Title

Grow the pie or have it? Using machine learning for impact heterogeneity in the Ultra-poor Graduation Model

Authors
Chowdhury, Reajul
Ceballos-Sierra, Federico
Sulaiman, Munshi

Publication Date
2021-07-21

DOI
10.26085/C3MG6Z

Series Name: WPS
Paper No.: 170
Issue Date: 21 Jul 2021

Grow the pie or have it? Using machine learning for impact heterogeneity in the Ultra-poor Graduation Model

Reajul Chowdhury, Federico Ceballos-Sierra and Munshi Sulaiman

Recommended Citation:

Chowdhury, Reajul, Ceballos-Sierra, Federico, Sulaiman, Munshi (2021): Grow the pie or have it? Using machine learning for impact heterogeneity in the Ultra-poor Graduation Model. CEGA Working Paper Series No. WPS-170. Center for Effective Global Action. University of California, Berkeley. Text. https://doi.org/10.26085/C3MG6Z

Grow the pie or have it?
Using machine learning for impact heterogeneity in the Ultra-poor Graduation Model[*]

April 11, 2021

Reajul Chowdhury, Federico Ceballos-Sierra, Munshi Sulaiman

Abstract

Anti-poverty interventions often face a trade-off between immediate reduction in poverty, measured by consumption, and building assets for longer-term gains. An "Ultra-poor Graduation" model, found effective on both dimensions, generally leans towards asset building. By using data from a large-scale RCT in Bangladesh, we find significant variation in impact on assets where the top quintile of gainers had an impact of 3.44 on their log of assets compared to the impact of 1.92 observed by the bottom quintile. We also find heterogeneity in household expenditure although the estimates are less robust across different estimation methods. Importantly, we find contrasts in characteristics of beneficiaries who made the most in assets vs. consumption. The results identify beneficiary characteristics that can be used in targeting households either to maximize impact on the desired dimension and/or to customize interventions for balancing the asset and consumption trade-off.

Keywords: Ultra-poor, Impact heterogeneity, Machine Learning, Bangladesh

JEL Classification: O12, I39

Introduction

Achieving the global ambition of ending poverty in all its forms everywhere by 2030, as postulated in the first sustainable development goal (SDG), will require scaling up successful poverty alleviating programs that not only lift people out of poverty but also can sustain the impacts over long-term. However, there is an inherent trade-off in anti-poverty programs between making immediate reduction in poverty (generally measured by consumption or expenditure) and encouraging asset accumulation for relatively longer-term change in poverty. This trade-off is indirectly discussed in poverty trap literature where the empirical evidence is mixed, but asset-based poverty dynamic generally found to be more salient than consumption or income-based measures (e.g. Ikegami et al. 2016; Carter and Barrett 2006; Quisumbing and Baulch 2013). In a more recent paper, Balboni et al (2021) find evidence of the poverty trap by looking at the impact of asset transfer, where being above a threshold results in asset accumulation, and those below face asset depletion. In terms of impact analysis of different types of interventions, a meta-analysis of cost-effectiveness of alternative livelihood support interventions by Sulaiman (2018) also reveals this trade-off whereby unconditional

[*] We are grateful to Dr. Benjamin Crost, Dr. Alex Winter-Nelson, and Dr. Imran Matin for their guidance and feedback on the analysis. Special thanks to Dr. Stefan Wager for prompt responses on use of the GRF package and interpreting the coefficients. We also acknowledge BRAC and LSE research team who worked on the project that our data comes from. All errors are ours. Corresponding author Reajul Chowdhury mac7@illinois.edu

cash transfers are more attractive in the short-run whereas more comprehensive interventions fare better in the long run. Therefore, being able to identify characteristics of households who are likely to have lower impacts on assets can improve the efficiency of programs that are focused on livelihoods development by giving more emphasis on their asset building through varying transfer amounts and/or technical supports.

