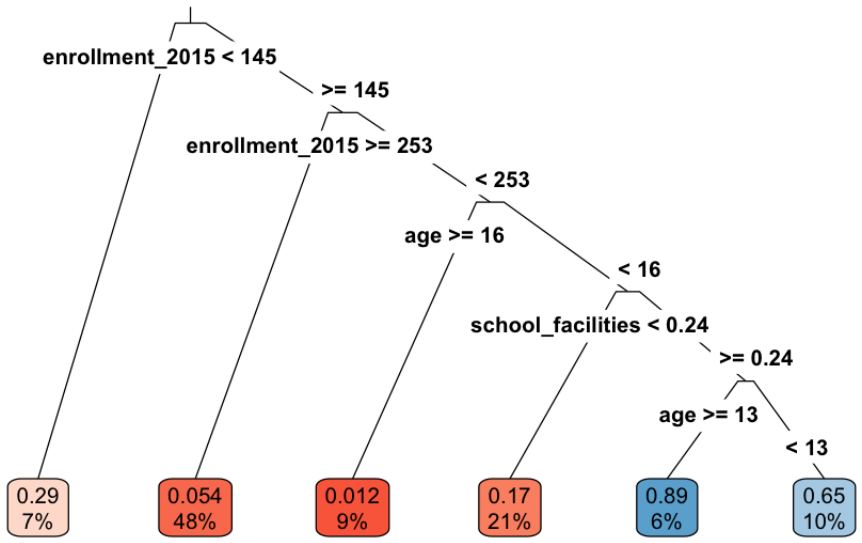
The challenge that we have when trying to further study these sources of heterogeneity is that provider information was not included in the public data. Consequently, the main purpose of our analysis is to see if we are able to uncover treatment heterogeneity that is qualitatively similar to that of Romero et al. (2020) by using only the publicly available information. Specifically, by using variables at the student, grade, or school level and without

using variables at group or provider level.

The first step we take in this direction is to explore how different methods uncover HTE in the data. To do this we compare different the results of different ML methods; specifically, we use a Causal Tree (Athey & Imbens, 2016), a S-Learner, a T-Learner, a Causal Forest (R-learner), and an X-Learner.

Causal trees (Athey & Imbens, 2016) are an interesting initial approach to the problem as it is a data driven approach to split the data based on possible heterogeneity. Figure 2 below shows the estimated splits and treatment heterogeneity found[4]. Our estimated tree can be used to illustrate how some school-level variables (enrollment and facilities) as well as some individual level variables (age) are important sources of heterogeneity.

Figure 2: Causal Tree



Note: Causal tree with minimum leaf size of 10 observations.

We can expand our analysis of heterogeneity by estimating the CATE for each observation. See code implementation for details. We split the observations into deciles using the esti-

[4]A caveat of this procedure, normal for ML methods in not very large samples, is that it was very sensible to our random choice of training data, as sometimes the tree never split. What we present in our results is an example of a choice of training and estimation datasets where heterogeneity was found.

mated CATE. To understand how different covariates affect the treatment effect, we examine how different observable characteristics map to the CATE deciles. Table 2 below shows some descriptive statistics by each decile of the CATE distribution for the T-Learner (our best performing ML model).[5] Observe that the gender column, which represents the share of males, appears negatively correlated with treatment effect deciles. Furthermore, the school facilities variable appears positively correlated with the treatment effect deciles and time to bank appears negative correlated. The other covariates analyzed do not appear to have a clear relationship. Despite these relationships, the underlying mechanisms for how covariates interact with treatment is unclear.

Table 2: Average of Covariate Within Each Decile of Group Treatment Effect (T-Learner)

t_decile	age	gender	stud_wealth	school_fac	time_to_bank	enrlmt_2015	rural
1	12.546	0.483	0.012	-0.571	114.539	296.431	0.900
2	12.497	0.445	-0.081	-0.282	98.676	285.438	0.894
3	12.550	0.487	0.457	0.205	48.920	367.608	0.506
4	12.958	0.469	-0.124	-0.053	64.977	275.995	0.772
5	13.011	0.442	-0.192	-0.325	52.501	274.410	0.779
6	13.166	0.414	-0.292	-0.553	75.972	299.151	0.833
7	12.979	0.410	-0.140	0.111	55.129	276.685	0.889
8	12.597	0.435	-0.228	0.671	63.372	207.328	0.887
9	12.956	0.414	-0.219	0.188	69.194	288.772	0.661
10	12.884	0.399	-0.116	0.391	71.088	201.755	0.778

