Machine Learning Who to Nudge

Causal vs Predictive Targeting in a Field Experiment on Student Financial Aid Renewal

Susan Athey
Stanford

Niall Keleher
RIPL

Jann Spiess
Stanford

August 12, 2021

Abstract

In many settings, organizations seek to target products or services to constituents using pre-existing information. In this paper, we estimate the value of targeting in the context of a large-scale field experiment with over 53,000 college students, where the goal was to use "nudges" to encourage students to renew their financial aid applications before a non-binding deadline. Our preferred approach uses a causal forest to estimate heterogeneous treatment effects, and then assigns students to treatment according to those estimated to have the highest treatment effects. We compare this to a policy where we target those students with a low predicted probability of renewing financial aid in absence of the treatment. Targeting based on predicted outcomes is clearly not optimal, but it is a common strategy used by organizations in practice when a treatment is new or there is limited historical data. We show that the causal method outperforms the predictive approach. We estimate that assigning using the causal method does better than assigning nudges randomly, while assigning based on low predicted baseline does significantly worse. In terms of the effectiveness of nudges, the estimated treatment effect heterogeneity from the random forest remains modest and noisy: students with an above-median estimated treatment effect are 25% to 65% more likely to file relative to below-median students across two baseline specifications and two experiment years. Our analysis suggests that the reminders to renew financial aid work best for students who would have been relatively likely to file for renewal even in the absence of the nudge, while they remain ineffective for students who are unlikely to file at baseline and are likely to drop out.

We thankfully acknowledge support from the Alfred P. Sloan Foundation and Schmidt Futures through the "Computational Applications for Behavioral Science" project with ideas42. Matthew Schaelling provided exceptional research assistance. For their support we also thank Octavio Medina, Rebecca Nissan, Rachel Rosenberg, and Josh Wright of ideas42, Vitor Hadad and Henrike Steimer of the Golub Capital Social Impact Lab at Stanford GSB, and the Office of Institutional Research and Assessment of the City University of New York.

1 Introduction

A growing number of randomized experiments set out to measure the effectiveness of behaviorally-informed nudges. Typically, these experiments are designed and analyzed to measure whether a nudge works well on average. In this paper, we utilize causal machine learning to move beyond average treatment effects towards optimal targeting of nudges. In a large-scale experiment that randomized behaviorally-informed reminders to increase student financial aid renewal applications, we estimate not just whether the nudge worked on average but whether it worked for some students better than for others. We then ask how such heterogeneous treatment effect estimates can improve the delivery of products and services.

Our application considers data from a field experiment among over 53,000 college students. The experiment aimed to measure the causal effect of behavioral nudges on timely applications for financial aid. Across two randomized controlled trials run in 2017 and 2018 by ideas42 and the City University of New York, enrolled students were randomly assigned to receive behaviorally informed text and email reminders about renewing their federal financial aid. The average treatment effect of the behavioral nudges was noteworthy. Students who received nudges were on average 6.4 ± 0.6 (2017) and 12.1 ± 0.7 (2018) percentage points more likely to submit their Free Application for Federal Student Aid (FAFSA) forms by the priority deadline, increasing early filing rates from 37% to 43% and 38% to 50%, respectively.

For whom were the nudges most effective, and thus whom should we target if there is a limited budget? A priori, it is not obvious. There are likely some students who are not affected by nudges, perhaps because they are already committed to file; and there are likely other students who are committed to not file, for example if they do not plan to attend the next year. Treatment effects will only be positive for those who are not yet committed. Within that group, there may be some types of students who are not responsive to nudges like the ones considered in our experiment.

The problem of whom to target for an intervention arises in many settings, ranging from prioritization of salespeople to allocation of advertising spend. A common approach in practice is based on predicting which individuals are most at risk of some undesirable outcome, such as customer churn. Predictive approaches are attractive because they can be applied with observational, historical data without the need to run an experiment; they can be used even for a treatment that has never been tried before. In the context of nudges to file financial aid forms, we could imagine forming a hypothesis about which type of student would be most influenced by a nudge and building a model to predict that outcome using data about student behavior in the absence of the treatment. If we hypothesized that students who were otherwise unlikely to file would be most influenced by the nudge, we would target the students with the lowest predicted probability of filing in the absence of the treatment.

However, in general it is an empirical question as to what type of student is most likely to be influenced. In this paper, we use causal machine learning estimate how treatment effects vary with individual characteristics, and we estimate the benefits of a policy that (counterfactually, under a budget constraint) assigns students with the highest treatment effects to receive the nudge. We further estimate the differences in expected outcomes between this policy and three alternative counterfactual

policies: (i) targeting students most likely to file, (ii) targeting those least likely to file, and (iii) random selection. We find that the policy that targets treatment effects performs about as well as one that targets students most likely to file, with both better than a random policy, while a policy that targets those least likely to file does very poorly. This finding mirrors Ascarza (2018), who studies the effect of a customer retention program on reducing churn in two field experiments, and shows that targeting based on predicted churn performs considerably worse than targeting based on estimated treatment effects.

Our results provide an example where naive targeting based on a machine-learning prediction of outcomes (particularly, one that targeted those who might have seemed to need the nudges most) performs substantially worse than targeting based on estimated treatment effect heterogeneity. On the other hand, if we had correctly guessed that it was effective to target those who were already most likely to file, we would have achieved similar performance. Our findings highlight the value of augmenting machine-learning algorithms, which provide powerful prediction tools, with careful causal inference to tackle policy problems. We also emphasize the role of an improved analytical toolkit in the design and analysis of randomized field experiments.

These results suggest two general conclusions. First, it clarifies the importance of integrating causal inference and randomized trials into machine learning to analyze and improve policy, rather than relying on predictive tools based on non-experimental baseline data alone. Second, the effects of treatment that are only small interventions, as is often the case for nudges, may not accrue mainly for those with low baseline outcomes, but rather for those who would already have been likely to obtain a better outcome in the absence of treatment – and just need to be "nudged" over the finish line.

What factors drive heterogeneity in treatment effects? We find that treatment effects vary systematically between students based on information available before reminders are sent. When using only information available before the start of the semester, this variation is modest across both years: the half of students with higher estimated treatment effects are, on average, around two (2017) or three (2018) percentage points more likely to respond to treatment than the lower half. Once we incorporate information available closer to the date reminders were sent, these differences become more pronounced in the 2018 cohort, pushing the difference between those with below-median predicted effect and those with above-median predicted effect to around seven percentage points.

There is no enrollment restriction on who can apply for FAFSA. Thus students could unenroll in a given year yet remain eligible to apply for FAFSA for the subsequent academic year, and indeed it is common for students to come and go from enrollment at CUNY, where many students are working alongside their studies. However, students who drop out midyear may be not only be harder for administrators to track but they may also be less likely to re-enroll for the subsequent academic year. Using administrative data, we are able to identify students that were enrolled at the start of the academic year yet dropped out by the time of the behavioral nudge treatment. For the randomized experiment, students that were unenrolled were still eligible for the behavioral nudges. This feature of the experimental design allows us to compare treatment effects conditional on predicted enrollment at the time of the treatment.

We find that enrollment status, once it becomes available, is highly predictive of treatment effects.

Students who are unenrolled at the time of the behavioral nudge campaign had smaller treatment effects. Using administrative data from the early part of the academic year, when policymakers would not know whether a student would be enrolled at the time of the nudge, we confirm that heterogeneity is plausibly related to enrollment: ordering people by their predicted probability to enroll explains a similar variation of treatment effects as estimated treatment effects themselves do. On the other hand, once we restrict ourselves to those students who are still enrolled when the treatment is sent out, we find only suggestive (2017) or no (2018) remaining systematic heterogeneity in treatment effects. The importance of enrollment as a treatment effect moderator is also apparent when we inspect which variables vary most with estimated treatment effects.

