The Potential for Improved Implementation of Microfinance through AI-Driven Targeting

Nicholas Lacoste

June 8, 2024

Abstract

Policy interventions – especially antipoverty – often provide social returns which are either disappointing in magnitude or cost-effectiveness. In many cases, the core of this issue may be attributed to treatment effect heterogeneity. That is, some program recipients exhibit high marginal benefit per-dollar of program cost while others are perhaps better served by alternative interventions. Effective program targeting based partially on desired impact is thus a critical addendum to typical need-based targeting. I apply modern machine learning (ML) approaches to assess the scope for improving microfinance – a large-scale antipoverty program with historically mixed returns – by targeting on predicted treatment effect. First, I utilize a set of RCTs to establish benchmark program impacts: confirming that average returns are small for a generic set of first- and second-order outcomes. Next, I develop a framework for employing the generalized random forest estimator of individualized treatment effects to perform subsample analysis by "counterfactually" reallocating treatment among the upper quantiles of the distribution. Consistent with large heterogeneity, I find that "gains" to impact targeting are often substantial: the average impact among the targeted subset is between \$60 – 480 (90 – 567%) larger for bi-weekly business profits, revenues, and expenditures and between \$147 – 711 (664 – 1,373%) larger for bi-weekly household income in case studies with high program take-up. Finally, I derive simplified targeting rules using GMM-matched policy trees on recipient characteristics in order to illustrate some "quick fixes" to the existing policy structure. Results suggest meaningful improvements may exist from a simple reallocation of treatment in the microfinance setting and other settings with a wide degree of heterogeneity.

1 Introduction

Empirical research which evaluates the impact of an event – be it a policy, randomized intervention, or natural experiment – typically focuses on identifying the average causal effect among the treated population. When researchers estimate non-transformative or noisy treatment effects, it is natural to conclude that the intervention has failed. This problem can be especially salient in instances where intervention costs per-recipient are high, such as social safety nets or wealth transfer programs. However, even in interventions which appear ineffective on average, there are often significant subsets of individuals who benefit from program receipt. This heterogeneity is thus a critical consideration when deciding program allocation: it is likely welfare-maximizing to continue treating those with large, positive impacts while simultaneously redirecting resources towards alternative programs for those who are unaffected (or negatively affected) by the treatment in question. The setting of economic development is one such area where effective treatment allocation is both critical and difficult to achieve, especially in the case of anti-poverty programs (Coady

et al., 2004; Ravallion, 2009). The targeting challenge policy administrators typically face is determining who is the "most deprived" in an efficient manner such that the "poorest of the poor" are more likely to be treated. There is a large accompanying body of literature outlining how governments or non-governmental organizations may target antipoverty interventions based on likely poverty status, using strategies such as proxy means tests or community targeting (Alatas et al., 2012), ordeal mechanisms (Alatas et al., 2013), "big data" (Blumenstock et al., 2015), or other. However, only recently have researchers begun to consider the merits of targeting based on impact magnitude, and it has been shown that (at least partially) targeting cash transfer programs towards those individuals where the marginal benefit of program receipt is higher is a welfare-maximizing strategy (Haushofer et al., 2022). This question of identifying high-impact individuals for optimal treatment allocation is a different question entirely from that of identifying the most in-need, and is an area where modern machine learning (ML) techniques are likely to lend themselves well as they provide attractive methods to non-parametrically characterize treatment effect heterogeneity in a disciplined manner (Athey and Imbens, 2017).

In this paper, I assess the scope for applying ML techniques to improve the efficiency and effectiveness of microfinance[1] by targeting based on predicted impact. Over the past 40 years, microfinance has become one of the most widespread anti-poverty programs in the world. Since the pioneering Grameen bank began offering small, low-interest loans to entrepreneurial women in 1983, microfinance institutions (MFIs) have reached over 140 million low-income clients, with outstanding loans of over $124 billion USD as of 2019 (Barometer, 2019). Despite being widely prevalent and well-studied, microfinance has a mixed track record for delivering meaningful improvements to business income, household income, consumption, or other measures relevant to welfare. In fact, the program has produced few positive average treatment effects for most relevant outcomes when considering entire samples of recipients across a host of settings, as outlined by Banerjee et al. (2015). However, economic intuition suggests that microcredit may create winners and losers. Those individuals likely to benefit may be individuals with greater capacity to repay loans and/or utilize loans in a profitable manner, whereas other individuals could be negatively impacted if (for example) they are more likely to default or are forced to prioritize repaying unprofitable debt. There is evidence for this possibility, as shown by Meager (2022), who used Bayesian hierarchy methods to characterize the treatment effects of microfinance across the income distribution and found that only those in the top 25% had positive treatment effects (while the bottom 75% had precisely null effects).

[1] I refer only to the micro-loan aspect of microfinance in this paper. Therefore I use the terms "microfinance" and "microcredit" interchangeably.

Between 2006 – 2007, Karlan and Zinman (2011) conducted a randomized controlled trial of microcredit access in the Philippines. Around the same time, six additional RCTs were conducted and published in a special issue of the *American Economic Journal: Applied Economics* assessing the effects of microcredit access in six additional developing countries (Angelucci et al., 2015; Attanasio et al., 2015; Augsburg et al., 2015; Banerjee et al., 2015; Crépon et al., 2015; Tarozzi et al., 2015). These six RCTs were conducted independently, but studied a coordinated set of outcomes. Due to the open data policies of *AEJ* and *Science*, all data from these studies are publicly available and are the primary data used in my analysis. These experimental datasets provide several key advantages to evaluating the potential gains from impact-targeting. Firstly, their randomization in credit access allows me to establish a unified set of first- and second-order outcomes across each experimental setting upon which to generate internally valid estimates for both the larger populations and well-defined population subsets. Second, their coordination generally prompted the authors to collect a rich set of baseline covariates which I can exploit to train targeting algorithms. Finally, their geographic and socio-economic diversity showcase the relative impacts of microfinance, and by proxy *targeting* the allocation of microfinance, across a variety of income and spatial distributions.

In order to quantify potential gains from impact targeting, my primary analysis proceeds in four steps. I begin by leveraging the experimental variation in credit access to establish benchmark *average* treatment effects upon my derived set of outcomes: business profits, revenues, and expenditures as well as household income and consumption. I generally confirm the findings of the original experiments: impacts of microcredit access are small and imprecisely estimated for most outcomes in most experimental settings.

Second, I employ new developments in causal machine learning (ML) to create "counterfactual" targeting algorithms. I define a partitioning algorithm within each experimental sample and use it to estimate individualized treatment effects with *generalized random forests* (GRFs) (Athey et al., 2019). ML techniques are not specifically designed to identify treatment effect parameters. Rather, they excel at prediction in cases where the set of covariates is large and the estimators must be modified in order to achieve the desirable properties of asymptotic normality and consistency (Mullainathan and Spiess, 2017). Consequently, there is a growing literature around the development of algorithms which produce consistent estimates of low-level treatment effects, particularly in "tree-based" models with the creation of the "casual tree" (Athey and Imbens, 2016), the extension "causal forest" (Wager and Athey, 2018), and most recently the generalized random forest (Athey et al., 2019) which is designed for observational settings with instrumental variables or experimental settings with imperfect compliance (among several other uses). In the case of heterogeneous treatment effects, GRFs may be used within semi-parametric and non-parametric regression to estimate *con-*

ditional average treatment effect (CATE) functions in a disciplined manner without relying on subsample analysis which can miss important covariate interactions and/or run into multiple hypothesis testing issues (Angrist, 2004).

Third, upon establishing predictions of individualized treatment effects and following additional guidelines by Chernozhukov et al. (2018) for a characterization of the within-sample heterogeneity, I identify "targeted" subsets within each population and reassess program impacts within these reallocated treatment groups. I document often substantially larger impacts among these populations, especially in case-studies with high program take-up and (therefore) more statistical power. The average impact among the targeted group is between \$60 – \$480 (90% – 567%) higher for bi-weekly business profits, revenues, and expenditures in terms of 2009 USD PPP. Similarly, in these high take-up studies, gains to bi-weekly household income range from \$147 – \$711 (664% – 1,373%), while gains to household consumption remain mixed. Conversely, in the studies with lower take-up rates, while targeting gains are often suggestively (sometimes meaningfully) higher, precise benchmark and/or counterfactual estimates are largely unachievable due to lack of power – an issue which plagued the original studies as well. The data used in my study are not particularly large for ML standards. Because of this, the targeting algorithms were most often successful when the "threshold" for a targeted individual was flexible. That is, the algorithms were not generally able to accurately predict the treatment effects themselves but were able to accurately rank individuals such that gains could be achieved by assessing the distribution of treatment effects and filtering out the lower quantiles.

Finally, while the data-driven subset analysis in Step 3 is meant to be illustrative of the average impact realized among the top of the treatment effect distribution, it cannot itself be considered a "gain." First, recent work has suggested that policy derivation based on "winners" is inherently limited by the degree to which the positive outcomes of these winners were overestimated based on chance (Andrews et al., 2024). Second, replicating an analysis such as this in the real world with observational data comes with implementation and administrative issues which may preclude a targeting scheme based purely on an algorithm. Therefore, the final aspect of my main analysis mobilizes the heterogeneity uncovered by Step 3 to derive new, simple targeting rules via policy trees based on doubly robust scores (Athey and Wager, 2021). These shallow-depth decision trees utilize covariates associated with large treatment effects to create "optimal" targeting rules and illustrate quick-fix reallocations rooted in the findings of the algorithm.

There are several additional questions which are important to a practical implementation of program targeting I explore. First, an MFI must decide upon which outcome to determine a treatment effect threshold

– i.e. should an MFI target based on likely business impact or on "downstream" outcomes such as consumption or income? While I cannot answer this question, it is key to understand the trade-offs between who is selected under an algorithm targeting a certain outcome vs. who is selected under a different outcome. Comparing the magnitudes of predicted treatment effects and the selected sample under different algorithms, I find that there are often both significant trade-offs and complementarities in the targeted populations based on the selected outcome. Additionally, I find that the "targeted" subset is not always more likely to borrow if offered a loan than the "untargeted" subset. This implies that some resources should likely be reallocated towards marketing in order to ensure that those who have large marginal benefits will actually borrow.

The final aspect of my analysis considers alternative algorithms to the generalized random forest for the prediction of individualized treatment effects. While the GRF provides numerous advantages other algorithms fail to provide (e.g. beneficial asymptotic properties of CATE estimates which allow for the estimation of standard errors, moment matching capabilities for policy trees, etc.), any predictive model may be modified to predict treatment effects and, in some cases, may actually *perform* better at predicting outcomes. I repeat Steps 2 and 3, replacing the GRF estimator with Support Vector Machines (SVMs) and OLS, the former being one of the ubiquitous "black box" prediction algorithms found commonly in the computer science literature. I find that OLS performs very well in the high take-up studies but fails dramatically in studies with low statistical power. Meanwhile, SVM performs similarly to the GRF, though generally produces smaller gains to average program impact.

The outline of this paper is as follows. Section 2 briefly discusses the microfinance industry and its' status as an anti-poverty program in the context of the experimental trials I use. I also discuss the RCTs I utilize in this analysis and their pros and cons for machine learning applications via their micro data. Section 3 outlines the empirical methods I use to set benchmark treatment effects in the form of intent-to-treats (ITTs), predict individual responses and apply counterfactual targeting, and derive policy trees. Section 4 discusses alternative predictor algorithms and supplementary analyses. Section 5 concludes.

2 Experimental Background and Data

2.1 Study Settings

In the 1990's, the microcredit industry had begun to transition into its "second generation." In the past, microcredit organizations essentially functioned as local lending groups where low-income individuals

– almost exclusively women in rural areas – pooled their funds into joint-liability loans to one another (Armendariz and Morduch, 2010). The transition into the second generation of microfinance saw the program become increasingly akin to more traditional credit markets. This included the introduction of more formal banks (both for-profit and non-profit) and microfinance institutions (MFIs) additionally extending credit access to men and in increasingly urban areas, often with individual-liability loan structures (Cull et al., 2009). At the time, there was great optimism around microfinance as an antipoverty program, with the belief that it could empower women, spur economic growth, and mitigate market failures by extending credit access to entrepreneurs who lacked collateral or credit history (Karlan and Zinman, 2011). However, there was little evidence for some time to support or contradict the dialog surrounding these programs.

Between 2006-2007, Karlan and Zinman (2011) conducted the first randomized control trial aimed at assessing the impact of microcredit access on various measures of welfare in the Philippines. This initial study attempted to evaluate the merits of the microcredit product to deliver on its' promises. Specifically, they evaluated if microcredit alleviated market failures, spurred business growth, improved subjective well-being, empowered women, substituted for insurance, and/or improved informal risk sharing by examining a set of outcomes which plausibly act as indicators for these definitions of welfare. Similarly, in January, 2015, the *American Economic Journal: Applied Economics* published a special issue highlighting an additional six RCTs[2] which were conducted around the same time as the Philippines study. These six studies took place in a coordinated fashion across six countries and five continents, each with the goal of assessing the impact of microcredit access on a similar set of welfare-related outcomes – though each aimed to provide a unique contribution to the microfinance literature. Due to the open data policies of *Science* and the *AEJ*, these experimental datasets are readily available for usage and five of the seven are the primary data I use in this analysis.