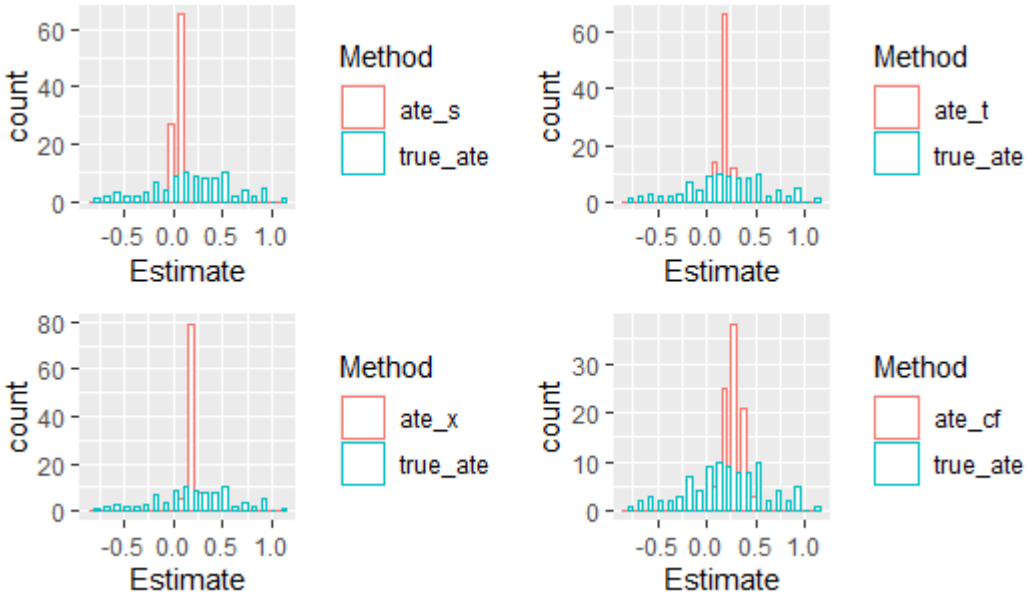
To have a better way to validate all ML methods that estimate heterogeneity in treatment effects, we take advantage of the matched-pairs design and see if the data driven methods can recover ATE effects for each group or matched pair. As these groups went through a valid randomization, it can serve as a comparison for ML methods that do not know the group information and are trying to uncover the heterogeneity of treatment effects across these groups. Moreover, we expect that the HTE algorithms will be able to find patters in

[5]We assessed the performance of the ML methods by cross-validating and calculating R-loss for each method. The T-learner had the lowest R-loss among our estimates.

observables that map to the provider heterogeneity found by Romero et al. (2020); noting again that provider information is not observable to us. The reason we believe such a map is feasible is that we observe the variable that used to assign schools across providers. Therefore, heterogeneity across providers, at least in principle, should be discoverable by heterogeneity across school facilities variables. Moreover, this mapping might give insights on the characteristics of students or schools that guided successful providers' decisions on where to operate; decisions which we know might be the source of possible confounding in this setting if group identifiers were not observed.

The details of the validation procedure we use is the following. First, we estimate the treatment effect for each matched-pair by a simple comparison of mean outcomes within each pair. We then train our estimated models to produce the CATE for each observation. For each group we average the CATE of each individual in the group to get the group CATE. If our assumption that the treatment effects only differ due to school and individual level observables the are group CATE should be similar to the ATE for a given group. We see how the distribution of the group average CATE compares to the distribution of group ATEs. The results from this exercise can be seen in Figure 3.

Figure 3: Group Heterogeneity in ATEs



We see that all ML methods produce estimates of Average CATE for groups which have a very low variance (across groups) while the group ATEs (the simple comparison of mean outcomes within group) has a lot more variance. This suggests that ML methods estimate HTEs that are shrunk towards the Average of the Treatment Effects. This may be bad feature if we this is due to overfiting bias. However, this may also be a good feature because if think the ML methods are correcting for the noise that may be present in simple mean comparisons by group (these estimates are noisy because the sample size with in each matched-pair is quite small).