In this paper, we look at this trade-off among the beneficiaries of an ultra-poor graduation model, which is considered as an effective approach for addressing poverty in the short run, and the impacts are also sustained over several years post interventions. Although the average effects of this model are generally positive for both asset accumulation and consumption, this does not imply the nonexistence of a trade-off between the two domains. Using machine learning tools, we investigate whether there are systematic differences between the participants of a graduation program in Bangladesh who gain more in household expenditure over asset accumulation and vice versa.

Several studies have shown robust evidence of the graduation model being successful in reducing extreme poverty in a wide range of contexts (Bandiera et al. 2017; Banerjee et al. 2015)[1]. The intervention model is composed of a sequence of supports including a grant of productive assets, hands-on coaching for 12-24 months, life-skills training, short-term consumption support, and access to financial services. The goal of this model is to develop micro-enterprise from the transferred assets while all the other components are related to protecting their enterprise and/or increasing productivity. Developed by BRAC, this model has shown significant impacts on household asset accumulation, consumption, labor supply, income, and food security status in Bangladesh (Bandiera et al. 2017; Banerjee et al. 2015). More importantly, these impacts sustain well beyond the 2-year intervention period. A six-country replication of the model has also shown similar positive results. Long-term follow-up studies show the impacts not only persist but also grow over 7 to 10 years in West Bengal (Banerjee et al. 2016; Duflo 2020) and up to 14 years in Bangladesh (Balboni et al. 2020). Evaluations of variations of the graduation model also produce similar positive results (Blattman et al. 2016; Gobin et al 2016; Sedlmayr et al 2020). Currently, the model has been adopted by various NGOs and in government social protection schemes in 75 countries by 2020 (Andrews et al. 2021).

Because of the comprehensive package, the graduation model generally costs substantially more than alternative poverty alleviation approaches (Sulaiman 2018). One of the avenues of improving cost-effectiveness is to better customize the supports to the needs of specific sub-groups within the target population. In fact, the existing studies find a large degree of heterogeneity even within the narrow group of the poorest households (e.g. Bandiera et al. 2017; Banerjee et al. 2015). However, these studies followed the traditional approach of evaluating heterogeneity with respect to only a few predetermined well-known covariates, and therefore, left other potential sources of heterogeneity unexplored. Understanding the source of heterogeneity in the effects of the Graduation Model has potential policy implications for implementing agencies to bring necessary changes in their targeting approaches as well as customizing intervention packages to fit the needs of different sub-groups of the extreme poor.

The typical approach of analyzing heterogeneous effects involves fitting a linear model which includes interactions between treatment and the covariates, essentially measuring the treatment effects for subgroups. However, this econometric approach is both inefficient and less robust as deciding on a few variables to create the subgroups involves the risk of overfitting the estimates (i.e. selecting only those variables on which we see heterogeneity) and throwing away the rich set of baseline

[1] Bandiera et al (2017) evaluate the model in Bangladesh and Banerjee et al (2015) evaluate the same approach in six countries.

information available (Chernozhukov et al. 2020). Execution of such model becomes more challenging when the number of covariates significantly increases; including all the potential interaction terms becomes infeasible unless the sample size is sufficiently large relative to the number of covariates and their interaction terms (Foster, Taylor, and Ruberg 2011; Green and Kern 2012; Imai and Ratkovic 2013; Schiltz et al. 2018). To overcome these weaknesses, several recent studies proposed using techniques from machine learning (ML) to better understand heterogeneous effects (Athey and Imbens 2017; Chernozhukov et al. 2020). This new and growing literature has proposed several parametric, semi-parametric, and non-parametric approaches that utilize a large array of covariates, are computationally feasible, and avoid the risk of overfitting. Consequently, the use of ML in randomized control trials (RCT) to make inferences on heterogeneous treatment effects is receiving increasing attention (Chernozhukov et al. 2020; Foster et al. 2011; Imai and Ratkovic 2013).