We then evaluate the heterogeneity our machine-learning approach was able to identify in terms of its implications for improving policy delivery. With only early information available, improvements over randomly targeting students are modest and comparable to the gain from sorting students by predicted enrollment. Once additional information becomes available, including updated enrollment information, the value of targeting increases. If we rank students by estimated treatment effects and send reminders to the top 50%, we can achieve 65% of the increase in early FAFSA filing that we would achieve by sending reminders to everybody. Once we consider only enrolled students, the only improvement from targeting appears to arise from a few students that are estimated to have a very small treatment effect and would thus not be targeted, but we cannot rule out that this latter result is due to noise.

Our analysis uncovers important challenges in applying machine learning to improve the analysis and targeting of nudges. The environment we study has a fairly low signal-to-noise ratio, and we find that treatment effect estimates are unlikely to be well-calibrated. Thus, describing heterogeneity requires additional diagnostic tools to avoid small-sample biases, such as group-wise analysis discussed by Chernozhukov et al. (2019).

We move beyond describing the performance of machine-learning policies in terms of raw predictive power and instead provide analogs to receiver-operating characteristic (ROC) diagnostics adapted to the problem of treatment assignment, where the performance of the model is quantified in policy-relevant units (following e.g. Rzepakowski and Jaroszewicz, 2012; Zhao et al., 2013; Hitsch and Misra, 2018).

We build upon a growing literature that combines causal estimation with prediction techniques from machine learning (Mullainathan and Spiess, 2017; Athey and Imbens, 2019) and discusses the application of machine learning to policy problems (Kleinberg et al., 2015). For the estimation of heterogeneous treatment effects, Athey and Imbens (2016) develop causal trees from regression trees to find subgroups of individuals with different treatment effects. Wager and Athey (2018) combine many such trees into a causal random forest, which Athey et al. (2019) extend to generalized random forests. Chernozhukov et al. (2019) discuss the analysis of heterogeneous treatment effects with arbitrary machine-learning estimators and suggest diagnostic tools. Hitsch and Misra (2018) consider targeting policies derived from estimates of heterogeneous treatment effects and show how their effect can be estimated from randomized trail data. Athey and Wager (2021) analyzes efficient estimation of targeted policies with constraints on the policy class.

We also connect to a literature on behaviorally-informed nudges (Sunstein and Thaler, 2008) and their empirical validation. In the context of student financial aid, behavioral science has informed multiple cost-effective strategies for increasing FAFSA submissions that the experiment we analyze in this paper is based on. Castleman and Page (2016) show that a simple text-based intervention that encouraged FAFSA submission increased sophomore year retention by 14%. ideas42 research at Arizona State University showed that behaviorally informed student reminder emails, which include devices to trigger loss aversion, plan making, and commitment, increased priority deadline FAFSA renewal by 11 percentage points, from 29% to 40% (ideas42, 2016).

2 Experiment and Data

This paper analyzes data from a multi-year experiment conducted in New York City. Students were randomly assigned to receive behaviorally informed text and email reminders to renew their federal financial aid. The field experiment, run in 2017 and 2018 by ideas42 and the City University of New York (CUNY), aimed at increasing applications for Free Application for Federal Student Aid (FAFSA) financial support by the June 30 priority deadline. Students randomly assigned to the control group received only business-as-usual emails from the college. Students assigned to the treatment groups also received supplementary behavioral emails and text messages. These emails and text messages were designed to trigger loss aversion, plan making, and commitment.[1] Figure 1 shows example text messages sent to students in the treatment group.

The experiment involved matriculated students from CUNY community colleges. Eligible students were those who had not yet renewed FAFSA in February of the study year. The 2017 study sample includes 25,167 students from three community colleges, of which 50% were randomly assigned to treatment. The 2018 sample includes 40,638 students from five community colleges, which were included in the intervention in two batches: an early batch of 30,627, of which 45% were assigned to each of the two treatment arms and 10% to control, and a late batch of 10,011 with a larger control group of 25% and roughly equal treatment groups. We pool the late and early cohorts from 2018 for a total combined fraction of 86% treated across the two treatment arms. Throughout our analysis of 2018 data, we adjust estimates for the varying propensity scores between early and late cohorts by inverse probability-weighted estimators.

Our data include baseline demographic, academic, and administrative information about the community-college students in the experiment. On average, students were around 24 years old, with a considerable standard deviation of almost seven years. Our sample includes more women than men, with 57% women in the 2017 experiment, and 56% (early schools) and 53% (late schools) women in 2018. A plurality of students was Hispanic (52% in 2017, 45% in 2018), followed by Black non-Hispanic students who made up around a third of the student body in this study. Almost 20% of students were enrolled part-time. Overall, we do not observe large imbalances between treatment and control groups; for nine baseline characteristics we tested across the 2017 and 2018 cohorts, only one variable, GPA for

[1] During 2017, the experiment had one treatment arm. The two treatment arms in the 2018 experiment differed in whether they used one-way texts or two-way texts that prompted students to respond. For this paper, we pool the two treatment arms in the 2018 study.

Text Message Content (using BMCC texts as an example):

Msg. #	Send Date and Time	Content
0	Wed, March 1 @ 6pm	**Part 1:** Hi {First Name}! This is the CUNY Student Persistence Team. To help you finish the year strong we will send you a few helpful texts. **Part 2:** Reply CANCEL if you don't want help setting yourself up for success. **Response to "cancel":** Thanks for letting us know. You will no longer receive texts from us.
1	Tues, March 14 @ 6pm	{First Name}, you must renew your FAFSA each year. This year it's easier -- you can use the same tax info as last year! Go to http://bit.ly/FAFSABMCC today.
2	Tues, March 28 @ 6pm	Renew your FAFSA and do it right the first time! Stop by the Financial Aid Lab (S115-C) and get help renewing today.
3	Wed, April 12 @ 6pm	Renew your FAFSA today! Many people renew in 30min or less at http://bit.ly/FAFSABMCC. Tip: use the IRS data retrieval tool to renew quickly.
4	Tues, April 25 @ 6pm	Unsure how to renew FAFSA? That's OK! Many students go before/after class to FinAid Lab (S115-C) for free help. Hrs: M/Th 10-6, F 10-5.
5	Tues, May 2 @ 6pm	{First Name Last Name}: FAFSA Status—NOT RENEWED. Renew now at http://bit.ly/FAFSABMCC
6	Wed, May 10 @ 6pm	{First Name}, our records show you haven't renewed your FAFSA. Need help? Get expert guidance at FinAid Lab (S115-C: Mon-Thurs 10-6, Fri 10-5).
7	Tues, May 16 @ 6pm	You filed FAFSA this academic year, but you must renew it to be eligible for aid next year. Don't miss out on free money! Renew now: http://bit.ly/FAFSABMCC
8	Wed, May 24 @ 6pm	Hi {First Name}, quick reminder--renew your FAFSA today at http://bit.ly/FAFSABMCC
9	Tues, May 30 @ 6pm	If you don't renew FAFSA, you'll likely pay more for college next year! Save $$ and renew now: http://bit.ly/FAFSABMCC

Figure 1: Example reminder text messages sent to students in the treatment group.

late 2018 schools, is significantly different between treatment and control at the 5% level. Estimated propensity scores are concentrated around their batch-wise mean and balanced between the respective treatment and control groups.

Across the two randomized controlled trials, those who received the treatment interventions were on average 6.4 ± 0.6 (2017) and 12.1 ± 0.7 (2018) percentage points more likely to submit their FAFSA forms by the priority deadline, increasing early filing rates from 37% to 43% and 38% to 50%, respectively. These estimates, which are based on simple averages between treatment and control units within batches, are robust with respect to two alternative, augmented inverse propensity weighted (AIPW) estimators that leverage covariate information to reduce noise. The first of these estimators assumes constant propensity scores within batches, while the second corrects for possible imbalances by estimating the propensity score. Both are based on random forest estimation of the outcome model and propensity score.