We employ a more formal test to show this distributions are not the same. Let F_n^{true} be the empirical distribution of ATEs across groups and let G_n^m be the empirical distribution of the predicted average of CATEs across groups for ML method m. Our null hypothesis is that both F_n^{true} and G_n^m are the same distribution while the alternative is that they are not the same distribution. We test this hypothesis using the Two-sample Kolmogorov–Smirnov test KS.

Table 3: Results from the Klomorgov-Smirnov test for different Forest Methods

	Method	Test_Stat	Conclusion	pvalue
1	S-Learner	0.60	Reject	0.00
2	T-Learner	0.41	Reject	0.00
3	X-Learner	0.45	Reject	0.00
4	Causal Forest	0.42	Reject	0.00

The results of the KS test are shown in Table 3 and clearly show that the ML learning group ATE and the simple matched-pair estimate of ATE in each group do not agree. The S-Learner does the worst in terms of fit quality. The T-Learner, X-Learner, and Causal Forest to equally well.

The are (at least) two reasons why the ML methods may have failed to match the group ATEs. First, there may be over-fitting on observables. If this is true it underscores a known drawback these methods. Second, the group ATEs we use to evaluate the ML methods may

be too noisy. In this case, we cannot say anything about the ML methods and would need a different approach altogether. Third, the treatment effects are not just heterogeneous due to the observables but because different providers do different things and what providers do is more important than the school-level covariates that we observe. In this case, we have not provided the ML algorithm with important information and the ML method cannot infer this information from the data. This case would be a cautionary tale against throwing ML methods at a problem without understanding the deep details. The computer is not (yet) able to violate Rubin's maxim "Design trumps analysis" (Rubin, 2008).

To get at our final goal of studying heterogeneity, we want to asses if the CATE estimates found by the ML methods reflect the heterogeneity found by (Romero et al., 2020) among providers. To do this, we take our group ATE estimates for the different methods and try to find clusters within them. In particular we try to find 8 clusters as there are 8 providers in the original study. The results of this comparison are shown in Table 4. Some good news come out of this analysis, as for example the heterogeneity found by the Causal Forest for the providers with the largest treatment effects is consistent in magnitude to the true provider ATEs. This result holds for some of the true postive ATE by provider, but it is also true that the ML methods did not match the possible negative effects some providers displayed. Here it might be important to point out that these possible negative effects found by Romero et al. (2020) come from very noisy estimates, while the large positive ones do not. Hence, the ML methods we test do a relatively good job finding the provider heterogeneity authors found only when data allowed for enough statistical precision. This is a nice result in terms of the power that ML methods might have in finding patters in the data in situations where confidentiality prevents complete sharing of information. But it can also come at the expense of concealing other patters due to over-fitting; in this case very policy relevant patters.

Another feature that it is important to point out is that, as we mentioned before, treatment effects are not just heterogeneous due to the observables but because different providers do different things. Our ML methods only account for heterogeneity coming from providers selecting schools with certain observable characteristics. So the fact that we are only able to reproduce positive heterogeneous effects might come from them being related to providers

selecting schools with certain observables where the treatment is very effective. Moreover, our failure to reproduce negative heterogeneity might suggest that negative effects come from some providers doing very different things in some schools; which we cannot account for in our methods as we do not have provider identity.

Table 4: The Center Values for Clusters Generated Using k-mean Clustering on Group ATEs

True Provider ATE	ATE	T-Learner	X-Learner	Causal Forest
-0.19	-0.60	0.10	0.13	0.12
-0.08	-0.22	0.13	0.15	0.19
0.14	0.03	0.17	0.17	0.25
0.18	0.18	0.19	0.18	0.29
0.27	0.30	0.20	0.20	0.34
0.36	0.45	0.22	0.21	0.38
0.38	0.60	0.25	0.23	0.41
0.42	0.90	0.30	0.27	0.46

5 Conclusion

Our analysis provides us with a deeper understanding of the examined ML methods. We find that the AIPW is effective for estimating ATEs when potential outcomes are correlated with observables. We were not able to use ML methods to estimate HTEs accurately in the absence of provider information. At most, we are able to get some magnitudes right for positive heterogeneous effects by providers without having this information available. Although our results are for the most part negative, they helps us formulate the questions one would need to answer before using ML methods in such applications. In particular, our study underscores the importance of understanding what observable can and cannot tell us about our treatment effects even if we are using sophisticated methods.