In this paper, we apply two ML approaches to investigate the heterogeneity in the effects of the graduation model from an RCT in Bangladesh. We use three rounds of panel data from 5,491 ultra-poor households, who were randomized into a treatment and control group. We begin with estimating the Conditional Average Treatment Effect (CATE) of the graduation program using the Honest Causal Forest (HCF) algorithm proposed by Susan Athey and Wager (2019). We favor the HCF method for two reasons: first, by construction, it allows us to flexibly model complex interactions and discontinuous relationships between independent variables, and second, it allows for valid hypothesis testing and the estimation of standard errors and confidence intervals. Next, we use the ML approach proposed by Chernozhukov et al. (2020), which involves estimating proxy predictors of CATE and then developing valid inference on key features of the CATE. While the causal forest method focuses only on tree-based random forest tools to produce consistent estimates of CATE and explore heterogeneity, the approach proposed by Chernozhukov et al. (2020) is more general and can be applied to any ML methods to predict and make inference on heterogeneous effects. Since both approaches resolve the fundamental problem of nonparametric inference of ML methods and propose strategies that produce uniformly valid inference, we use them as complementary to each other. Subsequently, we conduct classification analysis to identify baseline characteristics that are associated with the impacts to understand the trade-off and policy implications on customizing interventions.

For measuring heterogeneous impacts, we focus on two outcomes – household wealth and expenditure. Our results detect a large degree of heterogeneity in treatment effects on both assets and expenditures. However, the difference between the top and least gainers in household wealth is much larger than the corresponding difference in expenditure. Looking at heterogeneity by baseline characteristics, the results indicate a trade-off between the gains in wealth vs consumption. We find 15 common variables that are important determinants of impact heterogeneity in both assets and expenditure, and the direction of their relationships with the impact sizes for almost all the indicators are in opposite to each other between the two outcomes. In terms of specific indicators, the age of participants is found to be an important factor whereby top gainers of wealth are more likely to include older participants whereas those who show a high impact on consumption are more likely to consist of younger beneficiaries. This trend of older beneficiaries accumulating wealth while younger beneficiaries having higher consumption gain is in contrary to a common understanding of the model being less effective for older people. Another dimension of heterogeneity that comes out as significant is women's voice in household decision-making. We find that women with greater involvement in household decision-making at baseline are more likely to be in the high impact groups when it comes to expenditure and the other way around for wealth impact. Besides these participant characteristics, other factors showing significant heterogeneity in impacts are households' baseline level of savings, assets, expenditure, community-level variables of distance to market, and paved roads. We also find inequality in livestock ownership within the communities having significant impact heterogeneity whereby households with high expenditure gains are more likely to reside in communities with high asset inequality, but no significant difference in impact on assets.

Our results of impact heterogeneity going in opposite directions for expenditure and wealth by most of their baseline characteristics demonstrate the trade-off between achieving impacts on immediate poverty reduction by increasing consumption and more long-term impacts through asset accumulation. We also find household characteristics that are more likely to have impacts on consumption over the asset accumulation that can be used for potentially customize interventions, in the context of Bangladesh, to strengthen the long-term impacts of the model. One example of customization to improve cost efficiency can be reallocating greater staff time to low asset gainers instead of uniform coaching support that is currently being practiced. The rest of the paper is organized as follows. Section II gives a brief overview of the Graduation Model and the evidence of its impact. Section III describes the data used in this paper. Section IV description of the two ML approaches that we apply in this paper for exploring the heterogeneity in treatment effects. Study results are discussed in Section V and the conclusion in Section VI.

I. Graduation model and evidence of its impact

Over the last few decades, development organizations learned that bringing people out of ultra-poverty requires simultaneously addressing multiple constraints that they face in moving towards a sustainable livelihood. Building on this insight, BRAC, an NGO originating in Bangladesh, pioneered a program called Targeting the Ultra-poor (TUP) to build secure, sustainable, and resilient livelihoods for the ultra-poor (Matin et al. 2008; Morel and Chowdhury 2015). The approach in the TUP program, now better known as the "graduation model", is to combine multifaceted support services addressing both the immediate needs of the ultra-poor by giving them consumption supports, and their long-term need for a sustainable livelihood by providing them a grant of productive assets with technical skills training. This is complemented by a time-bound (typically 18-24 months) intensive coaching, access to finance, and health supports to both improve their productivity of the assets and protect them from distress sales due to shocks. BRAC started implementing the program in 2002 in Bangladesh.