3 Treatment-Effect Heterogeneity

Above we have documented a sizable average effect of behaviorally-informed reminders in this study. In this section, we use machine-learning tools to estimate treatment effects as a function of available individual covariates. We repeat the analysis for each study year (2017 and 2018) as well as for two sets of explanatory covariates – first, those available at time of randomization before the spring semester, and second, all information available halfway through the semester (which adds enrollment information and additional academic records) just before reminders were sent out. We report tests and diagnostics based on samples of 17,755 students for the 2017 study and 29,786 students for 2018. (A remaining quarter of the data remains available as an additional hold-out sample for future validation, which we do not use in the current study.)

In this study, our goal is to estimate conditional average treatment effects (CATEs), which are defined as average treatment effects conditional on observed covariates. Before describing how we estimate these treatment effects, we formally define the object of interest. We denote by $Y \in \{0, 1\}$ the random variable that expresses whether a student has filed by the priority deadline ($Y = 1$) or not ($Y = 0$), and $T \in \{0, 1\}$ for whether the student is in the treatment group ($T = 1$) or not ($T = 0$). We use standard potential-outcomes notation and write $Y(1)$ for the filing status a student would have had had they been assigned to treatment, and $Y(0)$ for the filing status had they been assigned to control. The treatment of that student is then $Y(1) - Y(0)$, and we are interested in how this treatment effect varies with some baseline student characteristics X. Specifically, we aim to estimate the conditional average treatment effect

$$\tau(x) = \mathrm{E}[Y(1) - Y(0) | X = x]$$

of students with characteristics $X = x$. When estimating $\tau(x)$, we face the challenge that for every student we only observe one of the potential outcomes $Y(1)$ and $Y(0)$, namely the realized filing decision $Y = Y(T)$ for their actual treatment status T. However, since treatment has been randomized, the

realization of $Y(1)$, $Y(0)$ and X are independent of T, so we can identify treatment effects from $\tau(x) = \mathrm{E}[Y|T = 1, X = x] - \mathrm{E}[Y|T = 0, X = x]$. In words, while we cannot compare outcomes within student, we can compare outcomes across similar students who have been treated or assigned to control, which yields the same conditional average effect when treatment in a randomized trial.

We estimate heterogeneous treatment effects with machine learning. Since treatment effects $Y(1)-Y(0)$ are not observed and therefore cannot be predicted directly by standard machine-learning methods, we employ the causal forest algorithm. The causal forest is an instance of generalized random forests (Athey et al., 2019) that is specifically adapted to solve the causal-inference problem of estimating conditional average treatment effects $\tau(x)$ in settings like ours. Causal forests recursively compute multiple partitions of the covariate space based on treatment effect heterogeneity, so that estimated average treatment effects vary as much as possible between subsets of the covariate space. To estimate the CATE $\tau(x)$ for a particular vector x of covariates, a weighted average treatment effect

$$\hat{\tau}(x) = \frac{\sum_{T_j=1} \hat{w}_j(x) Y_j}{\sum_{T_j=1} \hat{w}_j(x)} - \frac{\sum_{T_j=0} \hat{w}_j(x) Y_j}{\sum_{T_j=0} \hat{w}_j(x)}$$

of nearby observations (Y_j, T_j, X_j) is used, where weights $\hat{w}_j(x)$ are based on how often observations X_j share the same cell as the target vector x in different partitions. Moreover, because the process of constructing the partition to estimating treatment effects uses a careful rule for sample splitting ("honesty," Wager and Athey, 2018) that ensures that Y_i is not used in the estimation of $\hat{w}_j(X_i)$, the resulting estimates are guaranteed to be consistent and asymptotically normal.

Despite the theoretical guarantees, in practice the signal-to-noise ratio is often such that a very large sample is required for the theory to be a useful guide. For more realistic sample sizes, estimates $\hat{\tau}(x)$ of CATEs $\tau(x)$ are often miscalibrated. For a particular covariate vector x, the estimates $\hat{\tau}(x)$ may be biased towards the overall average treatment effect $\tau = \mathrm{E}[Y(1) - Y(0)]$, as there will not in general be enough observations with similar covariate vectors to a particular target. On the other hand, in a setting with very little true heterogeneity relative to the noise, sampling variation will still induce a distribution of estimated treatment effects over different covariate vectors, potentially overstating heterogeneity. Thus, it is important to assess the calibration of the estimates in a model-free way.

To do so, we first provide a calibration-based test for the existence of heterogeneous treatment effects based on Chernozhukov et al. (2019). Across both years and both sets of covariates, those available earlier and those later, we find statistically significant heterogeneity based on the slope estimate in a cross-fitted calibration regression of actual outcomes Y_i on treatment-effect estimates $\hat{\tau}(X_i)$ interacted with normalized treatment $T_i - p$, where p is the overall probability of being treated (Table 1).[2] We do not find any evidence for negative treatment effects, so reminders are unlikely to have caused any students not to file by the priority deadline.

Although we find evidence of heterogeneity, the model is not perfectly calibrated (the slope coefficients are substantially less than one), and as such we cannot assume that the magnitudes of our CATE

[2]To avoid biases, we estimate treatment effects $\hat{\tau}(X_i)$ for units (Y_i, T_i, X_i) within our sample by four-fold cross-fitting. Specifically, we randomly divide the sample into four folds, and for an observation i in a given fold estimate their CATE $\tau(X_i)$ by $\hat{\tau}(X_i)$ using a causal forest $\hat{\tau}$ fitted only on the other folds. We then run the calibration regression by fold and aggregate the resulting coefficient and standard error estimates.

Study year	Covariates used for estimation	Slope estimate	SE	t-stat	p-value
2017	Early (before semester)	0.4733	0.2545	1.8601	0.0314
	Late (mid-semester)	0.4767	0.2437	1.9561	0.0252
2018	Early (before semester)	0.5555	0.3212	1.7296	0.0419
	Late (mid-semester)	0.7338	0.2932	2.5024	0.0062

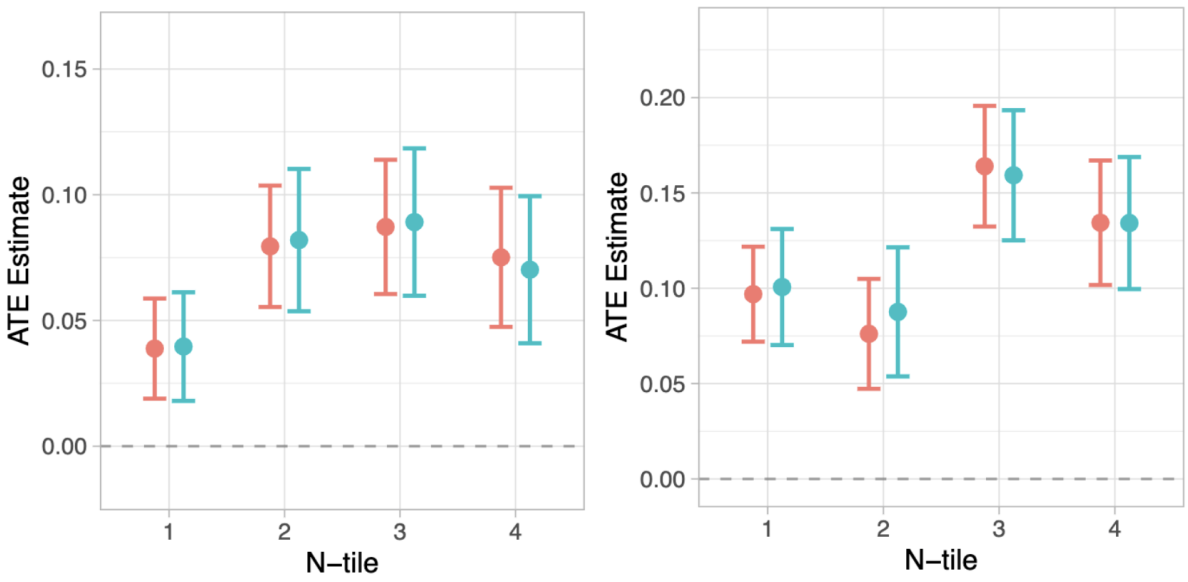
Table 1: Slope coefficient estimates for the calibration regression of actual outcomes on treatment-effect estimates interacted with normalized treatment following Chernozhukov et al. (2019).

estimates $\hat{\tau}(x)$ are unbiased for $\tau(x)$.