References

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, *27*(24), 4857-4873. doi: 10.1002/sim.3337

Lu, J., & Deng, A. (2017). On randomization-based causal inference for matched-pair factorial designs. *Statistics & Probability Letters*, *125*, 99–103.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68–78.

Romero, M., Sandefur, J., & Sandholtz, W. A. (2020). Outsourcing education: Experimental evidence from liberia. *American Economic Review*, *110*(2), 364–400.

Rubin, D. B. (2008, 09). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.*, *2*(3), 808–840. Retrieved from `https://doi.org/10.1214/08-AOAS187` doi: 10.1214/08-AOAS187

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Boston, Massachusetts: Cengage Learning.

Description of Analysis

Replication of ITT Effects

The following code supports the ITT estimation in our report. We implement three methods for estimating average treatment effects: * First, we estimate direct conditional ATEs for the assigned treatment group. The authors of the study we are replicating designed treatment assignment to be unconfounded, so we might expect that direct ATE estimates are a reasonable estimate of the underlying treatment effect. The estimated treatment effect is the average of $\tau(X)$, where

$$\tau(X) = E[Y_i(1) - Y_i(0)|X_i = x]$$

* Second, we estimate a weighted average treatment effect using propensity weights. In a non-randomized study, we might be concerned that treatment is correlated with certain features X. In our context, we know that assignment balanced school-level characteristics but we might be worried that student-level characteristics in particular are unbalanced in the treatment versus control groups. If unconfoudnedness holds, then it should be the case that

$$Y_i(1), Y_i(0) \perp W_i|e(X_i)$$

We use inverse propensity weighting to balance the influence of treated and control observations for a given set of features X_i, then calculate a weighted average treatment effect. Our results confirm randomization - there is substantial overlap between the propensity scores of treated and control individuals, so the IPW estimated ATE ends up close to the unweighted ATE. * Third, we calculate an augmented inverse-propensity weighted ATE. The AIPW estimator combines elements of conditional means and inverse propensity weighting to correct for possible bias due to misspecification. The estimated treatment effect for the AIPW estimator is

$$\tau = \mathbb{E}\left[W_i \frac{Y_i - \tau\left(1, X_i\right)}{e\left(X_i\right)} + \left(1 - W_i\right) \frac{Y_i - \tau\left(0, X_i\right)}{\left(1 - e\left(X_i\right)\right)} + \tau\left(1, X_i\right) - \tau\left(0, X_i\right)\right]$$