There are several key advantages which come from using this set of experimental datasets. First, due to their randomized nature, they allow for the credible estimation of benchmark treatment effects upon which I may attempt to improve using targeting algorithms. Second, because of their coordinated effort, most of them collected a full set of baseline and endline micro-level covariates to utilize in predicting treatment effects ex-ante and deriving a unified set of endline outcomes which can be evaluated across studies. Third, taken as a whole, they showcase outcomes which are representative of the global microfinance industry given their geographic scope, varying macroeconomic settings, targeting methods, loan structures, etc.

[2]Augsburg et al. (2015) – Bosnia and Herzegovina, Tarozzi et al. (2015) – Ethiopia, Banerjee et al. (2015) – India, Angelucci et al. (2015) – Mexico, Attanasio et al. (2015) – Mongolia, and Crépon et al. (2015) – Morocco

2.2 Original Experimental Designs

Each of the original studies worked in collaboration with a local MFI to generate randomized variation in credit access among a marginally creditworthy population. Specifically, five[3] of the seven experiments involved the MFI expanding to a new geographic area (or several geographic areas) which was determined to be a locality with a high concentration of probable microfinance customers. Within this experimental region, the MFI began offering their credit product exclusively in treatment villages or neighborhoods while refraining from offering it in control villages. The other two experiments[4] instead generated randomized variation in credit access by developing a rudimentary credit scoring system which identified the "marginally creditworthy" and randomized credit access among *individuals* who applied for a loan, conditional on the credit score obtained. There are several pros and cons to each type of randomization unit. The biggest benefit of randomizing credit access at the geographic level is that any spillover or general equilibrium effects which occur are internalized by the estimates obtained by regressing on endline outcomes. However, this type of randomization comes with a major limitation in that take-up rates are very low when one considers (nearly) the entire population as eligible, which yields low statistical power. Meanwhile, randomization at the individual-level yields much higher take-up because the experimental sample includes individuals who already expressed interest in credit, but conversely these experiments are not well-suited for internalizing spillovers.

Beyond the obvious requirement that each experiment I utilize should test the same treatment on a similar set of outcomes, I also require that each dataset contain a set of individual-level covariates such that I may predict treatment effects using baseline survey data. Two of the seven studies – India and Ethiopia – were unable to collect any sort of individual-level panel and are thus excluded from my analysis. I refer the reader to the well-done summary paper, Banerjee et al. (2015), and Karlan and Zinman (2011) for a complete characterization of the loan structures, borrower/lender characteristics, study environments, etc. of the remaining five papers. However, I do wish to highlight some statistics taken from those papers which are relevant to this study, displayed in Table 1. As we can see, the study settings encompass several factors which may impact the average effectiveness of the program and vary across experiments. For example, of the five studies, three offer more traditional group-liability loans, two offer "second generation" individual-liability loans, and one (Mongolia) tries both in a two-arm experiment. It is also important to note the time from treatment to the endline survey. The original papers (as well as my analysis) consider "downstream" outcomes in a reduced-form manner as primary measures of utility/welfare. However, while there is some

[3]Ethiopia, India, Mexico, Mongolia, Morocco
[4]Bosnia, Philippines

variation in the duration of the experimental window (ranging from about 1-2 years) across studies, they typically lack consistent re-interviews which would better capture short vs. medium vs. long-run impacts[5]. Nonetheless, all treatment effects estimated come with the caveat that the impacts measured on any particular outcome are a function of the window under which the individual was re-interviewed.

Some additional minor notes to discuss about a few of the datasets:

1. **Mexico**: This experiment attempted to create a very large panel sample, however improper implementation of the baseline survey largely prevented this. The original analysis used the resulting "endline-only" sample of approximately 16,000 individuals. However, in my analysis I cannot use this larger sample because it lacks the baseline covariates I require to train machine learning targeting models. The authors were able to procure a smaller panel sample of around 1,800 individuals. This panel sample is the estimating sample in my specifications regarding this experiment. Unfortunately, take-up rates are extremely low in the Mexico study, especially so in the panel sample, which dramatically hinders precision.

2. **Mongolia**: In this particular experiment, the authors were somewhat interested in the relative impacts of microcredit when the loans are "group-liability" vs. "individual-liability." The canonical microcredit product is a group-liability loan, but at the time many microlenders were changing this practice to include individual-liability. For this reason the authors conducted a 2-arm experiment: one where loan access was group-liability and another where the loan access was individual. However, results were not very interesting among the individual-liability group, so the original authors focused their main analysis only on the group-liability sample. For my purposes, I do not follow this. With machine learning, more data is typically necessary for the consistency of pattern recognition, and incorporating loan characteristics such as the type of liability into the algorithm is straightforward. For this reason, my benchmark results will differ slightly from the original paper's because I incorporate both treatment arms into the estimating sample.

2.3 Existing Results

The overall punchline when distilling results across these studies is that treatment effects are typically non-transformative in their magnitude and are often imprecisely estimated with a significant amount of likely heterogeneity. The issues largely stem from a lack of statistical power and low take-up rates, especially in the

[5]There were, of course, budgetary limitations which prevented the administration of multiple phases of endline surveys and equity concerns which prevented the prolonged existence of a control group.

Table 1: Selected Lender, Lendee, and Loan Characteristics

Study:	Bosnia and Herzegovina	Mexico	Mongolia	Morocco	Philippines
Household annual income (PPP USD)	$19,780	$7,828	$1,620	$5,059	$4,582
Lender	Unknown	Compartamos Banco	XacBank	Al Amana	First Macro Bank
Lender Org Type	N/A	For-profit	For-profit	Nonprofit	For-profit
Unit of randomization	by individual	by village	by village	by village	by individual
Liability type	individual	group	both	group	individual
Number of groups	N/A	238	40	162	N/A
Sample size at endline	995	16,560*	964	5,551	1,601
Time from treatment to endline	14 months	16 months	19 months	24 months	between 11-22 months
Loan term length	14 months (avg)	4 months	6 months (avg) group, 8 months (avg) individual	16 months (avg)	13 weeks
Interest rate	22% APR	110% APR	26.8% APR	14.5% APR	60% APR
Market interest rate	27.3% APR	145% APR	42.5% APR	46.3% APR	Not specified
Loan size (PPP USD)	$1,816 (avg)	$451 (avg)	$696 (avg) group, $472 (avg) individual	$1,082 (avg)	$220 (median)
Gender of recipients	Male/Female	Female	Female	Male/Female	Male/Female
Microentrepreneurs?	Yes	Yes	Yes	Yes	No

Notes: Statistics were taken from Banerjee et al. (2015) and from Karlan and Zinman (2011), therefore all dollar amounts are in 2015 USD PPP. Each of these papers has more detailed information on lenders and lender requirements, survey timing, applicant characteristics, etc.
* This is the size of the estimating sample the original authors used. However, for my case study I am limited to a smaller panel sample of 1,823 individuals due to baseline survey implementation issues with the original experiment.

three studies which randomized across geographic strata. The original papers concerned themselves primarily with estimating intent-to-treat (ITT) effects on a large set of "outcome families" which included a host of welfare-related measures including labor supply, business ownership, business success, income, consumption, health, female empowerment, etc. As explained in Banerjee et al. (2015), each study estimated between 36 – 82 unique treatment effects, with an average of 50. Of these, the share of impact estimates which were statistically significant ranged from just 6-39%. I again refer the reader to the summary paper for a full discussion of the original experiments.

3 Generalized Random Forests for Loan Targeting

The primary analysis proceeds in the following subsections:

1. Establish benchmark *average* causal effects in the form of intent-to-treats (ITT) for a set of unified and standardized outcomes for each of the five experimental datasets.

2. Train *causal forest (CF)* and *generalized random forest (GRF)* machine learning algorithms – designed to predict *individualized* treatment effects.

3. Estimate counterfactual *average* casual effects by restricting the estimating sample to subsets of individuals predicted to have larger impacts.

4. Derive simplified "decision rule" targeting policies based on the heterogeneity uncovered in Step 2.

3.1 Establishing Benchmark ITTs

As previously stated, each of the original studies estimated a large number of treatment effects. While each of the outcomes they considered are important in their own right, it is both practically and computationally infeasible to focus on them all. Therefore, I seek a smaller set of outcomes which satisfy some basic criteria: they are comparable across studies, they are relevant as potential target outcomes to a microfinance organization, and they are leading indicators of the other "downstream" outcomes one may be interested in. In each RCT, I derive (or directly extract from the data if available) five primary outcomes which are broad indicators for economic welfare and for which I set benchmark estimates: business profits, business revenues, business expenditures, household income, and household consumption. I often refer to the first three outcomes as the "business success" or simply the "business" outcomes and the last two as the "household" outcomes. Typically, the original papers estimated treatment effects somewhat related to these variables. For example, most of the papers estimated some ITT(s) related to self-employment outcomes,

which are similar to the business success measures.

Now, the area where I did *not* differ from the original papers is in the estimation procedure itself. Attanasio et al. (2015) explains that the decision between estimating treatment effects via an OLS regression on endline outcomes with a binary treatment indicator vs. difference-in-differences (DID) came down to statistical power: if autocorrelation within outcomes is small, then DID likely will limit statistical power. It was typically the case that autocorrelation values were below 0.5, thus each paper opted for the former specification. When extrapolating the original analyses to the my unified outcome set, I retain the authors' original estimation procedures as the treatment randomization permits me to do so. To illustrate an example from one of the studies where credit access was randomized across villages, the specification from Mongolia took the following form:

$$Y_{iv1} = \alpha + \tau W_{iv} + \beta \boldsymbol{X}_{i0} + Y_{iv0} + \varepsilon_{i1} \tag{1}$$

Where Y_{iv1} represents the endline outcome for individual i in village v, W_{iv} is an indicator equal to 1 if the individual is in a village with credit access, τ is therefore the ITT effect of interest, Y_{iv0} represents the outcome for individual i collected in the baseline survey, and $\boldsymbol{X}_{i0}$ is a vector of individual baseline controls. In some cases, strata fixed effects are included in equation (1), and standard errors are always clustered at the strata level. The specifications relevant to the two studies which randomized credit access among individuals (Bosnia and Philippines) are quite similar, the differences being that the treatment indicator, W_i, takes on a value of 1 if the individual was approved a loan conditional on them being in the sample of eligible applicants (that is, if they were determined "marginally creditworthy") and that standard errors are heteroskedasticity-robust rather than clustered. It is also key to note that treatment effects, while technically still interpreted as ITTs in the case of Bosnia and Philippines because compliance is imperfect in both directions (i.e. 97-100% of treated individuals complied, but a decent share of control individuals were able to obtain loans elsewhere), are much closer to treatment-on-the-treated (ATT) effects and may very cautiously be interpreted as such for intuition.

Also note that each of the original papers presented results using the currency of the local area. For my purposes I do not follow this because treatment effects across studies are difficult to compare and conceptualize due to varying exchange rates. Thus, my primary results are all presented using standardized currency variables representing 2009 USD PPP. Flow currency variables are converted to bi-weekly 2009 USD PPP. This process follows the advice of Meager (2022). For completeness, I include the analysis conducted with raw currency in Appendix Section C such that I have estimates which more closely match what the original

papers reported.

I display "first stage" results – the effects of credit access on borrowing propensity – in Appendix Tables A.1. Note that these tables are merely replications of the original studies, each of whom estimated these impacts prior.[6] Generally speaking, take-up differentials between the treatment and control villages are small in the studies which randomized credit access among geographic areas. In the cases of Mexico and Morocco, the take-up differential ranged from about 6pp to about 14pp, depending on the indicator of take-up used. While the differential was larger (47.5pp) in Mongolia, there is evidence of substitution in that control individuals were often found to obtain loans from other sources.

Table 2 displays the benchmark ITT estimates on my primary downstream outcomes. Generally speaking, the effects are small and often imprecisely estimated. In some cases, such as the Philippines case-study, they are even suggestively negative for several outcomes. These results are consistent with the findings of the original studies in both magnitude and precision (Banerjee et al., 2015).

3.2 Training Generalized Random Forests for CATE Mappings

3.2.1 GRF Implementation Process

The next step in the analysis is to produce a mapping of the conditional average treatment effect (CATE) function. That is, I seek a framework which can produce consistent *predictions* of individualized treatment effects such that I may perform counterfactual subgroup analysis along the distribution of predicted individual impacts. The primary algorithm I employ for this process is the generalized random forest (GRF) (Athey et al., 2019). The GRF is an adapted random forest which nonparametrically estimates a continuous mapping of covariates to outcomes such that point estimates within that mapping have the necessary asymptotic properties of consistency and normality – provided there is a valid identification strategy underlying the estimation of average treatment effects (ATT's).