Despite that, the CATE estimates may be reliable for assessing which units have higher treatment effects than others. Although it is impossible to assess the accuracy of a treatment effect estimate for a single observation, since we can only observe the outcome Y for an individual with one of the two possible treatment assignments $T \in \{0, 1\}$, it is of course possible to construct an unbiased estimate of the average treatment effect $\mathrm{E}[Y(1) - Y(0)|G]$ for a sufficiently large group of individuals, where the group $G = g(X)$ is defined by covariates X. We proceed by creating such groups based on our CATE estimates. We make use of cross-fitting to remove bias, as follows. We partition the data into folds, and let $k(i)$ be the fold that leaves out observation i. For each fold k, we estimate a mapping $\hat{\tau}_{-k}$ from covariates x to the CATE using data that leaves out observations in fold k. We then divide the covariate space into four groups based on the quartile of estimated treatment effects $\hat{\tau}_{-k}(x)$, yielding a mapping $g_{\hat{\tau}-k}$ from x to quartile identifiers that represents which quartile $\hat{\tau}_{-k}(x)$. We then apply this mapping to the observations in fold k, assigning units in the into four groups based only on the covariates of those units. Finally, we estimate average treatment effects $\mathrm{E}[Y_i(1) - Y_i(0)|g_{\hat{\tau}-k(i)}(X_i) = G]$ for each of the four groups G by the average difference between treated and control outcomes within that group. These estimates are unbiased estimates of the average treatment effect for the groups (recalling groups are defined by the covariates), since the outcomes of the units in fold k were not used in any part of the process of assigning units to groups. We finally average these over the folds. The resulting estimates are model-free, unbiased estimates of average treatment effects in each group.

Our resulting estimates of treatment effects are noisy, and the heterogeneity across groups is only moderate. Focusing on the model estimated with the later set of available information, Figure 2 plots unbiased estimates of the group-wise average treatment effect across quartiles of estimated treatment effects for the 2017 and 2018 cohort.

Both graphs document that treatment effects $\tau(x)$ vary across the distribution, although the forest-based estimates $\hat{\tau}(x)$ differ from unbiased estimates and even produce non-monotonic rankings. For 2017, the average treatment effect in the lowest quartile of estimated effects is significantly lower than those of the rest of the distribution, at around 4 percentage points, relative to the average of the remaining sample at around 8 percentage points. For 2018, the average outcome in the lower half of estimated treatment effects is around 7 percentage points lower than the average above the median. Effects for the set of covariates available earlier is noisier, especially for the 2018 data.



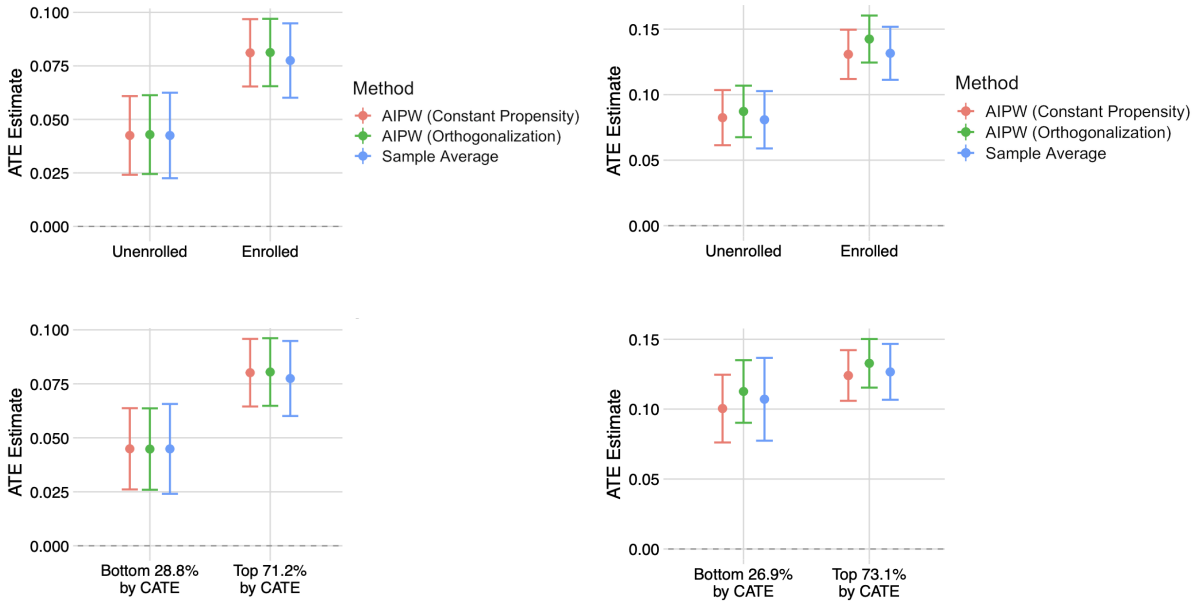
(a) 2017 (b) 2018

Figure 2: Average treatment effects by quartiles of estimated treatment effects. The x-axis divides the sample into quartiles of predicted cross-fitted treatment effects. The y-axis plots a simple difference-in-averages (cyan) as well as an augmented inverse-propensity weighted (red) estimator of the average treatment effect of the group, along with a 95% confidence interval.

We next inspect which variables vary the most along with treatment effects predicted by the random forest (Figure 4). When we order variables according to how much they vary across quartiles of estimated treatment effects, we find that variables related to enrollment are among those associated most with heterogeneity estimated from information available right before the intervention.

Our analysis suggests that most of the heterogeneity identified by the generalized random forest from these covariates does indeed come from some students dropping out and therefore being less responsive to the nudge. Indeed, if we calculate average treatment effects across enrollment status, we find that it partitions treatment effects as well as (2017) or even better than (2018) our treatment-effect estimates (Figure 3). Once we estimate heterogeneous treatment effects only among those who are still enrolled at the onset of reminders, we are unable to find sufficient evidence for additional heterogeneity (Figure 4). Only in the 2017 data do we find suggestive evidence of a group of enrolled students that responds less than other enrolled students. In the 2018 sample of enrolled students, we do not find any significant heterogeneity, and estimated effects are close to the average effect across quartiles of estimated treatment effects. This result suggests that there is not much, if any, systematic heterogeneity in how enrolled students react to the nudge. For policy purposes, this means that in 2018 we could likely not have done better in terms of targeting than sending reminders to the enrolled students.

We note that the results on enrollment as an important treatment-effect moderator are not mechanical. Indeed, reminders affect filing of both students who are enrolled and who are not enrolled at the time enrollment is measured for the spring semester. Since students may drop out and re-enroll, it remains



(a) 2017 (b) 2018

Figure 3: Average treatment effects by enrollment status (top) and by whether predicted cross-fitted treatment effects are below or above the quantile corresponding to the proportion of enrolled students (bottom), using all data up to the start of the intervention. The y-axis plots a simple difference-in-averages as well as two augmented inverse-propensity weighted ("AIPW") estimators, with batch-wise constant and with estimated propensity scores, of the average treatment effect within the group, along with a 95% confidence interval.