Confirming our intuition, the AIPW estimator is reasonably close to the IPW and unweighted ATE. When we omit group fixed effects from the estimate,

```
############################
# LOAD DATA
############################
  df <- read_csv('../sample.csv') %>%
    select(-X1) %>%
    mutate(gender = gender - 1) %>%
    filter(!is.na(NGO_donations),
           !is.na(student_wealth))
  itt_out <- list()


############################
# REPLICATE TABLE 3 OF ROMERO ET AL.
############################
  rf1 <- paste0('IRT_score ~ ',
                    # Group fixed effect
                    'factor(groupid) + ',
                    # Treatment
```

```
                    'W + ',
                    # Student fixed characteristics
                    paste('age', 'gender', 'student_wealth', 'factor(grade)', sep = ' + '),
                    ' + ',
                    # School fixed characteristics
                    paste('school_facilities', 'time_to_bank', 'enrollment_2015',
                          'rural', 'NGO_donations', sep = ' + '))

  # Regression function with no group fixed effect
  rf2 <- paste0('IRT_score ~ ',
                # Treatment
                'W + ',
                # Student fixed characteristics
                paste('age', 'gender', 'student_wealth', 'factor(grade)', sep = ' + '),
                ' + ',
                # School fixed characteristics
                paste('school_facilities', 'time_to_bank',
                      'enrollment_2015', 'rural', 'NGO_donations', sep = ' + '))

############################
# ITT FUNCTION
############################
ITT_est <- function(reg_fun) {
  ############################
  # Calculate ITT OLS, cluster standard errors at school level
  ############################
    lm_out <- lm(reg_fun, data = df)
    itt_unweighted <- coef_test(lm_out, vcov = "CR1",
                                cluster = df$schoolid, test = "naive-t")['W', ]
    itt_unweighted <- tibble(type = 'unweighted',
                      estimate = itt_unweighted$beta,
                      se = itt_unweighted$SE,
                      ci_low = itt_unweighted$beta - 1.96 * itt_unweighted$SE,
                      ci_high = itt_unweighted$beta + 1.96 * itt_unweighted$SE)

  ############################
  # 2 - IPW Estimator
  ############################
    W <- as.matrix(df %>% select(W))

      # Manually create indicator variables for X$grade
      grs <- matrix(nrow = nrow(W), ncol = length(unique(df$grade)), 0L)
      grades <- unique(df$grade)
      for(i in 1:length(grades)) {
        grs[ , i] <- as.numeric(grades[[i]] == df$grade)
      }
      X <- cbind(grs,
                 as.matrix(df %>%
                             select(school_facilities, time_to_bank, enrollment_2015,
                                    rural, NGO_donations, gender, age, student_wealth)))
    # Estimate propensity weights
    p <- glm(W ~ X, family = "binomial") %>%
        predict(type = 'response')
```

```r
  # Append propensity weights to df
  df_ipw <- df %>%
    # Calculate propensity weights
    mutate(weight = (W / p) + ((1 - W) / (1 - p)))

  # Estimate IPW regression
  lm_out <- lm(reg_fun, data = df_ipw, weights = weight)
  itt_ipw <- coef_test(lm_out, vcov = "CR1",
                               cluster = df_ipw$schoolid, test = "naive-t")['W', ]
  itt_ipw <- tibble(type = 'IPW',
                    estimate = itt_ipw$beta,
                    se = itt_ipw$SE,
                    ci_low = itt_ipw$beta - 1.96 * itt_ipw$SE,
                    ci_high = itt_ipw$beta + 1.96 * itt_ipw$SE)

############################
# 3 - Double robust estimator
############################
  aipw_fun <- paste0('IRT_score ~ ',
                    # Group fixed effect
                    'factor(groupid) + ',
                    # School fixed characteristics
                    paste('school_facilities', 'time_to_bank', 'enrollment_2015',
                          'rural', 'NGO_donations', sep = ' + '),
                    # Interaction
                    ' + W * (',
                    # Student fixed characteristics
                    paste('age', 'gender', 'student_wealth', 'factor(grade)', sep = ' + '),
                    ')')
  lm_out <- lm(aipw_fun, data = df)

    # Predict - all treated
    df_treatall <- df %>%
      mutate(W = 1)
    y_treatall <- predict(lm_out, df_treatall)

    # Predict - all control
    df_treatnone <- df %>%
      mutate(W = 0)
    y_treatnone <- predict(lm_out, df_treatnone)

    # Predict actual
    actual_pred = predict(lm_out, df)

    # Calculate AIPW estimate
    G <- y_treatall - y_treatnone +
      ((df$W - p) * (df$IRT_score - actual_pred)) / (p * (1 - p))
    tau.hat <- mean(G)
    se.hat <- sqrt(var(G) / (length(G) - 1))

    # Format output
    itt_aipw <- tibble(type = 'aipw',
                       estimate = tau.hat,
```

```
                          se = se.hat,
                          ci_low = tau.hat - 1.96 * se.hat,
                          ci_high = tau.hat + 1.96 * se.hat)

      # Only plot propensity score histogram if reg function contains
      # group fixed effects
      if(str_detect(reg_fun, 'group_id')) {
          plot_tib <- tibble(treatment = factor(W), score = p)
          ggplot(plot_tib) +
              geom_histogram(aes(x = score, y = stat(density), fill = treatment),
                             alpha = 0.3, position = 'identity') +
            labs(x = 'Propensity Score',
                 y = 'Density',
                 title = 'Logit Propensity Scores')
          ggsave(filename = 'Output/Propensity Histogram.png', device = 'png')
      }

      # Return output
      return(bind_rows(itt_unweighted,
                       itt_ipw,
                       itt_aipw))

}

############################
# FORMAT OUTPUT - ITT
############################
  # 1 - Propensity score overlap
  it1 <- ITT_est(rf1) %>%
    select(-contains('ci'))
  it2 <- ITT_est(rf2)  %>%
    select(-c(contains('ci'), type))
  itt_table <- cbind(it1, it2)


  # 2 - Table of ITT Estimates
  kable(itt_table, digits = 3, format = 'latex') %>%
    add_header_above(c(" " = 1, "Include Group FE" = 2, "Omit Group FE" = 2))
```