Consider a structural model which determines outcomes for individual i, where the treatment effect of

[6] Also note that Bosnia and Philippines have near-100% take-up given their experimental designs. The "take-up" indicator used in these first-stage regressions is the differential of outstanding loans at the endline survey.

Table 2: Benchmark ITTs - Outcomes of Interest

Panel A: Bosnia

	Business profits (USD 2009 PPP)	Business revenues (USD 2009 PPP)	Business expenditures (USD 2009 PPP)	Household income (USD 2009 PPP)	Household consumption (USD 2009 PPP)
Treatment	34.131	71.138	31.526	10.777	−5.603
	(24.531)	(43.945)	(25.955)	(36.790)	(19.064)
Control Mean	132.58	198.06	73.4	687.55	196.42
R^2	0.028	0.030	0.028	0.106	0.059
Num. obs.	994	994	994	994	994

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Panel B: Mexico

	Business profits (USD 2009 PPP)	Business revenues (USD 2009 PPP)	Business expenditures (USD 2009 PPP)	Household income (USD 2009 PPP)	Household consumption (USD 2009 PPP)
Treatment	−1.76	23.36	25.67*	−2.76	−8.56
	(15.44)	(18.96)	(14.70)	(16.64)	(13.69)
Control Mean	18.36	55.4	37.66	283.89	299.31
R^2	0.00	0.02	0.02	0.00	0.00
Num. obs.	1823	1823	1823	1684	1623

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Panel C: Morocco

	Business profits (USD 2009 PPP)	Business revenues (USD 2009 PPP)	Business expenditures (USD 2009 PPP)	Household income (USD 2009 PPP)	Household consumption (USD 2009 PPP)
Treatment	10.937	12.520*	40.344**	−25.032	−1.828
	(10.952)	(7.147)	(17.920)	(158.421)	(5.665)
Control Mean	69.53	48.54	169.92	2821.88	314.73
R^2	0.069	0.058	0.108	0.091	0.180
Num. obs.	3614	3614	3614	3614	3605
N Clusters	162	162	162	162	162

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Panel D: Mongolia

	Business profits (USD 2009 PPP)	Business revenues (USD 2009 PPP)	Business expenditures (USD 2009 PPP)	Household income (USD 2009 PPP)	Household consumption (USD 2009 PPP)
Treatment	−0.63**	0.17	0.85**	9.75	70.58
	(0.28)	(0.20)	(0.32)	(29.23)	(46.88)
Control Mean	-0.91	1.44	2.35	66.93	280.09
R^2	0.01	0.06	0.06	0.02	0.12
Num. obs.	961	961	961	955	961
N Clusters	40	40	40	40	40

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Panel E: Philippines

	Business profits (USD 2009 PPP)	Business revenues (USD 2009 PPP)	Business expenditures (USD 2009 PPP)	Household income (USD 2009 PPP)
Treatment	64.942	−85.432	−71.275	−107.108
	(55.302)	(181.491)	(151.189)	(196.402)
Control Mean	380.11	1525.29	1110.94	1778.12
R^2	0.037	0.027	0.026	0.041
Num. obs.	1113	1113	1113	1078

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: Treatment effects are all represented in 2009 USD PPP over a 2-week period. The regression specifications are all estimated via OLS, regressing endline outcomes on an indicator for treatment status. Controls are included exactly as specified in the original papers and in the case of Mongolia I additionally control for loan liability type. Standard errors are either clustered at the unit of randomization or are heteroskedasticity-robust if the unit of randomization was the individual.

loan receipt on outcome Y_i is allowed to vary continuously along a vector of observable baseline covariates:[7]

$$Y_i = \tau(\boldsymbol{X}_i)W_i + f(\boldsymbol{X}_i) + \varepsilon_i \tag{2}$$

$$\tau(\boldsymbol{X}_i) = E[Y_i(1) - Y_i(0)|\boldsymbol{X}_i = \boldsymbol{x}] \tag{3}$$

Where $\tau(\boldsymbol{X}_i)$ is the conditional average treatment effect: a function representing the difference in potential outcomes for individual i, conditional on their baseline covariates.

The traditional random forest algorithm (Breiman, 2001) is designed to estimate the conditional mean function: $m(\boldsymbol{x}) = E[Y_i|\boldsymbol{X}_i = \boldsymbol{x}]$. It achieves this by building B regression trees which greedily and recursively split the data along covariates, $x_k \in \boldsymbol{X}$, and thresholds, c, to maximize the squared-difference in means across the resulting partitions (a.k.a. child nodes), denoted $\bar{Y}_{C_p}$:[8]

$$\max_{x_k, c} \quad n_{C_1} n_{C_2} (\bar{Y}_{C_1} - \bar{Y}_{C_2})^2$$

Each tree is grown on a bootstrapped sub-sample of the training data and a random sub-sample of the covariate vector, $\boldsymbol{X}$. The conditional mean function is then estimated by (1) averaging outcomes, Y_i, of observations which fall in the same terminal leaf (L_b) as a given test point, $\boldsymbol{x}$, and then (2) averaging over trees:

$$\hat{m}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{|L_b(\boldsymbol{x})|} \sum_{i=1}^{n} Y_i \cdot \mathbb{1}\{\boldsymbol{X}_i \in L_b(\boldsymbol{x})\}$$

Wager and Athey (2018) show that we can express $\hat{m}(\boldsymbol{x})$ as a weighted-average of all observations, where the weights, $\alpha_i(\boldsymbol{x})$, are the share of total trees where i falls into the same terminal leaf as $\boldsymbol{x}$:

$$\hat{m}(\boldsymbol{x}) = \sum_{i=1}^{n} Y_i \alpha_i(\boldsymbol{x})$$

$$\alpha_i(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}\{\boldsymbol{X}_i \in L_b(\boldsymbol{x})}{|L_b|}$$

The generalized random forest (Athey et al., 2019) modifies the objective function of the random forest such that the algorithm searches for the covariate split that maximizes the squared difference in ATTs across

[7] Equation (2) is not simply a relaxation of equation (1), as equation (1) identifies the ITT which is the LATE multiplied by the share of compliers. This is a different structural model defining the heterogeneous treatment effect. In the case of an RCT with perfect compliance, equation (2) amounts to a relaxation of equation (1) because (1) would identify the ATT, and (2) identifies the conditional ATT. But in the case of an IV, they are fundamentally distinct as $\tau(\boldsymbol{x})$ in (2) represents the conditional local average treatment effect, not the conditional ITT.

[8] "Greedy" splitting refers to the process of selecting the optimal split relative to the current node only: never selecting suboptimal splits or reverting back to prior splits. "Recursive" splitting refers to continuous splitting along the parent node, then subsequently along child nodes, and so on until stopping criteria are met.

the resulting partitions:

$$\max_{x_k, c} \quad n_{C_1} n_{C_2} (\bar{\tau}_{C_1} - \bar{\tau}_{C_2})^2$$

We identify the CATE function in the GRF by finding the GMM solution to a (local) moment condition implied by the identification strategy. In the case of an RCT with perfect compliance, the relevant population moment is conditional exogeneity of treatment:

$$\begin{aligned} & E[W_i \varepsilon_i | \boldsymbol{x}_i] = 0 \\ \implies & E[W_i (Y_i - W_i \tau(\boldsymbol{x}_i)) | \boldsymbol{x}_i] = 0 \end{aligned}$$

At the parent node, P, all observations are weighted equally because no splits have been made. Therefore, the GMM solution is just the sample analog of the above moment where weights = 1, which corresponds to the simple OLS estimate of the average treatment effect:

$$\begin{aligned} & E[W_i Y_i - W_i^2 \tau(\boldsymbol{x}_i)) | \boldsymbol{x}_i] = 0 \\ \implies & \frac{1}{n} \sum_i W_i Y_i - \frac{1}{n} \sum_i W_i^2 \tau(\boldsymbol{x}_i) \\ \implies & \hat{\tau}_P = \frac{\sum_i W_i Y_i}{\sum_i W_i^2} \end{aligned}$$

In order to estimate the subgroup ATTs within potential child nodes in a computationally tractable manner, the GRF uses a first-order approximation:

$$\hat{\tau}_{C_p} \approx \hat{\tau}_P - \frac{\sum_{i \in C_p} W_i (Y_i - W_i \hat{\tau}_P)}{\sum_{i \in C_p} W_i^2}$$

After growing all the trees in the forest, the GRF then creates a mapping of the CATE function by considering any point, i, with covariates, $\boldsymbol{X}_i = \boldsymbol{x}$, and then re-estimating the OLS version of the CATE (the one from the parent node), adding the forest-defined weights to each observation, $\alpha_i(\boldsymbol{x})$:

$$\hat{\tau}(\boldsymbol{x}) = \frac{\sum_i^n \alpha_i(\boldsymbol{x}) W_i Y_i}{\sum_i^n \alpha_i(\boldsymbol{x}) W_i^2}$$

I also note a few implementation details. First, in practice, the GRF actually estimates the CATE using residualized (a.k.a. recentered) observations to improve prediction quality. That is, it estimates the conditional mean function and the propensity score function: $m(\boldsymbol{x}) = E[Y_i | \boldsymbol{X}_i]$ and $e(\boldsymbol{x}) = E[W_i | \boldsymbol{X}_i]$, using

separate random forests[9] in order to de-mean the observations before estimating the CATE:

$$\hat{\tau}(\boldsymbol{x}) = \frac{\sum_i^n \alpha_i(\boldsymbol{x})\tilde{W}_i\tilde{Y}_i}{\sum_i^n \alpha_i(\boldsymbol{x})\tilde{W}_i^2} \tag{4}$$

$$\text{Where:} \quad \tilde{Y}_i = Y_i - \hat{m}(\boldsymbol{x}); \quad \tilde{W}_i = W_i - \hat{e}(\boldsymbol{x})$$

The second implementation note is that when building the B "causal trees," two disjoint subsamples of the data are used: one subsample follows the aforementioned splitting process to define the tree structure, and the other subsample fills the tree nodes to produce $\hat{\tau}$ values. This is known as "honest" training and is a key component for yielding consistent and asymptotically normal point estimates.

The final implementation detail addresses noncompliance within the RCT samples I am using. Because take-up rates are low in the three studies which randomized credit access among villages (and because some control individuals were able to obtain loans in the two studies randomizing among applicants), the conditional exogeneity assumption of treatment itself is violated, and instead the policy acts as an instrument, Z. In this case, I seek the conditional *local* average treatment effect (CLATE) as expressed in Abadie (2003):

$$\tau(\boldsymbol{x}) = \frac{E[Y|\boldsymbol{X}, Z=1] - E[Y|\boldsymbol{X}, Z=0]}{E[W|\boldsymbol{X}, Z=1] - E[W|\boldsymbol{X}, Z=0]}$$

The relevant population moment to be solved locally via GMM is therefore:

$$\begin{aligned} & E[Z_i\varepsilon_i|\boldsymbol{X}_i = \boldsymbol{x}] = 0 \\ \implies & E[Z_i(Y_i - W_i\tau(\boldsymbol{x}) - f(\boldsymbol{x}))|\boldsymbol{X}_i = \boldsymbol{x}] = 0 \end{aligned}$$

Where the estimate $\hat{\tau}(\boldsymbol{x})$ (and the nuisance function $\hat{f}(\boldsymbol{x})$) are the soutions to the sample analong of the above moment:

$$\sum_i^n \alpha_i(\boldsymbol{x})\Big[Z_i\Big(Y_i - \hat{f}(\boldsymbol{x}) - \hat{\tau}(\boldsymbol{x})W_i\Big)\Big] = 0 \tag{5}$$

Recall that the purpose of this exercise is to produce *individualized* CLATE estimates, i.e. a treatment effect estimate for each observation in the experimental population. It is critical that these estimates are "out-of-bag." That is, they must come from GRFs that are not trained on the observation being assigned a prediction. I achieve this by applying a custom cross-validation algorithm with repeated partitioning and

[9] It is also key that these auxillary forests are cross-trained on observations *other* than observaton i.

retraining. The process is described in Algorithm 1, and is inspired by a somewhat similar partitioning algorithm used by Haushofer et al. (2022):

Algorithm 1

Steps
1. Split data into K folds
2. For each $k \in K$, do: a) Training data = $K' \leftarrow K/k$ other folds (i.e. all folds not in k) b) Train the forest by growing 2,000 causal trees and defining the GRF predictor of the individual treatment effect. c) Apply the predictor to k and generate a predicted individualized treatment effect i in iteration r: $\{\hat{\delta}_{ir}\}$
3. Repeat for 150 iterations, define the final individualized treatment effect as the average treatment effect over iterations: $\hat{\tau}_i = \frac{\sum_r \hat{\delta}_{ir}}{150}$ end

The process begins by splitting the data into an even set of K folds. Depending on the size of the experimental sample, I either set this value to 50 (in larger samples) or 100 (in smaller ones). In all cases, I set k – the number of folds in the hold-out sample – to be equal to 5, which corresponds to either 10% or 5% of the full sample. The algorithm then trains the GRF using the training sample, K', applying the trained model to the hold-out set, k, and repeating this process until each fold is held out exactly once. For example, when $K = 50$ and $k = 5$, this algorithm will train 10 GRFs in order to give every observation a predicted treatment effect using a forest which was trained on data that did not see its' outcome. I then repeat Steps 1 and 2 over 150 iterations in order to reshuffle the folds within k, which yields 150 predicted treatment effects for each observation in the experimental sample (note this includes *both* treatment and control individuals), from which I define the final predicted individual treatment effect (denoted $\hat{\tau}_i$) to be the mean across these 150 iterations.