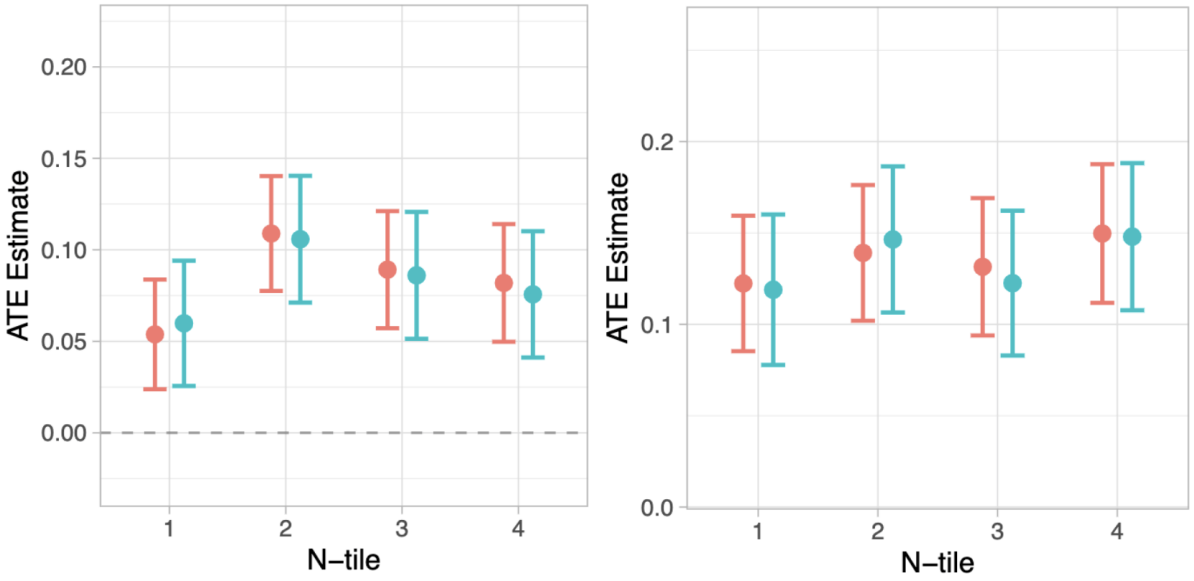
effective to provide reminders for unenrolled students. Rather, we consider it an interesting finding that there seems to be very little additional heterogeneity, making enrollment a powerful proxy for the effectiveness of reminders already by itself.

4 Effect of Targeted Policies

Building upon the estimates $\hat{\tau}(x)$ of heterogeneous treatment effects from the previous section, we now ask how we can use such information to improve the targeting of reminders. While most heterogeneity in conditional average treatment effects $\tau(x)$ can be predicted by enrollment once we know whether a student has dropped out, we now ask whether we could have used predictions of heterogeneous treatment effects to select students to send behaviorally-informed reminders before the beginning of the semester, when enrollment information was not yet available. We focus on the 2017 cohort.

Rather than evaluating the quality of our predictions $\hat{\tau}(x)$ in terms of raw prediction loss relative to $\tau(x)$, Figure 5 quantifies which proportion of total gain from the reminders we could have realized when targeting a given fraction of students.

Here, we leverage the insight that the outcome under alternative assignment policies can be esti-



(a) 2017 (b) 2018

Figure 4: Average treatment effects by quartiles of estimated treatment effects among enrolled students. The x-axis divides the sample into quartiles of cross-fitted estimated treatment effects. The y-axis plots a simple difference-in-averages (cyan) as well as an augmented inverse-propensity weighted (red) estimator of the average treatment effect within each quartile, along with a 95% confidence interval.

mated from existing trial data by exploiting randomized assignment, which permits the estimation of $\mathrm{E}[Y(\pi(X))] = \mathrm{E}[\mathrm{E}[Y|T=1,X]\,\pi(X) + \mathrm{E}[Y|T=1,X]\,(1-\pi(X))]$ for an assignment policy that maps characteristics x to an assignment $\pi(x) \in \{0,1\}$ (e.g. Hitsch and Misra, 2018). Specifically, we consider assignment policies

$$\hat{\pi}_t^{\text{causal}}(x) = \mathbb{1}(\hat{\tau}(x) \geq t)$$

that assigning all students to treatment whose estimated treatment effect is above some threshold t. For each threshold, t, we can estimate (using four-fold cross-fitting) the average outcome $U_t^{\text{causal}} = \mathrm{E}[Y(\hat{\pi}_t^{\text{causal}}(X))]$ we could have achieved using this policy. For every value of t, Figure 5 plots the resulting estimate $\hat{U}_t^{\text{causal}}$ of U_t^{causal} on the y-axis against the fraction of individuals in our sample with $\hat{\tau}(X_i) \geq t$, allowing us to evaluate and compare policies based on the total increase in FAFSA renewal relative to the number of students who are sent reminders.[3] We estimate that we could have increased FAFSA renewal at the priority deadline from 36.5% to 41% (realizing 80% of the gain) by targeting those 70% of students with the highest predicted effect. Future enrollment prediction, where we instead predict future enrollment and assign those students with the highest probability of enrollment, would have been comparatively successful, providing further evidence that heterogeneous treatment effects here are related to enrollment.

[3]Such policy ROC curves (also called "uplift curve", "profit curve", or "cost curve") have also been used to represent benefits of targeting at varying costs e.g. by Rzepakowski and Jaroszewicz (2012); Zhao et al. (2013); Hitsch and Misra (2018); Sun et al. (2021).

We compare targeting by causal treatment effects to a purely predictive targeting rule. Specifically, we predict the probability $f(x) = \mathrm{E}[Y(0)|X = x]$ that a student would have been to file by the priority deadline absent the behaviorally-informed reminders and give treatment first to those with the lowest predicted probability. This targeting rule has intuitive appeal since those who are least likely to file are those with the highest potential for the treatment to have a large effect. It can also be implemented efficiently using any off-the-shelf machine-learning predictor that predicts filing by the deadline from available variables in the absence of an experiment, since $f(x) = \mathrm{E}[Y|X = x, T = 0]$. However, the predictive policy

$$\hat{\pi}_b^{\text{predictive}}(x) = \mathbb{1}(\hat{f}(x) \leq b)$$

based on a random-forest prediction $\hat{f}(x)$ performs significantly worse than the policy based on the causal estimation of treatment effects. Indeed, we estimate that the outcome $\mathrm{E}[Y(\hat{\pi}_t^{\text{predictive}}(X))]$ of the prediction-based approach is considerably worse than assigning people randomly.

Assigning those to treatment that are least likely to renew at baseline is just one of many prediction-based assignment policies we could have considered. An alternative policy that sends nudges to those students who are *most* likely to renew at baseline (the "Inverse Baseline" policy in Figure 5) has performance that is comparable or even exceeds that of the causal policy, showing that nudges may be most effective for those who are already most likely to respond at baseline. Another natural baseline – targeting those with intermediate baseline probabilities of filing for financial aid, denoted by "Middle Baseline" in Figure 5 – seems not to do much better than random targeting in this case. While many of these approaches may have been plausible ex-ante, and one of the predictive policies performs well in this case, only through the ex-post evaluation from the experiment did we learn which one it was. The causal approach has the advantage that it directly estimates a policy that we can expect to work well based solely on the empirical relationships of covariates to treatment effects, rendering guessing a policy that may work well (or testing a large number of them explicitly) unnecessary.

5 Conclusion

The failure of a purely predictive policy relative to a method that combines causal inference and machine learning provides an example of the value of integrating careful experimentation, causal inference, and predictive modeling. By itself, predictive machine learning could have led to a bad policy, but a purpose-built algorithm run on a randomized experiment provides a coherent analysis that can inform future policy development. Our analysis also points to the challenges of evaluating existing experiments with machine learning. While sample and effect sizes seem large for estimating average treatment effects, an intervention designed to work well on average in an experiment powered for estimating averages makes the precise estimation of heterogeneous treatment effects statistically and technically challenging. Future experiments could also rely on an integration of targeting into their design.