Heterogeneous Treatment Effects

We are interested in using machine learning methods to estimate treatment effects in the PSL setting. Machine learning methods enable more flexible estimation of treatment effects, where the non-parametric flexibility allows the models to identify effects that might not have been captured under assumptions driving the previous analysis. We implement the following four methods to estimate treatment effects in the PSL setting:

- **S-Learner**: Estimate $Y(0)$, $Y(1)$ with a single model. In this application, I learn a single model $\hat{\mu}(z)$ using a single random forest that predicts Y_i from $Z_i = (X_i, W_i)$, then estimates the treatment effect for some feature vector $\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$. In general, S-learners work well when groups are of substantially different size because the learner pools information about both groups.
- **T-Learner**: The T-learner first separate models $\hat{\mu}_{(i)}(x)$ for treated and control individual $i \in \{0, 1\}$, then calculates a treatment effect as the difference $\hat{\tau}(x) = \hat{\mu}_{(1)}(x) - \hat{\mu}_{(0)}(x)$. To develop accurate models across the entire feature space, we need similarly distributions of treated and control observations across

$\mathcal{X}$. The T-learner will fail if the density of treated and control observations differ substantially

- **X-Learner**: The X-learner models $Y(0)$ and $Y(0)$ to estimate conditional average treatment effects on the treated and control observations. The model extracts the relationship between outcomes and features in separate forests for treated and control individuals, then uses the model to predict counterfactual outcomes $\hat{\mu}$. We then estimate treatment effects by regressing the difference of individual treatment effects on the covariates. The final estimate is a convex combination of the estimated CATE on treated and CATE on control observations, weighted by the estimated propensity score. The X-learner combines some of the benefits of the S- and T-learners: it fits models for treated and control observations but overcomes regularization bias by regressing the predicted treatment effect on the features. However, the X-learner does not learn treatment effect estimates using propensity scores, so it is still vulnerable to confounding
- **Causal Forest**: Estimates average treatment effects by learning multiple causal trees (partition feature space to maximize contribution to loss function, estimate constant ATE within each leaf, then average over trees to smooth). One of the benefits of the causal forest is the incorporation of propensity weighting to address confoundedness - of particular relevance in this problem.

Having laid out the benefits and disadvantages of each method above, we can predict which methods will perform well in this setting and which methods will struggle. In particular, given strong balance on observables in the treatment and control groups, we might expect that the T-learner will outperform the S- or X- learners.