Several non-experimental studies (e.g. Ahmed, Sulaiman, and Das 2009; Matin and Hulme 2003; Mallick 2013) found the program very effective in increasing household consumption, asset holdings, and self-employment among the ultra-poor. The holistic treatment of poverty in the graduation approach drew the attention of the donor community and other stakeholders in low-income countries. The model has later been replicated and adapted by at least 219 programs in 75 countries by NGOs, governments, and donor organizations (Banerjee et al. 2015; Andrews et al. 2021). Taking advantage of the large-scale replication of the model in low-income countries, several high-quality randomized trial studies have been conducted to assess the impact of the model (Bandiera et al. 2013, 2017; Banerjee et al. 2015). In Bangladesh, Bandiera et al. (2013) found that after four years of the program inception, the beneficiary households expanded their self-employment activities, increased labor supply, accumulated more productive assets, which led to increased household income and per capita consumption. A follow-up survey on the same households seven years after the program began, found that the long-term effect of the program is at least as large as the four-years effect (Bandiera et al. 2017). Banerjee et al. (2015) documented the findings from 6 randomized trial studies assessing the impact of the graduation model implemented in 6 countries. The study found the effect of the program on income, household asset accumulation, food security, and consumption similar to the Bangladesh study albeit with some variations across the 6 sites.

While the effects of the graduation model have been found to be positive and durable in a wide range of geographical and cultural contexts, the existing studies also reported a high degree of heterogeneity in the effects. For instance, Bandiera et al. (2017) showed that the effects on consumption, savings, and productive assets accumulation at 95th percentiles were at least 10 times larger than the effect at the 5th centile of the distribution. Similar significant variation in treatment effects on household income, consumption, food security, and financial inclusion was also reported in Emran, Robano, and Smith (2014); Raza, Das, and Misha (2012); and Banerjee et al. (2015). The large degree of heterogeneity in treatment effects implies that even with the narrow group of the ultra-poor, there could be subgroups who are not benefitting as much. However, the quantile regression approach in these studies has the limitations of only measuring impact across different percentile on a continuum of an outcome indicator and does not confirm whether these are associated with any baseline characteristics. This paper aims to identify and define these subgroups, that will help adopt changes in the graduation model to better fit the needs of these groups.

II. Data

We use the data from a cluster-randomized trial by Bandiera et al (2017) that assessed the impact of the graduation model implemented in Bangladesh by BRAC. Starting in 2007, the study randomly assigned 40 BRAC branch offices serving 1,309 villages to the treatment or control group. We use the data from three rounds of surveys - baseline in 2007 followed by midline in 2009 and endline in 2011.[2] The baseline survey was preceded by a participatory wealth ranking exercise in both treatment and control villages, which classified households into four groups: ultra-poor, near-poor, middle-class, and upper-class. Although the impact evaluation paper looked at spillover on these groups, we focus only on the ultra-poor group as they are the targeted beneficiaries in the graduation model and received support. Our final sample size comprises 5,315 households of whom 3,082 from treated branches and 2,233 from control branches. Our analysis explores heterogeneity in treatment effects on two outcomes: the value of per-capita wealth, and per-capita household expenditure. We use the log of both outcomes. Household wealth has been calculated summing the monetary value of land, business assets, non-business assets, and savings. The household expenditure outcome includes household expenses on food (both purchased and produced), fuel, cosmetics, entertainment, transportation, utilities, clothing, footwear, utensils, textiles, dowries, education, charity, and legal expenses. While our analysis is intended to capture possible trade-offs between wealth and consumption, it is worth mentioning that both outcomes are among the key performance indicators for the graduation model. We complement our main analysis of these two outcomes with the analysis of two additional outcomes – household savings, and self-employment income.

The treatment variable is a dummy indicating if a household resides in villages under a treated BRAC branch office. The covariates for heterogeneity include baseline information on respondents' characteristics, demographic and socio-economic characteristics at the household level, and several cluster-level characteristics. Initially, we started with 103 covariates and after filtering out variables with near-zero variance, multicollinearity, and a high number of missing values, we are left with 50 covariates. The list of these covariates along with some descriptive statistics are presented in Annex (Table A1).