GRFs (and any other machine learning algorithm for that matter) typically require some tuning of hyperparameters. I typically follow most of the defaults established in the 'grf' package in R – a stable CRAN package written by the original authors of the causal forest. A few of the most important hyperparameters include: the share of the training sample used in honest estimation (I use 50%), and the minimum number of observations in each child node in order for a split to occur (I use 5). These defaults tend to be more optimized on larger training samples, so I also experiment with some additional regularization as outlined in a GitHub post by one of the authors[10] which includes holding out only 10% of the training sample for honesty and increasing the minimum size of a node to 10 observations. The results under both sets of hyperparameters were qualitatively similar.

[10] https://github.com/grf-labs/grf/issues/120#issuecomment-327276697

3.2.2 Individualized Effect Distributions and Outcome Trade-offs

In order to visualize the heterogeneity uncovered by the GRF training procedure, Figure 1 displays the resulting $\hat{\tau}_i$ distributions – which for this figure have been standardized to z-scores to allow for comparability across outcomes – in one of the experimental samples (Mongolia) along the main diagonal. The distributions for business revenue, business expenditures, and household consumption appear more-or-less Gaussian, however the distributions of business profits and household income seem to have sizeable left-tails. The bottom-left and top-right sections of Figure 1 display correlations of predicted individualized treatment effects across outcomes. In several cases, there appear to be significant trade-offs in predicted impact based on the outcome selection. For example, if a microfinance organization were to select loan eligibility based on the likelihood of increased business expenditures, then (at least over the duration of time in this particular experiment) one would expect to see decreased business profits from those same individuals. Appendix Figures B.2 display these results for the other case studies. When distilling these results across experiments, there are several instances of trade-offs as well as complementarities. In Bosnia, Mexico, and Morocco, the business success outcomes tend to be positively correlated – indicating complementarity along those outcomes in most cases. However, there does not appear to be a strong general relationship between business success and household income or consumption – except in the case of Morocco, where business success is positively related to household income but negatively related to consumption.

3.2.3 Covariates Driving Heterogeneity

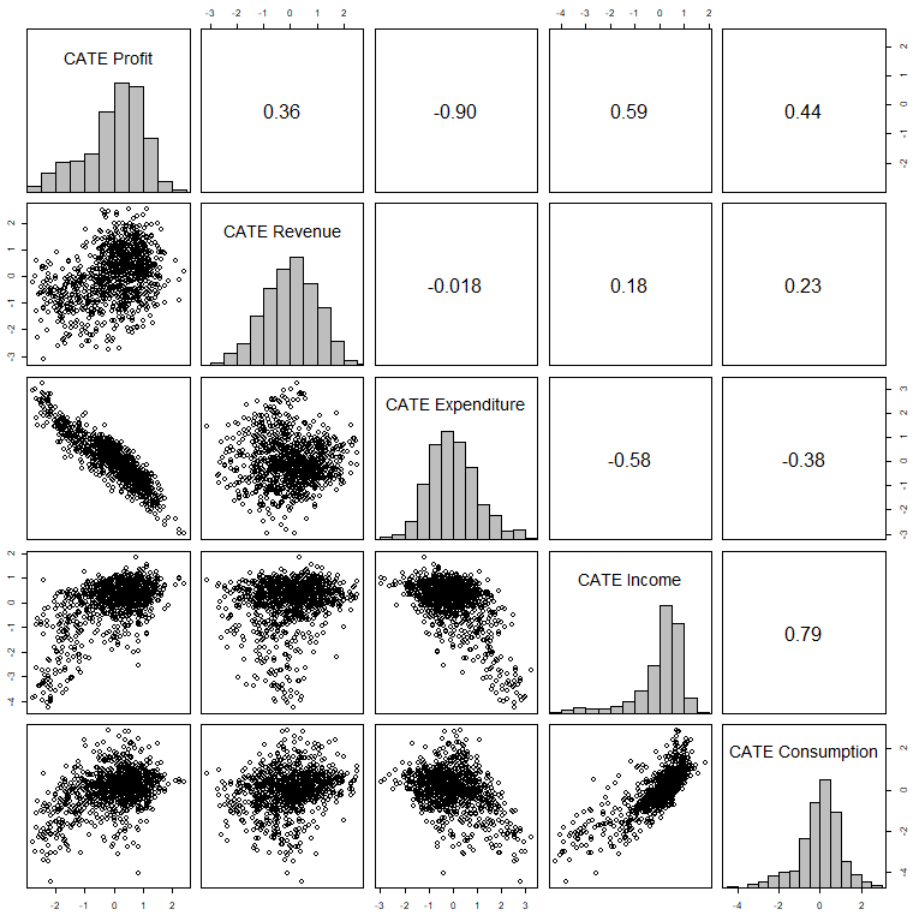
In order to paint a picture of which baseline variables are driving heterogeneity in $\hat{\tau}_i$, I report two sets of results. First, I display Variable Importance Factors (VIFs) in Table 3. VIFs are a simple average of the number of trees which split along a given covariate, weighted by the depth of the split within a tree such that earlier splits are weighted more heavily than later splits.[11] As such, VIF scores closer to 1 roughly indicate that a given covariate is a more informative predictor of the CLATE because a greater share of trees use it to create partitions. A limitation of this measure is that the VIF does not indicate *in what way* a given covariate is informative or how it may be informative non-linearly or interactively with other covariates.

The scores in Table 3 and Appendix Tables A.3 (for the remaining case studies) also serve to display the full list of covariates included in the GRF algorithms. I selected variables by hand among the set of baseline covariates collected in each paper. This was an attempt to limit the size of the covariate list given the relatively small size of the datasets.[12]. I attempted to select variables which were possibly relevant, had

[11] The formula for variable importance of variable x_k is: $VIF(x_k) = \sum_j \frac{1}{j^2}\left[\sum_j \frac{\sum_{trees} \#\text{ depth-}j\text{ splits on } x_k}{\sum_{trees} \#\text{ total depth-}j\text{ splits}}\right]$

[12] A "rule of thumb" is that the number of covariates should not be much greater than the log of the number of observations

Figure 1:
Correlation Across Targeting Algorithms
Mongolia



Notes: This figure shows the correlation of standardized $\hat{\tau}_i$ predictions across targeting algorithms for each of the main outcomes. For example, the top left scatter plot shows the correlation of predictions for the targeting algorithm trained on the "profit" outcome vs. an alternative model's predictions which train on the "revenue" outcome in terms of z-scores. In other words, it shows the likelihood that an individual with a given predicted $\hat{\tau}_i$ on business profit would have a predicted $\hat{\tau}_i$ of similar magnitude from the algorithm which predicts business revenues. High positive correlation between two algorithms hints at complementarity when selecting the appropriate targeting algorithm, whereas high negative correlation hints at a trade-off within the timeline of the experiment. The top-right panel shows the respective correlation coefficients, and the main diagonal shows the distribution of predicted treatment effects as generated by each algorithm.

some notable variation, and together showcased a variety of different elements a microfinance organization could realistically collect from loan applicants. I did not attempt to use data-driven methods for variable selection because it led to problems in another study which initially tried this approach (Haushofer et al., 2022).

I will highlight a few results from Table 3 and Appendix Tables A.3. First, VIF scores are often similar across outcomes within a given experiment, indicating that covariates tend to be either consistently informative or uninformative in predicting CLATEs in a general sense. In all cases, the baseline levels of the business success outcomes are predictive of treatment effects. For example, the baseline levels of business expenditures are a top-5 predictor of CLATEs in nearly all case studies. Other predictors which are consistently informative across case studies include: age, baseline value of assets, expenditure on durable goods, and baseline household income.

The second set of results I report which aim to help understand the patterns of heterogeneity are covariate means within different quartiles of the $\hat{\tau}_i$ distribution. Table 4 displays the results of this exercise for the algorithm which attempts to optimize household consumption. G4 represents the top 25% of the individualized treatment effect distribution and G1 represents the bottom 25%. Appendix Tables A.2 display these same results for the other experiments. To highlight some of these values in Table 4, keeping in mind the "important" variables according to VIF: households with larger predicted impacts (G4) tend to spend about $41 more bi-weekly on their business than those in G1, they own about $6,300 more in assets, and their bi-weekly incomes are about $64 higher.

3.3 Estimating Counterfactual ITTs – Subsample Analysis on $\hat{\tau}_i$ Distribution

3.3.1 Results with Ex-Ante Information Only

The next step in the analysis is to fully characterize the heterogeneity by performing subsample analysis on the upper ends of the individualized effect distribution. That is, I aim to again leverage the randomization of credit access and apply the linear model in equation (1) to evaluate the quality of the non-parametric mapping of the CLATE distribution produced by the GRFs. This procedure simply involves defining a "targeted subsample," denoted $M(\boldsymbol{X}) = 1$, and re-estimating equation (1) in the spirit of a "counterfactual" alternative policy under which credit access is randomly offered among only this population subset.

$$Y_{iv1}^{m} = \alpha + \tau^{m} W_{iv}^{m} + \beta \boldsymbol{X}_{i0}^{m} + Y_{iv0}^{m} + \varepsilon_{i1}^{m} \tag{6}$$

in the data (Chernozhukov et al., 2018). Following this rule strictly would limit my covariate list to around 16-20 variables. In most cases narrowing down the list of possibly informative predictors to 20 is difficult, so I generally allow for up to about 30.

Table 3: Mean Variable Importance Factors (VIF)
Mongolia

	Covariate	Profit	Revenue	Expenditures	Income	Consumption
1	Age	0.04	0.03	0.03	0.04	0.04
2	Value of assets owned	0.06	0.09	0.07	0.07	0.10
3	Total business expenditures	0.08	0.06	0.07	0.06	0.07
4	Total business profits	0.08	0.05	0.05	0.12	0.06
5	Total business revenues	0.10	0.05	0.09	0.11	0.07
6	Owns at least 1 business	0.00	0.00	0.00	0.00	0.00
7	Gave transfer to relatives in last year	0.00	0.00	0.00	0.00	0.00
8	Total hours of wage employment (household)	0.03	0.02	0.02	0.01	0.01
9	Number enterprises owned	0.00	0.01	0.00	0.00	0.01
10	Respondent (female) owns the business	0.00	0.00	0.00	0.00	0.00
11	Expenditure on tempation goods	0.03	0.04	0.03	0.02	0.04
12	Owns computer, phone, or satellite	0.01	0.00	0.00	0.00	0.00
13	Size of debt at baseline	0.04	0.04	0.03	0.03	0.04
14	Monthly durables consumption	0.07	0.09	0.08	0.03	0.06
15	Completed grade 8	0.00	0.00	0.00	0.01	0.00
16	Household income from business profits	0.11	0.10	0.12	0.11	0.10
17	Household income from food production	0.03	0.03	0.03	0.05	0.02
18	Monthly food consumption	0.06	0.06	0.05	0.06	0.05
19	Owns ger (tent)	0.00	0.00	0.00	0.00	0.01
20	Household wage income	0.02	0.04	0.02	0.02	0.02
21	Total hours enterprise employment (household)	0.02	0.05	0.02	0.02	0.02
22	Owns house	0.00	0.00	0.00	0.00	0.00
23	Individual liability	0.00	0.00	0.00	0.00	0.00
24	Has outstanding loans	0.00	0.00	0.00	0.00	0.00
25	Monthly nondurables consumption	0.05	0.05	0.05	0.03	0.05
26	Received transfer in last year	0.00	0.00	0.00	0.00	0.00
27	Loan amount	0.03	0.02	0.03	0.02	0.01
28	Total household consumption	0.03	0.03	0.03	0.04	0.06
29	Total household income	0.08	0.09	0.09	0.07	0.09
30	Total monthly consumption	0.03	0.03	0.03	0.04	0.06
31	Owns TV	0.00	0.00	0.00	0.00	0.00
32	Owns vehicle	0.00	0.00	0.00	0.00	0.00

Notes: This table displays the mean VIF across iterations for each of the GRF targeting algorithms I apply. For example, the first column represents the VIFs for the algorithm trained to target those with large $\hat{\tau}_i$ regarding business profits. The values themselves reflect the share of trees which split on a certain variable in the final model, weighted by the depth of the split. Therefore, larger values indicate a larger share of trees found that variable to be informative in predicting $\hat{\tau}_i$.