We learn each model using the entire dataset, given that we are not concerned about out of sample predictions for this exercise (our objective is to recreate within-group average treatment effects without sharing group ids with the model).

```
############################
# PREPARE DATA
############################
  # Split data frame into outcome, features, treatment
  W <- df$W
  Y <- df$IRT_score
  X <- select(df, -W, -IRT_score,-studentid) %>%
    fastDummies::dummy_cols(select_columns = c('grade','schoolid'),
                            remove_first_dummy = TRUE) %>%
    select(-grade,-schoolid, groupid) %>% as.matrix() %>%
    as('dgCMatrix')

############################
# S LEARNER
############################
  # 1 - Calculate S-Learner (see slide 22 of Lecture 4)
  s_learn <- regression_forest(cbind(W, X), Y)
  pred_s_0 <- predict(s_learn, cbind(0, X))$predictions
  pred_s_1 <- predict(s_learn, cbind(1, X))$predictions
  pred_s_oob <- predict(s_learn)$predictions
  pred_s_0[W == 0] <- pred_s_oob[W == 0]
  pred_s_1[W == 1] <- pred_s_oob[W == 1]
  pred_s <- pred_s_1 - pred_s_0


############################
# T LEARNER
############################
  # 2 - Calculate T-Learner (see slide 21 of Lecutre 4)
  tf0 <- regression_forest(X[W==0,], Y[W==0])
```

```
  tf1 <- regression_forest(X[W==1,], Y[W==1])
  tf.preds.0 <- predict(tf0, X)$predictions
  tf.preds.1 <- predict(tf1, X)$predictions
  tf.preds.0[W==0] <- predict(tf0)$predictions #OOB
  tf.preds.1[W==1] <- predict(tf1)$predictions #OOB
  pred_t <- tf.preds.1 - tf.preds.0

############################
# X LEARNER
############################
  # A - Predict Y_i from X_i when W_i == 0 (use T-learner forest 0),
  # Learn \tau_1 by predicting delta from X_i when W_i == 1
  yhat0 = predict(tf0, X[W==1,])$predictions
  xf1 = regression_forest(X[W==1,], Y[W==1]-yhat0)
  xf.preds.1 = predict(xf1, X)$predictions
  xf.preds.1[W==1] = predict(xf1)$predictions

  # B - Swap: Predict Y_i from X_i when W_i == 1 (use T-learner forest 1),
  # Learn \tau_0
  yhat1 = predict(tf1, X[W==0,])$predictions
  xf0 = regression_forest(X[W==0,], yhat1-Y[W==0])
  xf.preds.0 = predict(xf0, X)$predictions
  xf.preds.0[W==0] = predict(xf0)$predictions

  # C - Estimate the propensity score - regression forest
  # of W on X
  propf = regression_forest(X, W)  # , tune.parameters = TRUE)
  ehat = predict(propf)$predictions

  # D - Estimate \hat{\tau}(x)
  pred_x = (1 - ehat) * xf.preds.1 + ehat * xf.preds.0


############################
# CAUSAL FOREST
############################
  cf <- causal_forest(X, Y, W, num.trees = dim(X)[1])
  pred_cf <- predict(cf)$predictions

############################
# WITHIN-GROUP ATE
############################
  # Estimate ATE within each matched pair of schools
  group_ate <- df %>%
    # Collapse to pair-treatment group level
    group_by(groupid, W) %>%
    mutate(avg_score = mean(IRT_score),
           sq_error = (avg_score - IRT_score) ^ 2) %>%
    summarize(avg_score = mean(avg_score),
              count = n(),
              sum_error = sum(sq_error)) %>%
    ungroup() %>%
    # Collapse to pair level to estimate pair treatment effect, error
```

```
    mutate(avg_score = if_else(W == 0, -1 * avg_score, avg_score),
           st_err = sum_error / (count - 1)) %>%
    group_by(groupid) %>%
    summarize(treatment_effect = sum(avg_score),
              st_err = sum(st_err),
              count = sum(count)) %>%
    ungroup() %>%
    mutate(st_err = sqrt(st_err))

  ############################
  # Average treatment effects in group
  ############################
    ate_est <- tibble(groupid = df$groupid,
                      ate_s = pred_s,
                      ate_t = pred_t,
                      ate_x = pred_x,
                      ate_cf = pred_cf)
    ate_est <- ate_est %>%
      group_by(groupid) %>%
      summarize_all(mean) %>%
      inner_join(group_ate %>%
                   select(groupid, count, treatment_effect) %>%
                   rename(true_ate = treatment_effect),
                 by = 'groupid') %>%
      select(groupid, count, true_ate, everything())

  # Export output
  write.csv(ate_est, '../Intermediate/2020.06.04 compare ate estimates.csv')
```