Our results come with important caveats that limit statistical power, generalizability, and policy appli-

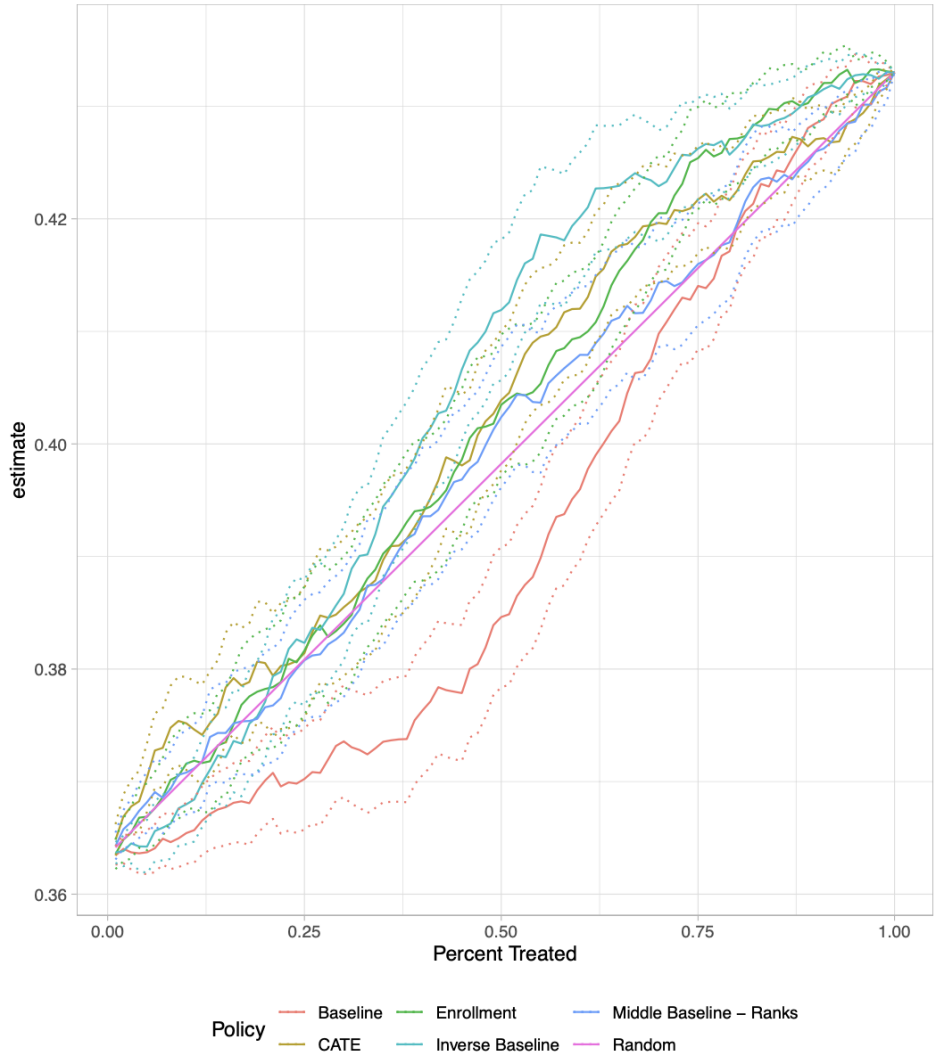


Figure 5: Total estimated FAFSA renewal rate (y-axis) by targeting a given fraction (x-axis) of students according to different cross-fitted predictions in the 2017 data with early covariates, including a prediction of outcomes absent treatment ("Baseline"), a prediction of future enrollment ("Enrollment"), and a prediction of treatment effects ("CATE"). Shown are augmented inverse-propensity weighted estimates with 95% confidence intervals that represent the pointwise uncertainty of the difference in renewal rate relative to the random policy that assigns the same fraction to treatment.

cability. Since this experiment was not designed for heterogeneous-treatment effect analysis, treatment arms were chosen to work well on average, rather than for specific subgroups. Overall treatment effects are moderate since they come from relatively modest nudges, and the experiment was powered to detect average effects rather than effects on many subgroups. Finally, sending behaviorally informed reminders is cheap and does not appear to have any negative treatment effects in the experiment, so while these results can help target reminders to those for whom they will work best, the main effect remains limited to avoiding inundating students with reminder texts and emails for who the effect would be small.

We believe that overcoming these shortcomings in future studies requires designing experiments *ex ante* to estimate heterogeneous treatment effects in the first place. This includes designing individual treatment arms that are likely to affect different people differently so that differentiated treatments can be matched to appropriate individuals and situations. It also involves updating power analyses to the higher sample size demands for estimating heterogeneous treatment effects, rather than average effects alone. Finally, policies based on heterogeneous treatment effects will be particularly important when treatment delivery is costly, or we need to make choices between treatments for which none dominates others across individuals.

References

Ascarza, E. (2018). Retention futility: Targeting High-Risk customers might be ineffective. *J. Mark. Res.*, 55(1):80–98.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. U. S. A.*, 113(27):7353–7360.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annu. Rev. Econom.*, 11(1):685–725.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Ann. Stat.*

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Castleman, B. L. and Page, L. C. (2016). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence. *J. Hum. Resour.*, 51(2):389–415.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2019). Generic machine learning inference on heterogenous treatment effects in randomized experiments.

Hitsch, G. J. and Misra, S. (2018). Heterogeneous treatment effects and optimal targeting policy evaluation. *SSRN Electronic Journal*.

ideas42 (2016). Meeting the FAFSA priority deadline. Technical report.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *Am. Econ. Rev.*, 105(5):491–495.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *J. Econ. Perspect.*, 31(2):87–106.

Rzepakowski, P. and Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and information systems*, 32(2):303–327.

Sun, H., Du, S., and Wager, S. (2021). Treatment Allocation under Uncertain Costs.

Sunstein, C. R. and Thaler, R. (2008). *Nudge*. Yale University Press.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.*, 113(523):1228–1242.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L. J. (2013). Effectively Selecting a Target Population for a Future Comparative Study. *Journal of the American Statistical Association*, 108(502):527–539.

A Additional Tables and Figures

	Control $N = 12,658$	Treatment $N = 12,480$	p-value	N
COLLEGE_IN_INTERVENTION_SPR:			0.69	25138
0	3953 (31.229%)	3928 (31.474%)		
1	8705 (68.771%)	8552 (68.526%)		
AGE	23.680 (6.635)	23.556 (6.518)	0.19	19153
GENDER:			0.51	25138
Men	5396 (42.629%)	5372 (43.045%)		
Women	7262 (57.371%)	7108 (56.955%)		
ETHNICITY:			0.89	25138
American Indian or Native Alaskan	39 (0.308%)	33 (0.264%)		
Asian or Pacific Islander	1072 (8.469%)	1091 (8.742%)		
Black, Non-Hispanic	4107 (32.446%)	4067 (32.588%)		
Hispanic, Other	6558 (51.809%)	6422 (51.458%)		
White, Non-Hispanic	882 (6.968%)	867 (6.947%)		
TRANSFER:			0.32	25138
0	9104 (71.923%)	8874 (71.106%)		
1	591 (4.669%)	584 (4.679%)		
'Missing'	2963 (23.408%)	3022 (24.215%)		
FT_PT_STATUS:			0.39	25138
FULL-TIME	6610 (52.220%)	6415 (51.402%)		
PART-TIME	2441 (19.284%)	2473 (19.816%)		
'Missing'	3607 (28.496%)	3592 (28.782%)		
GPA_CUMU_BF	2.475 (0.952)	2.472 (0.947)	0.84	14633
CRD_CUMU_ATMPT_BF	18.706 (16.565)	18.785 (16.524)	0.75	17939
CRD_CUMU_EARN_BF	17.475 (15.661)	17.387 (15.505)	0.70	17939

Table 2: Balance table for the 2017 FAFSA experiment.