Table 4: $\hat{\tau}_i$ Quartile Covariate Means
Mongolia – Household Consumption

variable	G1	G2	G3	G4	G4-G1	G4-G2	G4-G3	estimate.variable
Value of assets owned	9648.55	8345.37	9754.88	15455.14	6314.61	6444.49	5912.76	assets_all_stan_BL
Size of debt at baseline	494.12	603.63	695.53	1307.46	761.17	724.30	464.26	currentdebt_stan
Total monthly consumption	197.52	198.89	223.28	434.82	231.01	234.00	205.35	totalc_stan_BL
Total household consumption	197.52	198.89	223.28	434.82	231.01	234.00	205.35	total_hh_consumption_stan_BL
Loan amount from XacBank	479.15	476.21	500.90	705.42	162.55	216.65	278.46	loan_amount_xac_stan
Total household income	82.11	76.67	90.27	145.15	63.72	72.57	51.10	total_hh_income_stan_BL
foodconsumption_stan_BL	70.65	59.64	67.33	121.99	47.20	60.25	52.33	foodconsumption_stan_BL
Household wage income	41.62	55.23	64.94	113.28	63.19	57.34	47.77	hhwageinc_stan_BL
Monthly nondurables consumption	38.31	37.90	45.34	87.11	46.21	49.17	39.72	nondur_consumption_stan_BL
Total business expenditures	11.96	17.77	18.67	52.34	40.76	27.79	31.46	biz_expenditures_stan_BL
Total business revenues	24.59	26.69	24.64	60.19	36.15	32.19	28.56	biz_revenues_stan_BL
Total hours of wage employment (household)	24.84	27.43	29.12	38.74	13.90	12.48	9.87	BLhours_wage
Household income from food production	7.26	6.14	6.19	14.08	7.84	8.53	7.00	food_inc_stan_BL
Expenditure on tempation goods	5.64	5.56	6.99	13.31	7.68	7.76	5.77	BLtempt_stan
Household income from business profits	31.21	12.68	4.62	20.20	-11.42	4.27	15.51	entprofit_stan_BL
Monthly durables consumption	4.32	4.56	4.78	10.33	6.09	5.93	5.40	dur_consumption_stan_BL
Total hours enterprise employment (household)	63.64	48.19	43.01	53.34	-12.23	-0.63	11.56	hours_ent_BL
Total business profits	11.00	8.21	3.77	11.46	1.74	3.25	7.38	biz_profits_stan_BL
Age	39.24	39.98	39.76	40.88	0.82	0.45	1.40	age_BL
Gave transfer to relatives in last year	0.34	0.38	0.41	0.53	0.20	0.17	0.11	BLgavetrans_y
Owns computer, phone, or satellite	0.15	0.13	0.17	0.26	0.11	0.16	0.09	compphonesat_BL
Owns vehicle	0.32	0.25	0.26	0.38	0.01	0.13	0.11	vehicle_BL
Owns ger (tent)	0.25	0.15	0.14	0.22	-0.04	0.08	0.04	ger1_BL
Owns house	0.35	0.30	0.29	0.37	-0.02	0.07	0.06	house1_BL
Number enterprises owned	0.35	0.28	0.25	0.35	-0.04	0.12	0.11	BLnrents
Has outstanding loans	0.62	0.62	0.64	0.68	0.04	0.04	0.04	loan_baseline
Received transfer in last year	0.05	0.07	0.12	0.11	0.05	0.04	-0.03	rectrans_BL
Owns TV	0.97	0.92	0.90	0.93	-0.05	0.00	-0.02	tv1_BL
Completed grade 8	0.88	0.88	0.85	0.87	-0.03	-0.02	0.03	eduhigh_BL
Respondent (female) owns the business	0.43	0.40	0.32	0.36	-0.06	-0.03	0.04	BLsoleent
Owns at least 1 business	0.77	0.60	0.48	0.55	-0.27	-0.10	0.06	BLenterprise

Notes: This table displays the means of each covariate within each quartile of the predicted treatment effect distribution for the GRF which targets on household consumption. This table contains the full list of covariates used for Mongolia. All stock currency variables are standardized to 2009 USD PPP, and flow currency variables are bi-weekly 2009 USD PPP.

Where the superscript m denotes the ITT among the targeted subset, which is identified as:

$$\tau^m = \tau|_{M(\boldsymbol{X})=1} = E[Y|M(\boldsymbol{X}) = 1, W = 1] - E[Y|M(\boldsymbol{X}) = 1, W = 0]$$

One point of clarification as to why I estimate equation (6) as opposed to an equation such as:

$$Y_{iv1} = \alpha + \tau_1 W_{iv} + \tau_2\Big(W_{iv} \cdot \mathbb{1}\{M(\boldsymbol{X})\}_{iv}\Big) + \theta \cdot \mathbb{1}\{M(\boldsymbol{X})\}_{iv} + \beta \boldsymbol{X}_{i0} + Y_{iv0} + \varepsilon_{i1}$$

In this alernative specification, $\mathbb{1}\{M(\boldsymbol{X})\}_{iv}$ serves as an indicator function which equals 1 if individual i is in the targeted subset, and $\tau_1 + \tau_2 = \tau^m$ from equation (6). While this alternative specification does allow me to directly test if τ_2 – the additional impact the average targeted individual receives from treatment over-and-above an untargeted individual – is statistically significant, ultimately the key question is if τ^m is statistically different from the benchmark τ. This alternative specification does not allow for that direct test, and therefore only serves to add confusion.

It is important to note that Hirano and Porter (2009) and Athey and Imbens (2019) point out that defining a targeted subsample based on a threshold – e.g. treating everyone where the predicted treatment effect is positive ($\hat{\tau}_i \geq 0$) or treating everyone except the bottom quartile ($\hat{\tau}_i \geq \hat{\tau}_{i,0.25}$), etc. – is suboptimal when the objective function of the planner is more complex than an unrestricted maximization of the average treatment effect. For example, if there are allocation restrictions on who should be treated, or if the objective is an arbitrary function of multiple outcomes, then the optimal policy will be a non-linear mapping of covariates to treatment status. Nonetheless, in this section I explore what I tentatively call "gains" to the average program impact by defining two targeted subsamples: the top 75% of the $\hat{\tau}_i$ distribution and the top 25% of the distribution for the given outcome.[13] Using quantiles of the distribution to define the treatment threshold is more flexible than setting an integer threshold such as $\hat{\tau}_i \geq 0$. The CLATE mapping is known to be consistent, but not necessarily unbiased in small samples (Athey et al., 2019). Therefore, selecting quantile thresholds weakens the precision requirement so that the GRF algorithms need only be able to *rank* individuals accurately rather than estimating a precise treatment effect magnitude in order to effectively reallocate the policy. And in fact, when I estimate equation (6) by counterfactually treating anyone with positive individualized treatment effects, the results remain noisy. This is consistent with Meager (2022), who found precisely null treatment effects for the bulk of the distribution using these same datasets and very different methodologies.

[13] I display treatment vs. control balance tests among these targeted subsamples in Appendix Tables A.7. In all cases covariates remain balanced within target samples.

Table 5 displays the main results of this exercise for each case study: a direct comparison of ITT estimates from the targeted subsamples vs. the benchmark (observed) sample. Column (1) re-states the estimates from Table 2; Columns (2) and (5) display the estimates of $\hat{\tau}^m$ for the targeted samples focusing on the top 25% of the individualized effect distribution and the top 75%, respectively; Columns (3) and (6) simply show the difference in ITT point estimates from the targeted samples vs. the benchmark; and Columns (4) and (7) indicate if the *gain itself* (i.e. the difference in ITT point estimates) is statistically significant using a two-sample t-test. I posit that a litmus test for a successful targeting application should consider each of (1) was the algorithm able to generate (at least suggestive) point estimate gains? (2) are these gains economically meaningful? (3) does the targeted sample generate greater precision than the benchmark sample? (4) is the *gain itself* a significant deviation from the benchmark? The different case studies were each successful in passing some of these tests, but to varying degrees.

The first question in the litmus test: "can the targeting algorithms generate point estimate gains in any sense?" is the most liberal definition of a successful targeting algorithm. The GRF procedure was able to generate at least suggestive point estimate gains for business-related outcomes (profits, revenues, and expenditures) in at least one of the two target subsamples, and 3/5 of the studies generated gains in the household-related outcomes (income and consumption) with only Mexico and Mongolia generating suggestive losses by those measures. When considering the economic magnitude of these gains, many of them are fairly large. The gains generated from the high take-up studies (Bosnia and Philippines) and Mexico often suggest counterfactual treatment effects which are between $\approx$ 90–1300% larger than the benchmark impacts with non-trivial improvements to PPP. Morocco and Mongolia also generate large suggestive gains in terms of percentage-gain, however the corresponding PPP gains are typically smaller and likely non-transformative.

The third and fourth measures of successful targeting – do the algorithms gain precision within their own estimating samples and are the gains a significant deviation from the benchmark? – are stricter definitions and are difficult to achieve in a setting of low take-up. Within the studies which randomized at a geographic unit and conversely had low take-up rates (Mexico, Morocco, Mongolia), there was only one study where a targeted sample achieved statistical significance where the benchmark sample did not: Morocco increased precision in business revenues and household income. There are otherwise no notable increases in precision and thus, the gains themselves generally remain statistically insignificant despite their often-large magnitude. However, in the studies which randomized among applicants (and thus had very high take-up rates), the targeting algorithms were much more likely to satisfy these criteria. Bosnia generated increased precision in

4/5 outcomes (minus consumption) and the treatment effect gains were statistically significant at the 10% level in 3/4 of these. Similarly, Philippines generated increased precision in 3/4 outcomes and treatment effect gains were statistically significant in 2/3 of these in the first targeted subsample.

Table 5: Treatment Effect Gains Comparison (GRF)

Panel A: Bosnia

Outcome	Benchmark ITT	Counterfactual ITT (Top 25%)	Gain ($)	Significant?	Counterfactual ITT (Top 75%)	Gain ($)	Significant?
Profits	34.13	58.51	24.38	Yes	125.47	91.34	Yes
Revenues	71.14	88.47	17.33	No	195.12	123.98	No
Expenditures	31.53	52.83	21.30	Yes	164.06	132.53	Yes
Household Income	10.78	48.89	38.11	Yes	158.64	147.86	Yes
Household Consumption	-5.60	-24.11	-18.51	No	-5.39	0.21	No

Panel B: Mexico

Outcome	Benchmark ITT	Counterfactual ITT (Top 25%)	Gain ($)	Significant?	Counterfactual ITT (Top 75%)	Gain ($)	Significant?
Profits	-1.76	23.21	24.97	No	9.53	11.29	No
Revenues	23.36	89.66	66.30	No	33.84	10.48	No
Expenditures	25.67	86.58	60.91	No	37.60	11.93	No
Household Income	-2.76	-4.71	-1.95	No	-5.25	-2.49	No
Household Consumption	-8.56	-25.20	-16.64	No	-18.65	-10.09	No

Panel C: Morocco

Outcome	Benchmark ITT	Counterfactual ITT (Top 25%)	Gain ($)	Significant?	Counterfactual ITT (Top 75%)	Gain ($)	Significant?
Profits	10.94	-12.21	-23.15	No	14.67	3.73	No
Revenues	12.52	-11.03	-23.55	No	17.06	4.54	No
Expenditures	40.34	85.89	45.55	Yes	41.96	1.62	No
Household Income	-25.03	232.43	257.46	Yes	70.38	95.41	No
Household Consumption	-1.83	-0.54	1.29	No	-3.42	-1.59	No

Panel D: Mongolia

Outcome	Benchmark ITT	Counterfactual ITT (Top 25%)	Gain ($)	Significant?	Counterfactual ITT (Top 75%)	Gain ($)	Significant?
Profits	-0.63	-0.97	-0.34	No	-0.52	0.11	No
Revenues	0.17	0.12	-0.05	No	0.20	0.03	No
Expenditures	0.85	0.77	-0.08	No	0.97	0.12	No
Household Income	9.75	-48.60	-58.35	Yes	-15.05	-24.80	No
Household Consumption	70.58	123.07	52.49	No	61.89	-8.69	No

Panel E: Philippines

Outcome	Benchmark ITT	Counterfactual ITT (Top 25%)	Gain ($)	Significant?	Counterfactual ITT (Top 75%)	Gain ($)	Significant?
Profits	64.94	125.28	60.34	Yes	79.54	14.60	No
Revenues	-85.43	88.00	173.43	No	699.72	785.15	Yes
Expenditures	-71.28	21.46	92.74	No	248.28	319.56	No
Household Income	-107.11	79.39	186.50	No	604.69	711.80	Yes

Notes: This table directly compares the benchmark treatment effects estimated in Table 2 to the counterfactual estimates from the GRF-defined targeted subsamples estimated via equation (6). For each outcome in each case study, I display the treatment effect coefficients from the benchmark sample (Column 1), the counterfactual treatment effects from filtering the bottom 25% and 75% of the $\hat{\tau}_i$ distribution (Columns 2 and 5, respectively), the treatment effect gains in USD PPP over the benchmark from both counterfactual samples (Columns 3 and 6, respectively), and the result of a two-sample t-test of significance of the difference between the counterfactual ITT and the benchmark ITT (Columns 4 and 7, respectively), where a "Yes" indicates the gain is significant at the 10% level.