Plot Causal Tree

To visually examine heterogeneity in treatment effects, we learn and plot a causal tree. The causal tree makes each split in order to maximize the different in average treatment effect between leaves but also upwardly weighs balance in leaves. The causal tree is also regularized as to not overfit and produce a tree of too great a depth.

```
#STEP 1. Split the dataset
# Diving the data 40%-40%-20% into splitting, estimation and validation samples
split_size <- floor(nrow(df) * 0.5)
split_idx <- sample(nrow(df), replace=FALSE, size=split_size)

# Make the splits
df_split <- df[split_idx,]
df_est <- df[-split_idx,]

#STEP 2. Fit the tree
#use reg_fun for model

ct_unpruned <- honest.causalTree(
  formula=rf2,            # Define the model
  data=df_split,              # Subset used to create tree structure
  est_data=df_est,            # Which data set to use to estimate effects

  treatment=df_split$W,       # Splitting sample treatment variable
```

```
  est_treatment=df_est$W,     # Estimation sample treatment variable

  split.Rule="CT",            # Define the splitting option
  cv.option="TOT",            # Cross validation options
  cp=0,                       # Complexity parameter

  split.Honest=TRUE,          # Use honesty when splitting
  cv.Honest=TRUE,             # Use honesty when performing cross-validation

  minsize=10,                 # Min. number of treatment and control cases in each leaf
  HonestSampleSize=nrow(df_est)) # Num obs used in estimation after building the tree

#STEP 3. Cross-Validate
# Table of cross-validated values by tuning parameter.
ct_cptable <- as.data.frame(ct_unpruned$cptable)

# Obtain optimal complexity parameter to prune tree.
selected_cp <- which.min(ct_cptable$xerror)
optim_cp_ct <- ct_cptable[selected_cp, "CP"]

# Prune the tree at optimal complexity parameter.
ct_pruned <- prune(tree=ct_unpruned, cp=optim_cp_ct)

#STEP 4. predict point estimates
tauhat_ct_est <- predict(ct_pruned, newdata=df_est)

#STEP 5. Compute standard errors
# Create a factor column 'leaf' indicating leaf assignment
num_leaves <- length(unique(tauhat_ct_est))  # There are as many leaves as there are predictions
df_est$leaf <- factor(tauhat_ct_est, labels = seq(num_leaves))

# Run the regression
# ols_ct <- lm_robust(IRT_score ~ 0 + leaf + W:leaf, data=df_est)

# ols_ct_summary <- summary(ols_ct)
# te_summary <- coef(ols_ct_summary)[(num_leaves+1):(2*num_leaves), c("Estimate", "Std. Error")]

#STEP 6. Predict point estimates on test set
# tauhat_ct_test <- predict(ct_pruned, newdata=df_test)


rpart.plot(
  x=ct_pruned,        # Pruned tree
  type=3,             # Draw separate split labels for the left and right directions
  fallen=TRUE,        # Position the leaf nodes at the bottom of the graph
  leaf.round=1,       # Rounding of the corners of the leaf node boxes
  extra=100,          # Display the percentage of observations in the node
  branch=.1,          # Shape of the branch lines
  box.palette="RdBu") # Palette for coloring the node
```

Histograms of treatments by different methodologies

To compare heterogeneity across our different models we plot historgrams of the group-level treatment effect distribution.

```
par(mfrow=c(2,3))
hist(ate_est$ate_cf, main="Causal Forest: Group-level average of CATE's")
hist(ate_est$ate_t, main="T-learner: Group-level average of CATE's")
hist(ate_est$ate_s, main="S-learner: Group-level average of CATE's")
hist(ate_est$ate_x, main="X-leaner: Group-level Difference-in-means")
hist(ate_est$true_ate, main="Matched pairs: Group-level Difference-in-means")
# computing difference in means by group
# ATE for each group using causal forest -> histogram
# Compare the distribution using
```

Comparison of covariates by decile on groups (which are pairs of schools (which have many observations each))

To analyze how covariates vary across treatment effect deciles we form tables for the covariate values across treatment effect deciles.

```
# set number of quantiles
ntiles = 10
#define covariates to get summary statistics
covariates = c('age', 'gender', 'student_wealth', 'school_facilities', 'time_to_bank', 'enrollment_2015

# merge
df_group <- ate_est %>% inner_join(df
                      %>% group_by(groupid) %>% summarise_at(covariates, mean))

#create ranks
df_group$t_decile = ntile(df_group$ate_t, ntiles)
df_group$x_decile = ntile(df_group$ate_x, ntiles)
df_group$decile = ntile(df_group$true_ate, ntiles)


#generate summary statistics
sum_stats_t <- df_group %>% group_by(t_decile) %>% summarise_at(covariates, mean)
sum_stats_x <- df_group %>% group_by(x_decile) %>% summarise_at(covariates, mean)
sum_stats_a <- df_group %>% group_by(decile) %>% summarise_at(covariates, mean)

#print tables
print(sum_stats_t)
print(sum_stats_x)
print(sum_stats_a)

#print tables
xtable(sum_stats_t, digits = 3, type = 'latex')
xtable(sum_stats_x, digits = 3, type = 'latex')
xtable(sum_stats_a, digits = 3, type = 'latex')
```