	Control $N = 3226$	Treatment 1 $N = 13698$	Treatment 2 $N = 13610$	p-value	N
AGE	23.412 (6.566)	23.625 (6.764)	23.425 (6.601)	0.07	23164
GENDER:				0.40	30534
Men	1464 (45.381%)	6062 (44.255%)	5998 (44.071%)		
Women	1762 (54.619%)	7636 (55.745%)	7612 (55.929%)		
ETHNICITY:				0.83	30534
American Indian or Native Alaskan	13 (0.403%)	69 (0.504%)	68 (0.500%)		
Asian or Pacific Islander	503 (15.592%)	2065 (15.075%)	2000 (14.695%)		
Black, Non-Hispanic	971 (30.099%)	4112 (30.019%)	4019 (29.530%)		
Hispanic, Other	1419 (43.986%)	6092 (44.474%)	6147 (45.165%)		
White, Non-Hispanic	320 (9.919%)	1360 (9.928%)	1376 (10.110%)		
TRANSFER:				0.59	30534
0	2319 (71.885%)	9809 (71.609%)	9776 (71.830%)		
1	132 (4.092%)	593 (4.329%)	535 (3.931%)		
'Missing'	775 (24.024%)	3296 (24.062%)	3299 (24.240%)		
FT_PT_STATUS:				0.67	30534
FULL-TIME	1706 (52.883%)	7350 (53.657%)	7342 (53.946%)		
PART-TIME	643 (19.932%)	2692 (19.653%)	2601 (19.111%)		
'Missing'	877 (27.185%)	3656 (26.690%)	3667 (26.943%)		
GPA_CUMU_BF	2.538 (0.934)	2.508 (0.942)	2.500 (0.948)	0.28	19020
CRD_CUMU_ATMPT_BF	21.635 (17.409)	21.285 (16.829)	21.275 (16.995)	0.63	22334
CRD_CUMU_EARN_BF	19.330 (15.676)	19.001 (15.169)	18.968 (15.317)	0.58	22334

(i) Early schools

	Control $N = 2497$	Treatment 1 $N = 3802$	Treatment 2 $N = 3699$	p-value	N
AGE	23.742 (6.448)	23.906 (6.724)	24.076 (6.803)	0.23	7866
GENDER:				0.50	9998
Men	1200 (48.058%)	1770 (46.554%)	1745 (47.175%)		
Women	1297 (51.942%)	2032 (53.446%)	1954 (52.825%)		
ETHNICITY:				0.54	9998
American Indian or Native Alaskan	9 (0.360%)	10 (0.263%)	6 (0.162%)		
Asian or Pacific Islander	195 (7.809%)	313 (8.233%)	278 (7.516%)		
Black, Non-Hispanic	814 (32.599%)	1241 (32.641%)	1219 (32.955%)		
Hispanic, Other	1100 (44.053%)	1717 (45.160%)	1646 (44.499%)		
White, Non-Hispanic	379 (15.178%)	521 (13.703%)	550 (14.869%)		
TRANSFER:				0.36	9998
0	1799 (72.046%)	2753 (72.409%)	2692 (72.776%)		
1	151 (6.047%)	258 (6.786%)	213 (5.758%)		
'Missing'	547 (21.906%)	791 (20.805%)	794 (21.465%)		
FT_PT_STATUS:				0.18	9998
FULL-TIME	1309 (52.423%)	2109 (55.471%)	2019 (54.582%)		
PART-TIME	497 (19.904%)	689 (18.122%)	682 (18.437%)		
'Missing'	691 (27.673%)	1004 (26.407%)	998 (26.980%)		
GPA_CUMU_BF	2.408 (0.935)	2.459 (0.935)	2.494 (0.936)	0.02	6223
CRD_CUMU_ATMPT_BF	22.097 (17.048)	22.089 (17.149)	22.415 (17.162)	0.74	7305
CRD_CUMU_EARN_BF	19.888 (15.367)	19.900 (15.465)	20.184 (15.339)	0.74	7305

(ii) Late schools

Table 3: Balance tables for the 2018 FAFSA experiment.

Method	ATE	SE
OLS (IPW)	0.0641	0.0061
AIPW (Constant propensity)	0.0687	0.0053
AIPW (Orthogonalization)	0.0686	0.0053

(a) 2017

School timeline	Method	ATE	SE
All	OLS (IPW)	0.1209	0.0074
	AIPW (Constant Propensity)	0.1182	0.0065
	AIPW (Orthogonalization)	0.1182	0.0065
Early	OLS (IPW)	0.1213	0.0091
	AIPW (Constant propensity)	0.1198	0.0079
	AIPW (Orthogonalization)	0.1200	0.0080
Late	OLS (IPW)	0.1201	0.0109
	AIPW (Constant propensity)	0.1134	0.0099
	AIPW (Orthogonalization)	0.1132	0.0100

(b) 2018

Table 4: Overall average treatment effects, estimated by simple (propensity-adjusted) differences in averages ("OLS") as well as by an Augmented Inverse Propensity Score estimator ("AIPW") based on random forests with constant and flexible propensity score, respectively.

Top 10 covariates (by variable importance) and their means by quartile of CATE

Average covariate values in each ntile:
Top 10 covariates with widest standardized distribution

covariates	ntile_constprop1	ntile_constprop2	ntile_constprop3	ntile_constprop4
COLLEGE_IN_INTERVENTION_SPR_X0	0.625	0.326	0.091	0.033
	(0.006)	(0.005)	(0.003)	(0.002)
COLLEGE_IN_INTERVENTION_SPR_X1	0.375	0.674	0.909	0.967
	(0.006)	(0.005)	(0.003)	(0.002)
ASAP_missing	0.619	0.322	0.087	0.03
	(0.006)	(0.005)	(0.003)	(0.002)
MISSING_FROM_ENR_DATA	0.621	0.323	0.09	0.032
	(0.006)	(0.005)	(0.003)	(0.002)
ASAP_X0	0.259	0.497	0.73	0.818
	(0.005)	(0.006)	(0.005)	(0.004)
GPA	1.768	2.255	2.416	2.603
	(0.013)	(0.011)	(0.01)	(0.01)
AGE	22.22	22.15	23.5	27.03
	(0.031)	(0.046)	(0.066)	(0.095)
FAFSA_DEPEND_STATUS_Dependent	0.738	0.724	0.617	0.406
	(0.005)	(0.005)	(0.006)	(0.006)
FAFSA_DEPEND_STATUS_Independent	0.262	0.276	0.383	0.594
	(0.005)	(0.005)	(0.006)	(0.006)
CAA	76.01	74.99	72.74	68.19
	(0.092)	(0.121)	(0.136)	(0.187)

Note:
Colors are assigned according to where the ntile's mean value lands on the standardized empirical distribution of it's variable: (x - mean(x))/sd(x)

Standardized distribution is colored from a scale of +/-0.804

Figure 6: Variables in the 2018 FAFSA study ordered buy how much they vary across estimated quartiles of treatment effects (late covariates).

B Evaluation of Assignment Policies

In the main paper, we compare the precision of different treatment-effect estimates in terms of the average outcomes that can be achieved when we use them for targeting. In this section, we discuss estimation and inference on these average outcomes, which we use to obtain Figure 5.

Consider treatment assignment policies π that map characteristics $X = x$ to probabilities $\pi(x) \in [0, 1]$ of being treated. (This may include policies derived from treatment-effect estimates and random assignment, in which case $\pi(x) \equiv q$ with q the probability of assignment.) When treatment is assigned completely randomly (or randomly with known propensity score that only depends on X) in the existing data and X is observed, then the average outcome $\mathrm{E}[\pi(X)Y(1) + (1 - \pi(X))Y(0)]$ under this policy, the total lift $\mathrm{E}[\pi(X)(Y(1) - Y(0))]$ relative to baseline, and the average treatment effect $\frac{\mathrm{E}[\pi(X)(Y(1)-Y(0))]}{\mathrm{E}[\pi(X)]}$ of those assigned to treatment are all identified, since $\mathrm{E}[Y(1)|X] = \mathrm{E}[Y|X, T = 1], \mathrm{E}[Y(0)|X] = \mathrm{E}[Y|X, T = 0]$ are.