3.3.2 Assessing Maximum Achievable Gains

Table 5 defined GRF-targeted subsamples by performing iterative filtration along the distribution of predicted treatment effects. This process is meant to partially simulate setting the targeting threshold based on ex-ante information. However, the "optimal" policy is the one which (assuming the objective function is solely determined by the ITT of the outcome in question) is the one such that the ITT is maximized in the target subsample. The ITT-maximizing policy is unattainable by using only ex-ante information, and in the next section I will redefine the "optimal" policy to be the one which minimizes ex-ante "empirical regret." Nonetheless, the ITT-maximizing policy is a good target to shoot for, and Table 6 displays the counterfactual treatment effects from the best possible cutoff threshold. Specifically, I iteratively filter from the 10^{th} percentile to the 90^{th} percentile in 1-pp increments and report the largest estimated gain along with the respective cutoff quantile. One notable pattern is that in the two studies with high take-up rates (Bosnia and Philippines), it is typically the case that the largest gains come from dropping a comparatively smaller percentage of the sample. It is also interesting to note that the largest gains in the remaining studies typically came from dropping much more of the population. This again highlights the difficulty of generating precise estimates of individualized treatment effects for the "middle" of the distribution in instances with low statistical power.

3.4 Deriving a Regret-Minimizing Policy

While the subsample analyses in Section 3.3 are useful for characterizing the heterogeneity of treatment effects in these experiments, they stop short of "deriving" an optimal policy based on ex-ante information. Arbitrarily determining a cutoff threshold (e.g. treating only the top 75% of the $\hat{\tau}_i$ distribution or treating those where $\hat{\tau}_i > 0$) is almost certainly suboptimal in practice for several reasons. First, even if one were to abstract away from budgetary limitations and assume the objective function seeks to maximize a single ITT, selecting the welfare-maximizing threshold for data outside of these experiments is impossible. The thresholds which yielded the maximum gains to the ITT are very unlikely to be externally valid as the optimal thresholds when the target population becomes more general. Second, is it conceptually simpler to derive a targeting policy based on a small set of rules founded in a fully realized algorithm rather than requiring the algorithm itself be implemented. I will discuss later how my approach can generalize to non-experimental data and/or incorporate common predictive algorithms outside of GRFs, but I posit that providing a simple "plug-and-play" policy which still achieves internally valid ITT gains can provide more feasible fixes to the program.

Table 6: Maximum Gains via GRF

Panel A: Bosnia

	Outcome	Baseline ITT	Max Gain ($)	Max Gain (%)	Quantile Dropped	Statistically Significant?
1	Profits	34.13	122.28	358.29	15.00	Yes
2	Revenues	71.14	235.37	330.86	13.00	Yes
3	Expenditures	31.53	299.07	948.54	10.00	Yes
4	Household Income	10.78	287.67	2668.54	10.00	No
5	Household Consumption	-5.60	102.31	1827.01	10.00	Yes

Panel B: Mexico

	Outcome	Baseline ITT	Max Gain ($)	Max Gain (%)	Quantile Dropped	Statistically Significant?
1	Profits	-1.76	29.97	1702.68	79.00	No
2	Revenues	23.36	93.16	398.81	90.00	No
3	Expenditures	25.67	197.12	767.91	90.00	Yes
4	Household Income	-2.76	39.58	1434.14	90.00	No
5	Household Consumption	-8.56	-1.32	-15.42	10.00	No

Panel C: Morocco

	Outcome	Baseline ITT	Max Gain ($)	Max Gain (%)	Quantile Dropped	Statistically Significant?
1	Profits	10.94	11.05	101.02	46.00	No
2	Revenues	12.52	7.39	59.04	62.00	No
3	Expenditures	40.34	58.08	143.98	76.00	Yes
4	Household Income	-25.03	327.89	1310.01	77.00	No
5	Household Consumption	-1.83	5.82	317.89	88.00	No

Panel D: Mongolia

	Outcome	Baseline ITT	Max Gain ($)	Max Gain (%)	Quantile Dropped	Statistically Significant?
1	Profits	-0.63	0.22	35.22	23.00	No
2	Revenues	0.17	0.19	113.22	50.00	No
3	Expenditures	0.85	0.50	58.58	90.00	No
4	Household Income	9.75	-4.03	-41.34	11.00	No
5	Household Consumption	70.58	144.61	204.88	90.00	No

Panel E: Philippines

	Outcome	Baseline ITT	Max Gain ($)	Max Gain (%)	Quantile Dropped	Statistically Significant?
1	Profits	64.94	90.60	139.51	90.00	Yes
2	Revenues	-85.43	2126.57	2489.26	11.00	Yes
3	Expenditures	-71.28	554.29	777.62	10.00	No
4	Household Income	-107.11	809.67	755.92	19.00	Yes

Notes: This table displays the maximum gains from the GRF-defined treatment effect distribution. I select the ITT-maximizing quantile threshold from the $\hat{\tau}_i$ distribution via a grid search over each 1-pp increment from the 10th percentile to the 90th percentile. Column 2 displays the benchmark ITTs pulled from Table 2. Columns 3 and 4 display the gain generated from the best model in USD and percentage-gain, respectively. Column 5 indicates the cutoff threshold of the best model – i.e. the best model dropped the bottom X-percent of the distribution. The final column indicates if the treatment effect gain estimated by the best model is statistically significant at the 10% level according to a two-sample t-test.

Athey and Wager (2021) provide a practical implementation of their algorithm designed to produce shallow-depth decision trees which act as treatment assignment functions. In their approach, we seek a policy, π, which maps baseline covariates to a treatment assignment decision, $\pi : \boldsymbol{X}_i \rightarrow \{0, 1\}$. The optimal assignment function is the policy within the set of finite-depth policies, $\hat{\pi} \in \Pi$, such that empirical "regret" is minimized, which is defined as the difference in total utility from the best possible policy, π', vs. the derived policy, $\hat{\pi}$:

$$R(\pi) = \max_{\pi'} \Big\{ E[Y_i\big(\pi'(\boldsymbol{X}_i)\big)] \Big\} - E[Y_i\big(\hat{\pi}(\boldsymbol{X}_i)\big)]$$

In practice, the best policy, π', is unobservable. However, the authors show that their approach leads to bounds on $R(\pi)$ in the order of $\sqrt{VC(\Pi)/n}$, where $VC(\Pi)$ is the Vapnik–Chervonenkis dimension of the set of feasible treatment assignment functions.[14] By default, the GRF creates nonparametric approximations of the conditional mean function, $m(w, \boldsymbol{x}) = E[Y_i(w)|W_i, \boldsymbol{X}_i]$, and the propensity score function, $e(\boldsymbol{x}) = P[W_i = 1|\boldsymbol{X}_i]$. Using these, we can construct doubly-robust scores (Ψ_i) representing individual contributions to the ITT by plugging into the popular augmented inverse-probability weights estimator (Robins et al., 2017):

$$\hat{\Psi}_i = \hat{m}(1, \boldsymbol{X}_i) - \hat{m}(0, \boldsymbol{X}_i) + \big(Y_i - \hat{m}(1, \boldsymbol{X}_i)\big) \frac{W_i}{\hat{e}(\boldsymbol{X}_i)} - \big(Y_i - \hat{m}(0, \boldsymbol{X}_i)\big) \frac{1 - W_i}{1 - \hat{e}(\boldsymbol{X}_i)} \tag{7}$$

The derived policy follows as the solution to a custom objective function of the outcomes and the doubly-robust scores:

$$\hat{\pi} = \text{argmax} \Big\{ \frac{1}{n} \sum_{i}^{n} \Big(2\pi(\boldsymbol{X}_i) - 1 \Big) \hat{\Psi}_i \Big\} \tag{8}$$

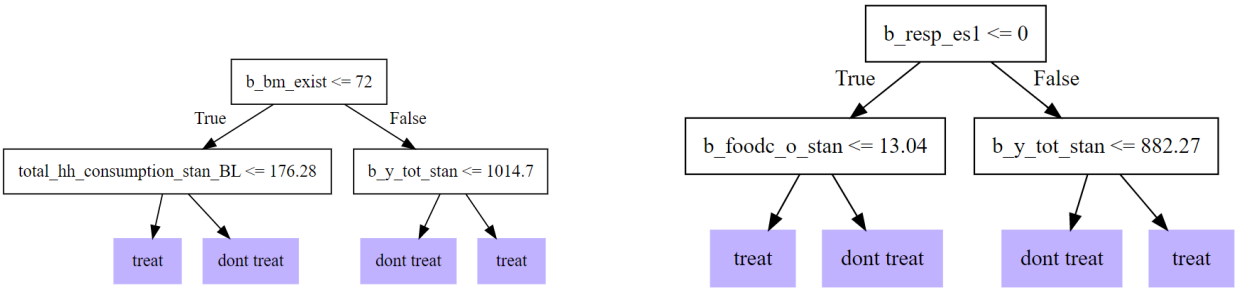
I apply the above approach to derive depth-2 decision tree policies using the Bosnia case study and two of the main outcomes: business profits and household income. First, I train a single GRF for each of the two outcomes. Second, I derive the treatment assignment functions using 50% of the sample, and I estimate the counterfactual ITT by estimating equation (9) on the remaining 50% of the sample, where the superscript pt indicates the ITT among the policy tree-defined target subsample:

$$Y_{iv1}^{pt} = \alpha + \tau^{pt} W_{iv}^{pt} + \beta \boldsymbol{X}_{i0}^{pt} + Y_{iv0}^{pt} + \varepsilon_{i1}^{pt} \tag{9}$$

Figure 1 displays the resulting treatment assignment functions from this exercise. The tree targeting the ITT effect on business profits uses three covariates in the decision process: if the business has existed for less than 72 months ("b_bm_exist") but the household spends less than \$177 on bi-weekly consumption ("total_hh_consumption_stan_BL"), the decision is to provide that household with credit access. Conversely,

[14] $VC(\Pi)$ represents the maximum number of observations that can be correctly classified by a function within the set of depth-d decision trees.

if the business is older than 72 months and their bi-weekly income ("b_y_tot_stan") is at least \$1,014 (i.e. annual income of $\approx$ \$24,300), the household will obtain credit acess. For the tree targeting household income, we first consider if the individual is employed ("b_resp_es1"). If they are employed and their bi-weekly income is at least \$882 (annual income $\approx$ \$21,200), we provide them with credit access. If the respondent is unemployed, they may still obtain credit access if they spend less than \$13 bi-weekly on food outside of the home (e.g. restaurants) – perhaps indicating that some measure of frugality is associated with positive outcomes in the event of unemployment.



(a) Business Profits

(b) Household Income

Figure 1: Policy Trees for Each Target Outcome

Table 7 compares the treatment effect estimates obtained via equation (9) to the benchmark and counterfactual estimates previously discussed. ITT gains among the tree-defined target subsample remain fairly large, but are smaller than the gains generated from the better of the two GRF-defined target subsamples. This is to be expected considering the policy tree is of low depth relative to the complex process of the GRF.

Treatment Effect Gains Comparison (Policy Trees)

	Outcome	Benchmark Estimate	Counterfactual (top 75%)	Gain (\$)	Counterfactual (top 25%)	Gain (\$)	Policy Tree	Gain (\$)
1	Profits	34.13	58.51	24.38	125.47	91.34	111.54	77.41
2	Income	10.77	48.89	38.12	158.64	147.87	97.92	87.15

Notes: This table directly compares the gains created from the causal-forest generated targeting indices to the policy-tree generated targeting indices.

Table 7: Policy Tree ITT Gains Comparison

4 Generalizing Subsample and Policy Derivation Exercises

The aim of this section is to explore some key additional questions regarding the generalizability of the framework and results laid out in Section 3. Specifically, I examine three major factors which play a role

in how the approach and/or results may be extrapolated in a general setting. First, I explore whether one requires a causal mapping of the CLATE function to produce treatment effect gains or if we may substitute the GRF for one of the many general-purpose predictive algorithms found in the computer science literature. Second, I consider the borrowing propensity of targeted vs. untargeted microfinance clients. Finally, I test for the presence of general equilibrium and spillover effects onto nonborrowers within treated villages using the three case studies that randomized across geographic strata.

4.1 Alternative Algorithms: OLS and SVMs

The first main question one may ask is if the GRF or another "causal" machine learning approach is necessary for replication of these results in a more general external population. While the GRF and other causal ML methods provide several key advantages to understanding the patterns of heterogeneity in the data including the construction of doubly robust scores and consistent point estimates of CLATEs/CATEs, they are not strictly necessary in order to produce a predictor of the CLATE/CATE function itself (Chernozhukov et al., 2018). In fact, any "generic" predictive algorithm can construct an estimate of the CLATE function simply by predicting an outcome in the presence of treatment, $\hat{Y}_i(1)$, and an outcome in the absence of treatment $\hat{Y}_i(0)$, then defining the individualized treatment effect as their difference: $\hat{\tau}_i \equiv \hat{Y}_i(1) - \hat{Y}_i(0)$. While this mapping is not necessarily consistently estimated, it may still be useful in certain settings where prediction is the main goal, which includes this setting.

Training general-purpose predictive algorithms comes with several considerations not necessarily common in the average economist's toolkit (Athey and Imbens, 2019). Some of these include hyperparameter tuning, regularization[15], and the relative sparsity of the covariate space to the sample size. In all cases, it is key to carefully select the algorithm itself in order to achieve predictive validity (i.e. out-of-bag accuracy according to the selected loss function). This is all to say that certain popular methods which rely on stochastic gradient descent along the loss function (e.g. neural networks) are not universally appropriate and tend to require large amounts of data. In my setting, I test two alternative predictive algorithms: Support Vector Machines (SVMs) and a simple linear model estimated via OLS. In all cases I follow the procedure outlined in Algorithm 1, replacing the GRF predictor of CLATEs with the OLS/SVM predictor, then performing subsample analysis by estimating equation (6) with the ML-defined targeted subsample.

[15]Regularization is not uncommon to economics, as (for example) the ubiquitous LASSO regression is simply a linear model with regularization. However, fretting over the appropriate degree of regularization or the loss function itself is a much more important issue in prediction problems.

4.1.1 OLS Predictor

OLS is commonly used in practice when conducting proxy means tests for targeting the "poorest of the poor" in cash transfer and other anti-poverty programs (Alatas et al., 2012). However, OLS is unregularized by default, meaning there is no penalty for overfitting. In settings with noisy data, OLS is prone to overstate the impacts of certain covariates on the CLATE (conversely, this could be beneficial in settings with greater statistical power but fewer observations). In training the OLS as a predictor, I include the same set of covariates as in the GRF, but additionally include a full set of covariate interactions and quadratic terms (where applicable) in order to allow the OLS to capture non-linearities the GRF and SVMs will attempt to capture by default. Table 8 displays the "gains" via OLS-defined targeted subsamples, contrasted with the gains from the GRF-defined subsample. In all cases I use the better of the two target samples (i.e. the larger gain between targeting the top 75% or the top 25%). OLS performs very well in the high signal-to-noise case studies (Bosnia and Philippines), but very poorly in the three settings with lower statistical power (Mexico, Morocco, Mongolia). Again, this highlights the benefits and costs of using simple, unregularized models as predictors of CLATES: in settings where the model *could* make credibly strong claims to accurately map the CLATE function, it was successful; in settings where there was a large amount of noise, the model severely overfit to the training data and was not successful.

4.1.2 SVM Predictor

SVMs are one of the common general-purpose predictive algorithms which are not very different from regularized maximum likelihood models (Vapnik, 2013). Therefore, they provide an alternative to forest-based approaches as they require less information than algorithms which rely on stochastic gradient descent but are still known to perform well in high-dimensional settings. I again use the same set of covariates as in the GRF case, and I follow most of the default hyperparameters in the 'e1071' R package, with 5-fold cross validation to confirm they were optimal. Table 9 displays the results of this exercise: the targeting gains from the SVM-defined subsample contrasted with the GRF-defined subsample. The SVM performs similarly to the GRF, however is ultimately unable to produce statistically significant gains to the ITT in nearly all instances. This indicates that the forest-based approach of the GRF is generally better suited to prediction in these case studies.

4.2 The Relative Borrowing Propensity of Targeted Subsamples

The second key generalization question is whether those targeted under GRF-defined subsamples are more or less likely to borrow than those untargeted. It is of consequence to understand if those with large

Table 8: Counterfactual ITT Estimates with OLS $\hat{\tau}$ Predictor

Panel A: Bosnia

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (OLS)	Significant?
1	Profits	34.13	219.06	125.47	184.93	Yes
2	Revenues	71.14	300.63	195.12	229.49	Yes
3	Expenditures	31.53	109.79	164.06	78.26	No
4	Income	10.77	199.89	158.64	189.12	Yes
5	Consumption	-5.60	114.22	-5.39	119.82	Yes

Panel B: Mexico

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (OLS)	Significant?
1	Profits	-1.76	-4.37	10.98	-2.61	No
2	Revenues	23.36	27.46	89.66	4.10	No
3	Expenditures	25.67	-23.73	86.58	-49.40	No
4	Income	-2.76	-96.82	-4.71	-94.06	No
5	Consumption	-8.56	-18.30	-18.65	-9.74	No

Panel C: Morocco

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (OLS)	Significant?
1	Profits	10.94	-10.99	14.67	-21.93	No
2	Revenues	12.52	3.53	17.06	-8.99	No
3	Expenditures	40.34	22.95	41.96	-17.39	No
4	Income	-25.03	-766.54	232.43	-741.51	No
5	Consumption	-1.83	-7.20	-0.54	-5.37	No

Panel D: Mongolia

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (OLS)	Significant?
1	Profits	-0.63	-0.55	-0.52	0.08	No
2	Revenues	0.17	0.01	0.36	-0.16	No
3	Expenditures	0.85	0.72	0.97	-0.13	No
4	Income	9.75	-11.41	-15.05	-21.16	No
5	Consumption	70.58	73.86	68.94	3.28	No

Panel E: Philippines

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (OLS)	Significant?
1	Profits	64.92	164.58	125.24	99.66	Yes
2	Revenues	-85.43	487.27	399.19	572.70	Yes
3	Expenditures	-71.28	86.10	21.46	157.38	No
4	Income	-107.11	774.40	604.69	881.51	Yes

Notes: This table replicates the regressions highlighted in Table 5 using CLATE predictors generated from alternative targeting algorithms from causal forest: OLS and Support Vector Machines (SVM). This table displays the results for OLS.

Table 9: Counterfactual ITT Estimates with SVM $\hat{\tau}$ Predictor

Panel A: Bosnia

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (SVM)	Significant?
1	Profits	34.13	122.60	125.47	88.47	No
2	Revenues	71.14	200.26	195.12	129.12	No
3	Expenditures	31.53	90.76	164.06	59.23	No
4	Income	10.77	0.24	158.64	-10.53	No
5	Consumption	-5.60	-31.55	-5.39	-25.95	No

Panel B: Mexico

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (SVM)	Significant?
1	Profits	-1.76	2.81	10.98	4.57	No
2	Revenues	23.36	96.26	89.66	72.90	No
3	Expenditures	25.67	68.15	86.58	42.48	No
4	Income	-2.76	-3.56	-4.71	-0.80	No
5	Consumption	-8.56	-5.45	-18.65	3.11	No

Panel C: Morocco

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (SVM)	Significant?
1	Profits	10.94	23.82	14.67	12.88	No
2	Revenues	12.52	19.80	17.06	7.28	No
3	Expenditures	40.34	48.47	41.96	8.13	No
4	Income	-25.03	715.77	232.43	740.80	Yes
5	Consumption	-1.83	-30.92	-0.54	-29.09	No

Panel D: Mongolia

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (SVM)	Significant?
1	Profits	-0.63	-0.59	-0.52	0.04	No
2	Revenues	0.17	0.02	0.36	-0.15	No
3	Expenditures	0.85	1.05	0.97	0.20	No
4	Income	9.75	15.09	-15.05	5.34	No
5	Consumption	70.58	80.70	68.94	10.12	No

Panel E: Philippines

	Outcome	Benchmark Estimate	Counterfactual Estimate	Gain (CF)	Gain (SVM)	Significant?
1	Profits	64.92	99.29	125.24	34.37	No
2	Revenues	-85.43	44.45	399.19	129.88	No
3	Expenditures	-71.28	-29.45	21.46	41.83	No
4	Income	-107.11	-121.34	604.69	-14.23	No

Notes: This table replicates the regressions highlighted in Table 5 using CLATE predictors generated from alternative targeting algorithms from causal forest: OLS and Support Vector Machines (SVM). This table displays the results for SVM.

predicted treatment effects are more or less likely to borrow if offered a loan than those with small predicted treatment effects. Any trade-offs between take-up probability and treatment effect magnitude might suggest a reallocation of funds toward marketing on the part of the MFI, which could positively increase program reach. In order to investigate this, I estimate a linear probability model for each GRF-defined subsample where I regress "target" status on take-up for each individual. The results of this are presented in Table 10. I omit the two studies which randomized credit access among applicants because those results come from samples that are selected on the basis of expressed borrowing interest. While there are no consistent cases of borrowing propensity divergence between the two groups, there are some significant differences which occasionally appear. For example, in the case of Mongolia, those targeted under the "profit" model are 9pp less likely to borrow than those who are untargeted. Conversely, those targeted under "expenditures" are 17pp more likely to borrow than their untargeted counterparts.

4.3 Within-Village Spillover Effects

A final question is related to the internal validity of the counterfactual ITTs from the studies that randomized credit access among geographic strata (Mexico, Mongolia, Morocco). A key assumption in mapping CLATEs is that the allocation of loans does not affect the ITT outside of the direct impact of receiving a loan on one's own outcomes. In other words, if there are within-strata spillovers, then the observed outcomes from individuals in treatment villages will not reflect the counterfactual outcomes of those same individuals if credit access is reallocated among them. In a perfect world, I would be able to estimate this equation via OLS to assess the size of the spillovers in a given study:

$$Y_{iv} = \alpha + \beta_1[W \times \mathbb{1}(\text{Loan})]_{iv} + \beta_2[W \times \mathbb{1}(\text{No Loan})]_{iv} + \beta_3[\mathbb{1}(\text{No Loan})]_{iv} + \varepsilon_{iv}$$

Where $\beta_2 \approx 0$ would indicate small spillovers as this parameter represents the additional impact of being in a treatment region but rejecting a loan over being in a control region and rejecting a loan. However, in order for β_2 to be an unbiased estimate of the true spillover parameter, there cannot be selection into treatment status. This is because (almost) everyone in the control villages is observed as "No Loan," so β_2 compares individuals who rejected a loan to individuals who were not offered a loan. If the expected counterfactual outcomes of individuals who rejected loans do not represent the expected counterfactual outcomes of individuals who were not offered a loan, then β_2 will be biased. So, I must first segment individuals in the control group such that we can directly compare never-taker treatment individuals with control group individuals who are likely to also be never-takers. Modifying the approach taken in Crépon et al. (2015), I estimate models of credit demand using a logit regression, employing Algorithm 1 for the creation of cross-trained individualized scores.

Table 10: ML Targets vs. Borrowing Propensity

Panel A: Mexico

	Profit Algorithm	Revenue Algorithm	Expenditure Algorithm	Income Algorithm	Consumption Algorithm
ML Targeted	0.05	0.05*	0.01	0.04	−0.01
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)
R^2	0.01	0.01	0.00	0.00	0.00
Num. obs.	663	663	663	663	663

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Panel B: Morocco

	Profit Algorithm	Revenue Algorithm	Expenditure Algorithm	Income Algorithm	Consumption Algorithm
ML Targeted	−0.02	−0.03	−0.02	0.07***	−0.03*
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
R^2	0.00	0.00	0.00	0.01	0.00
Num. obs.	1792	1792	1792	1792	1792

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Panel C: Mongolia

	Profit Algorithm	Revenue Algorithm	Expenditure Algorithm	Income Algorithm	Consumption Algorithm
ML Targeted	−0.09*	−0.03	0.17***	0.00	−0.02
	(0.05)	(0.04)	(0.04)	(0.06)	(0.05)
R^2	0.01	0.00	0.02	0.00	0.00
Num. obs.	701	701	701	701	701

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: This table shows the results from estimating the following linear probability model for each targeting algorithm: $1(TookLoan)_i = \alpha + \beta * 1(Targeted)_{ia} + \varepsilon_{ia}$. Where $1(TookLoan)_i$ is an indicator function = 1 if individual i was observed as accepting a loan, and $1(Targeted)_{ia}$ is an indicator function = 1 if the individual was targeted via algorithm a. I only use the experimental treatment sample here as including control individuals would skew these estimates because they are prevented from borrowing. The reported coefficients are the β values for the respective algorithm. Positive coefficients suggest that the sub-sample of individuals who are targeted under algorithm a are also more likely to borrow than the sub-sample who are untargeted under algorithm a. Coefficients of 0 suggest that there are no differences in borrowing tendencies between the targeted sample vs. untargeted sample, and conversely, a negative coefficient suggests that the targeted sub-sample is less likely to borrow than the untargeted sub-sample.

I then use the resultant propensity scores to estimate the following specification among the experimental treatment sample:

$$Y_{iv} = \alpha + \beta_1[W \times \mathbb{1}(\text{High Score})]_{iv} + \beta_2[W \times \mathbb{1}(\text{Low Score})]_{iv} + \beta_3[\mathbb{1}(\text{Low Score})]_{iv} + \varepsilon_{iv} \tag{10}$$

Where "High Score" individuals are in the top 30% of the propensity score distribution and "Low Score" individuals are in the bottom 30%. Now, β_2 represents the effect of being a likely never-taker in a treatment village over being a likely never-taker in a control village. Table 11 shows the results of the regressions estimated via these scores. While it is difficult to make any strong statements regarding the "no-spillovers" assumption based on these results, the estimates are typically not statistically significant on "low likelihood" individuals (except in the case of Mongolia). However, the point estimates are certainly not zero, adding a caveat to the subsample analyses conducted on these case studies.