Focusing on the case of average outcomes as in Figure 5, we write $U(\pi) = \mathrm{E}[\pi(X)Y(1)+(1-\pi(X))Y(0)$ for the expected outcome under this policy, which is identified by $U(\pi) = \mathrm{E}[\pi(X)Y|T = 1] + \mathrm{E}[(1 - \pi(X))Y|T = 0]$ and could be estimated by its sample analogue

$$\hat{U}(\pi) = \frac{\sum_{i=1}^n T_i \pi(X_i) Y_i}{\sum_{i=1}^n T_i} - \frac{\sum_{i=1}^n (1 - T_i)(1 - \pi(X_i)) Y_i}{\sum_{i=1}^n (1 - T_i)} \tag{1}$$

$$= \frac{1}{n} \sum_{i=1}^n \frac{T_i}{\hat{p}} \pi(X_i) Y_i - \frac{1 - T_i}{1 - \hat{p}} (1 - \pi(X_i)) Y_i \quad \text{with} \quad \hat{p} = \frac{\sum_{i=1}^n T_i}{n} \tag{2}$$

similarly to Hitsch and Misra (2018), who consider the case of a known propensity score and non-stochastic assignment.

In our implementation for Figure 5, we are specifically interested in making inference on differences $U(\pi) - U(\bar{\pi}_q)$ in outcomes of a policy π that assigns students to treatment based on some rule (such as by ranking by estimated treatment effects) relative to the baseline policy $\bar{\pi}_q(x) \equiv q$ that assigns a *random* fraction q to treatment. We note that for any two policies π and $\bar{\pi}$, we have $U(\pi) - U(\bar{\pi}) = \mathrm{E}[(\pi(X) - \bar{\pi}(X))Y(1) - Y(0)] = \mathrm{E}[(\pi(X) - \bar{\pi}(X))\tau(X)]$, and for this specific baseline policy $\bar{\pi}_q$, we find $U(\bar{\pi}_q) = \mathrm{E}[Y(0)] + q\,\mathrm{E}[Y(1) - Y(0)] = \mathrm{E}[Y|T = 0] + q\,\mathrm{E}[\tau(X)]$.[4] Since $\mathrm{E}[Y|T = 0]$ is readily estimated, we therefore now focus on efficient estimation and valid inference on weighted average treatment effects $\tau_w = \mathrm{E}[w(X)\tau(X)]$, where weights can be negative. Once we have established estimation and inference for τ_w, we can estimate all quantities of interest via

$$U(\pi) - U(\bar{\pi}_q) = \tau_{\pi - \bar{\pi}_q}, \qquad U(\bar{\pi}_q) = \mathrm{E}[Y|T = 0] + \tau_1, \qquad U(\pi) = U(\pi) + (U(\pi) - U(\bar{\pi}_q)).$$

To improve efficiency and robustness of our estimate, as well as to ensure valid inference later on, we consider an augmented inverse propensity weighted (AIPW) estimator of $\tau_w = \mathrm{E}[w(X)\tau(X)]$. Specifically, we assume that we have a consistent estimate $\hat{f}(x)$ of $\mathrm{E}[Y|X = x]$, a consistent estimate

[4] When propensity scores are non-constant, estimating $\mathrm{E}[Y(0)]$ will require additional care, and can be achieved by propensity-score weighting.

$\hat{\tau}(x)$ of $\tau(x)$, and a consistent propensity score estimate $\hat{p}(x)$ of $\mathrm{E}[T|X=x]$ available. The propensity score may be assumed to be constant when units are randomized unconditionally, in which case we may want to set $\hat{p}(x) \equiv \frac{\sum_{i=1}^n T_i}{n}$ as above. We assume that $\hat{f}, \hat{\tau}, \hat{p}$ are all fitted on separate data or using k-fold cross-fitting. Writing $\hat{f}_1(x) = \hat{f}(x) + (1-\hat{p}(x))\hat{\tau}(x)$, $\hat{f}_0(x) = \hat{f}(x) - \hat{p}(x)\hat{\tau}(x)$, the AIPW estimator

$$\hat{\tau}_w^{\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^n w(X_i) \underbrace{\left(\hat{\tau}(X_i) + \frac{T_i - \hat{p}(X_i)}{\hat{p}(X_i)(1-\hat{p}(X_i))}(Y_i - \hat{f}_{T_i}(X_i))\right)}_{=\hat{\tau}^{\mathrm{AIPW}}(Y_i, T_i, X_i)}$$

This estimator is $\sqrt{n}$ consistent and asymptotically Normal under standard regularity conditions, and it is exactly unbiased when the propensity score is known and remains consistent even when treatment-effect and outcome estimates are not. We can consistently estimate its asymptotic variance by

$$\hat{\sigma}_w^2 = \frac{1}{n}\left(w(X)\ \hat{\tau}^{\mathrm{AIPW}}(Y_i, T_i, X_i) - \hat{\tau}_w^{\mathrm{AIPW}}\right)^2$$

to obtain standard error estimates $\frac{\hat{\sigma}_w}{\sqrt{n}}$ and a corresponding 95 % confidence interval $\hat{\tau}_w^{\mathrm{AIPW}} \pm 1.96 \cdot \frac{\hat{\sigma}_w}{\sqrt{n}}$.

Applying this estimator to the estimation for Figure 5, we can estimate

$$\hat{U}^{\mathrm{AIPW}}(\bar{\pi}_q) = \frac{\sum_{i=1}^n (1-T_i) Y_i}{\sum_{i=1}^n (1-T_i)} + q\ \frac{1}{n}\sum_{i=1}^n \hat{\tau}^{\mathrm{AIPW}}(Y_i, T_i, X_i)$$

$$\hat{U}^{\mathrm{AIPW}}(\pi) = \hat{U}^{\mathrm{AIPW}}(\bar{\pi}_q) + \underbrace{\frac{1}{n}\sum_{i=1}^n (\pi(X_i) - q)\ \hat{\tau}^{\mathrm{AIPW}}(Y_i, T_i, X_i)}_{=\hat{\Delta}(\pi)},$$

$$\widehat{\mathrm{SE}}\left(\hat{U}^{\mathrm{AIPW}}(\pi) - \hat{U}^{\mathrm{AIPW}}(\bar{\pi}_q)\right) = \sqrt{\frac{1}{n^2}\sum_{i=1}^n \left((\pi(X_i) - q)\ \hat{\tau}^{\mathrm{AIPW}}(Y_i, T_i, X_i) - \hat{\Delta}(\pi)\right)^2}.$$

We note that, in particular, the estimation outcome of the random policy does not include any additional randomization, which would only add noise to the estimation. Further, we obtain the estimator in Equation 1 of $U(\pi)$ when we set $\hat{p}(x) \equiv \frac{\sum_{i=1}^n T_i}{n}, \hat{\tau}(x) \equiv 0, \hat{f}(x) \equiv 0$, which is generally inefficient.

So far, we have considered fixed policies π. However, in our application, policies are themselves estimated, such as the policy

$$\hat{\pi}_{\hat{t}}(x) = \mathbb{1}(\hat{\tau}(x) \geq \hat{t})$$

where treatment effects $\hat{\tau}(x)$ and the cutoff $\hat{t}$ chosen to achieve a given proportion q in treatment are all noisy. We use cross-fitting to avoid biases from estimation in this case. Specifically, we estimate the quantities of interest separately on each fold, using rankings estimated based on the other folds only, and then average aggregate across all folds. Under regularity conditions is the covariance between estimates from different folds of a lower order than the variance we estimate, allowing us to combine estimates and variance estimates across folds to obtain valid inference.