Table 11: Spillover Estimates via Logit Propensity Scores

Panel A: Mexico

	Business profits (USD PPP 2009)	Business revenues (USD PPP 2009)	Business expenditures (USD PPP 2009)	Household income (USD PPP 2009)	Household consumption (USD PPP 2009)
Treat x High Score	47.49	108.34*	61.86*	7.77	5.91
	(35.07)	(60.67)	(32.28)	(25.22)	(15.06)
Treat x Low score	−38.99	−13.09	25.86	−1.61	−11.56
	(35.14)	(15.52)	(36.77)	(32.73)	(21.17)
R^2	0.01	0.02	0.02	0.00	0.01
Num. obs.	1060	1060	1060	972	948

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Panel B: Morocco

	Business profits (USD 2009 PPP)	Business revenues (USD 2009 PPP)	Business expenditures (USD 2009 PPP)	Household income (USD 2009 PPP)	Household consumption (USD 2009 PPP)
Treat x High Score	29.31	39.52*	55.94	346.85	0.70
	(30.86)	(20.71)	(49.94)	(390.06)	(10.29)
Treat x Low score	−12.62	−2.73	45.47	−450.25	−17.41
	(19.25)	(10.53)	(36.31)	(313.24)	(11.37)
R^2	0.10	0.09	0.13	0.12	0.23
Num. obs.	2168	2168	2168	2168	2162
N Clusters	162	162	162	162	162

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Panel C: Mongolia

	Business profits (USD 2009 PPP)	Business revenues (USD 2009 PPP)	Business expenditures (USD 2009 PPP)	Household income (USD 2009 PPP)	Household consumption (USD 2009 PPP)
Treat x High Score	−0.32	0.48	0.94	−14.55	13.37
	(0.72)	(0.36)	(0.83)	(30.79)	(44.97)
Treat x Low score	−0.68*	0.48*	1.13**	−24.86	71.65**
	(0.35)	(0.28)	(0.46)	(40.05)	(27.19)
R^2	0.04	0.06	0.08	0.02	0.38
Num. obs.	572	572	572	572	572
N Clusters	40	40	40	40	40

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Notes: This table displays results from the following regression: $Y_{iv} = \alpha + \beta_1[W \times \mathbb{1}(\text{High Score})]_{iv} + \beta_2[W \times \mathbb{1}(\text{Low Score})]_{iv} + \beta_3[\mathbb{1}(\text{Low Score})]_{iv} + X_{iv} + \varepsilon_{iv}$. Following Crepon et al. (2015), I segment the sample into the top and bottom 30% of propensity scores based on the logit predictor to classify individuals into "High" or "Low" borrowing likelihood. The table above displays the coefficients β_1 and β_2. Note that there are no spillover estimates for Bosnia due to the unit of randomization.

5 Conclusion

While machine learning has generated hype in both public discourse and in scientific research for its' predictive capabilities and data-driven modeling approaches, in order for any benefits to materialize we must first understand where and when to apply it. In the field of economic development, recent research has shown the need for targeting anti-poverty programs, at least in some settings, based on predicted treatment effects rather than relative need. These "prediction policy"-type questions, where relevant dimensions of treatment effect heterogeneity do not necessarily need to have strong theoretical backing, are perfect settings for the application of these techniques. Microfinance has proven to be such a setting where, not only are treatment effects substantially heterogeneous across subpopulations, but identifying individuals whose treatment effects are positive is likely to be difficult with theory because the factors which determine one's responsiveness to a loan may be multi-dimensional and complex.

This study assesses the feasibility of enhancing the microfinance targeting process using machine learning to identify those with larger treatment effects. It provides a proof-of-concept for selecting relevant outcomes, establishing benchmark treatment effects, implementing a targeting procedure which identifies positively impacted sub-populations, and generating gains in the average impact across this targeted population. Successful implementation of this procedure in an applied setting may lead to increased program cost-effectiveness, prevent negative effects associated with loan defaults or poor investments, and better identify the characteristics of successful loan recipients.

I use RCT data from five independent studies, applying the micro-data on applicant characteristics, lender characteristics, loan structure, etc. to nonparametrically map the conditional local average treatment effect function with generalized random forests and other machine learning algorithms. With this mapping, I conduct counterfactual subsample analysis on the treatment effect distribution, isolating often large impacts to the average treatment effect within targeted populations. With the heterogeneity uncovered from my machine learning framework, I derive illustrative "plug and play" targeting rules designed for easy implementation in similar settings. These simple decision rules still yield substantial gains to the average program impact without the need to train and deploy a full targeting algorithm.

I also explore trade-offs across algorithms. Individuals targeted under algorithms trained to maximize business-related outcomes tend to be targeted under all outcomes in that "outcome family," but tend to be untargeted under models trained to maximize household income or consumption. Additionally, those

targeted are sometimes less likely to borrow than those who are untargeted – a key finding to a practical implementation of expanding program reach efficiently. Finally, I discuss trade-offs across algorithms designed to map the CLATE function. Under certain scenarios, OLS, SVMs, or other general-purpose predictive algorithms may perform well and serve as better predictors of the CLATE function than causal methods – though they lack beneficial asymptotic properties.

While this analysis provides many promising results, it also showcases the limitations of data-driven methods. Machine learning is not a cure-all, and in-fact is far from guaranteed to generate gains to the average treatment effect. These experimental settings where data are both small and noisy are cautionary tales, and are lower bounds (of sorts) for the types of data where ML may provide useful results. Nonetheless, my hope is that this study has provided a useful framework for mobilizing modern nonparametric methods to improve the effectiveness of interventions by leveraging experimental or quasi-experimental data.

References

Abadie, A. (2003, April). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics 113*(2), 231–263.

Adato, M., M. R. Carter, and J. May (2006, February). Exploring poverty traps and social exclusion in South Africa using qualitative and quantitative data. *The Journal of Development Studies 42*(2), 226–247. Publisher: Routledge _eprint: https://doi.org/10.1080/00220380500405345.

Agrawal, A., J. Gans, and A. Goldfarb (2019). *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Alatas, V., A. Banerjee, R. Hanna, B. A. Olken, R. Purnamasari, and M. Wai-Poi (2013, June). Ordeal Mechanisms In Targeting: Theory And Evidence From A Field Experiment In Indonesia.

Alatas, V., A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias (2012, June). Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review 102*(4), 1206–1240.

Andrews, I., T. Kitagawa, and A. McCloskey (2024, February). Inference on Winners*. *The Quarterly Journal of Economics 139*(1), 305–358.

Angelucci, M., D. Karlan, and J. Zinman (2015, January). Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco. *American Economic Journal: Applied Economics 7*(1), 151–182.

Angrist, J. D. (2004). Treatment Effect Heterogeneity in Theory and Practice. *The Economic Journal 114*(494), C52–C83. Publisher: [Royal Economic Society, Wiley].

Armendariz, B. and J. Morduch (2010, April). *The Economics of Microfinance, second edition*. MIT Press. Google-Books-ID: XSo3AgAAQBAJ.

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

Athey, S. and G. W. Imbens (2017, May). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives 31*(2), 3–32.

Athey, S. and G. W. Imbens (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics 11*(1), 685–725. _eprint: https://doi.org/10.1146/annurev-economics-080217-053433.

Athey, S., J. Tibshirani, and S. Wager (2019, April). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178. Publisher: Institute of Mathematical Statistics.

Athey, S. and S. Wager (2021). Policy Learning With Observational Data. *Econometrica 89*(1), 133–161. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15732.

Attanasio, O., B. Augsburg, R. De Haas, E. Fitzsimons, and H. Harmgart (2015, January). The Impacts of Microfinance: Evidence from Joint-Liability Lending in Mongolia. *American Economic Journal: Applied Economics 7*(1), 90–122.

Augsburg, B., R. De Haas, H. Harmgart, and C. Meghir (2015). The Impacts of Microcredit: Evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics 7*(1), 183–203. Publisher: American Economic Association.

Banerjee, A., E. Duflo, R. Glennerster, and C. Kinnan (2015). The Miracle of Microfinance? Evidence from a Randomized Evaluation. *American Economic Journal: Applied Economics 7*(1), 22–53. Publisher: American Economic Association.

Banerjee, A., D. Karlan, and J. Zinman (2015, January). Six Randomized Evaluations of Microcredit: Introduction and Further Steps. *American Economic Journal: Applied Economics 7*(1), 1–21.

Bangdiwala, S. I., A. Bhargava, D. P. O'Connor, T. N. Robinson, S. Michie, D. M. Murray, J. Stevens, S. H. Belle, T. N. Templin, and C. A. Pratt (2016, June). Statistical methodologies to pool across multiple intervention studies. *Translational Behavioral Medicine 6*(2), 228–235.

Barometer, M. (2019). Microfinance Barometer. https://www.convergences.org/wp-content/uploads/2019/09/Microfinance-Barometer-2019_web-1.pdf.

Bhattacharya, D. and P. Dupas (2012, March). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics 167*(1), 168–196.

Blumenstock, J., G. Cadamuro, and R. On (2015, November). Predicting poverty and wealth from mobile phone metadata. *Science 350*(6264), 1073–1076. Publisher: American Association for the Advancement of Science.

Breiman, L. (2001, October). Random Forests. *Machine Learning 45*(1), 5–32.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fernández-Val (2018, June). Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India. *NBER Working Paper Series No. 24678*.

Coady, D., M. E. Grosh, and J. Hoddinott (2004). *Targeting of Transfers in Developing Countries: Review of Lessons and Experience*. World Bank Publications. Google-Books-ID: l3ppSymUipQC.

Crépon, B., F. Devoto, E. Duflo, and W. Parienté (2015). Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco. *American Economic Journal: Applied Economics 7*(1), 123–150. Publisher: American Economic Association.

Cull, R., A. Demirgüç-Kunt, and J. Morduch (2009, March). Microfinance Meets the Market. *Journal of Economic Perspectives 23*(1), 167–192.

Haushofer, J., P. Niehaus, C. Paramo, E. Miguel, and M. W. Walker (2022, June). Targeting Impact versus Deprivation. *NBER Working Paper Series No. 30138*.

Hirano, K. and J. R. Porter (2009). Asymptotics for Statistical Treatment Rules. *Econometrica 77*(5), 1683–1701. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6630.

Jawadekar, N., K. Kezios, M. C. Odden, J. A. Stingone, S. Calonico, K. Rudolph, and A. Zeki Al Hazzouri (2023, February). Practical Guide to Honest Causal Forests for Identifying Heterogeneous Treatment Effects. *American Journal of Epidemiology*, kwad043.

Jyotsna, J. and M. Ravallion (2005). *Insurance Against Poverty*. OUP Oxford. Google-Books-ID: mFvcugNZQUMC.

Karlan, D. and J. Zinman (2011). Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation. *Science 332*(6035), 1278–1284. Publisher: American Association for the Advancement of Science.

Karlan, D. and J. Zinman (2019, July). Long-Run Price Elasticities of Demand for Credit: Evidence from a Countrywide Field Experiment in Mexico. *The Review of Economic Studies 86*(4), 1704–1746.

Knittel, C. R. and S. Stolper (2019, December). Using Machine Learning to Target Treatment: The Case of Household Energy Use.

Lokshin, M. and M. Ravallion (2004, September). Household Income Dynamics in Two Transition Economies. *Studies in Nonlinear Dynamics & Econometrics 8*(3). Publisher: De Gruyter.

Meager, R. (2022, June). Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature. *American Economic Review 112*(6), 1818–1847.

Mullainathan, S. and J. Spiess (2017, May). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives 31*(2), 87–106.

Ravallion, M. (2009, August). How Relevant Is Targeting to the Success of an Antipoverty Program? *The World Bank Research Observer 24*(2), 205–231.

Robins, J. M., L. Li, R. Mukherjee, E. T. Tchetgen, and A. v. d. Vaart (2017, October). Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics 45*(5), 1951–1987. Publisher: Institute of Mathematical Statistics.

Tarozzi, A., J. Desai, and K. Johnson (2015). The Impacts of Microcredit: Evidence from Ethiopia. *American Economic Journal: Applied Economics 7*(1), 54–89. Publisher: American Economic Association.

Vabalas, A., E. Gowen, E. Poliakoff, and A. J. Casson (2019, November). Machine learning algorithm validation with a limited sample size. *PLOS ONE 14*(11), e0224365. Publisher: Public Library of Science.

Vapnik, V. (2013, June). *The Nature of Statistical Learning Theory*. Springer Science & Business Media. Google-Books-ID: EqgACAAAQBAJ.

Wager, S. and S. Athey (2018, July). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association 113*(523), 1228–1242. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2017.1319839.