III. ML Method

[2] Bandiera et al (2017) also used a fourth round of survey conducted 2014. However, we do not use this since some of the households from control group were also treated after the endline.

Our empirical strategy combines two machine learning approaches; the honest causal forest algorithm proposed by Wager and Susan Athey (2018), and an agnostic approach proposed by Chernozhukov et al. (2020). The honest casual forest method builds on the causal tree algorithm proposed by Susan Athey and Imbens (2016), which partitions the data into a set of subgroups such that treatment effect heterogeneity across subgroups is maximized (Athey and Imbens 2016). In estimating the treatment effect, the causal tree algorithm follows an "honest" approach, whereby one sample is used to construct the partition (i.e. building the tree) and another to estimate treatment effects for the subgroups. More specifically, the causal forest algorithm starts by drawing a random subsample of training data and then splitting the training data into two halves *I* and *J*. The algorithm then grows a tree by using the *J*-sample data to partition the data space, while holding out the *I*-sample data for within-leaf estimation. When choosing a split, the algorithm seeks to maximize the difference in treatment effect $[\tau(\mathrm{X})]$ between the two child leaves. The treatment effect is estimated simply by taking the difference between the outcomes of the treated and control observations within a leaf:

$$\hat{\tau}(\mathrm{X}) = \frac{1}{|\{i: \mathrm{W}_i = 1, \mathrm{X}_i \in L\}|} \sum_{i:\mathrm{W}_i=1,\mathrm{X}_i \in L} \mathrm{Y}_i \quad - \quad \frac{1}{|\{i: \mathrm{W}_i = 0, \mathrm{X}_i \in L\}|} \sum_{i:\mathrm{W}_i=0,\mathrm{X}_i \in L} \mathrm{Y}_i \qquad (1)$$

In equation 1, W is the treatment indicator taking value 1 for treated observations, X is the covariate space, Y is the outcome variable, and L is the leaf within a tree. Wager and Susan Athey (2018) showed that the honest approach of tree building produces consistent estimates by eliminating bias in the CATE and enables centered confidence intervals that allow for valid statistical inference.

While the causal forests approach uses only one specific ML tool i.e. tree-based algorithm and relies on an honest approach to produce consistent estimates of CATE and explore heterogeneity, Chernozhukov et al. (2020) proposed a different approach that allows applying generic ML methods to estimate causal effects and draw a statistical inference. The empirical strategy of this approach starts by building a proxy predictor of CATE using generic ML methods, and then develop valid inference on some key features of the CATE based on this proxy predictor. Instead of obtaining consistent estimation and uniformly valid inference on the CATE itself, this approach focuses on providing valid estimation and inference on certain features of CATE. Referring to it as an agnostic approach, Chernozhukov et al. (2020) argued that by focusing on key features of CATE rather than CATE itself, this approach avoids making strong assumptions about the properties of ML estimators and yet obtain uniformly valid inference on some features of the estimators. Particularly, this approach targets to develop valid inference on three features namely – Best Linear Predictor (BLP), Sorted Group Average Treatment Effects (GATES), and Classification Analysis (CLAN).

The algorithm of this agnostic approach involves repeatedly splitting the data into two samples, namely the main sample ($Data_M$) and the auxiliary sample (Data_A), and for each split training ML methods and predict the outcome variable on the treated and untreated observations separately using the Data_A. Applying the trained ML models on Data_A, the algorithm then estimates two potential outcomes [Y(0), Y(1)], and obtain treatment effect estimates, S(X), and baseline effect estimates, B(X), for each observation in Data_M. The baseline effect and the treatment effect are estimated using the following simple equations:

$$B(X) = E[Y|W=0, X] \qquad (2)$$

$$S(X) = E[Y|W=1, X] – E[Y|W=0, X] \qquad (3)$$

The algorithm then involves testing for heterogeneity in the treatment effects using the following weighted ordinary least squared (OLS) or the Best Linear Predictor (BLP) model:

$$\mathrm{Y_i = \alpha\, X_i + \alpha_1\, B(X_i) + \beta_1\, (W_i - p(X_i)) + \beta_2\, [W_i - p(X_i)]\, [S_i\,(X_i) - ES) + \varepsilon,}$$
$$\text{with weights } \omega(\mathrm{X}) = \frac{1}{\{\mathrm{p(X)\,[1-p(X)]}\}} \qquad (4)$$

where ES = $\frac{1}{M}\Sigma S(X_i)$, and p(X)= $\frac{1}{N}\,\Sigma p(W_i = 1\,|X)$ or $\frac{N\,treated}{N}$ for a randomized trial study. In our case W = $\frac{3082}{5315}$ = 0.58.

β_1 in equation 4 indicates the average treatment effect. β_2, the main coefficient of our interest, indicates the degree to which the estimated treatment effect, $S_i(X)$, serves as a proxy for the true treatment effect or CATE. Rejecting the null hypothesis $\beta_2 = 0$ means that there is heterogeneity and $S_i(X)$ is a relevant predictor.

The second feature, Group Average Treatment Effect (GATE), involves dividing the main sample into non-overlapping groups G_1 to G_K based on the predicted treatment effect $S_i(X)$. If we decide to have k=5, then the resulting group G_1 will be the 20% of the data with the lowest treatment effect estimates and G_5 will be the group with the highest treatment effect estimates. The GATE parameters are estimated as follows:

$$\mathrm{E[\,S_i(X)\mid G_k\,]}, \quad k = 1, \ldots..\, K \qquad (5)$$

Where G_k is non-overlapping intervals dividing the $S_i(X)$ into *K* groups.

Finally, the third feature, Classification Analysis (CLAN), helps to characterize the most and least affected groups by identifying the baseline covariates on which the groups differ from each other. Assuming g(Y, X) is a vector of characteristics of an observational unit, the average characteristics of the most and least affected groups can be denoted by the following parameters:

$$\Upsilon_1 = \mathrm{E\,[g\,(Y, X)\mid G_1]} \text{ and } \Upsilon_k = \mathrm{E[g\,(Y, X)\mid G_k]} \qquad (6)$$

Our main results comprise of the treatment effects estimation from the causal forest method, and then using those estimations to construct the group average treatment effect (GATE) and the classification analysis (CLAN). We also complement our analysis of causal forest results with the results from other generic ML methods. We use the causal forest as our preferred method of estimation for two reasons. First, the treatment effect estimates from the causal forest are unbiased and allow for valid statistical inference. The 'honesty' approach used in the causal forest model addresses the fundamental problem of causal inference and allows for a direct estimation of causal effect while eliminating bias from the estimates. Second, causal forest follows a data-driven approach in identifying the most important variables from a large set of predictors used in growing the forest. We use this subset of predictors to perform the classification analysis to characterize the most and least affected groups and avoid the clumsiness of using the large set of baseline predictors for classification analysis.

We fit our models to reflect the heterogeneity in treatment effect at the household level rather than at the cluster level (e.g. branch or village level). Since the treatment was randomly assigned at the branch office level, we assume that the branch level effect or village level effect on TUP beneficiaries is normally distributed. Therefore, we train the causal forest model without clustering by branch or villages. However, we include subdistricts fixed effects, which was used for stratifying the treatment-control assignment. We also include some spot-level Gini coefficients as covariates in our model to see if the distribution of wealth, income, productive assets, and household assets in the immediate neighborhood of the beneficiaries influence treatment effect and heterogeneity in treatment effect.

We grow the causal forest following the generalized random forest (GRF) framework proposed by Susan Athey, Tibshirani, and Wager (2019). We first orthogonalize the treatment and the outcome variables by fitting a regression forest to estimate the expected outcome marginalizing over treatment (see Susan Athey, Tibshirani, and Wager 2019 for detail). Using the estimates from this regression forest, the GRF then makes out-of-bag predictions to be used as inputs in our causal forests. Following Basu et al. (2018) and Susan Athey and Wager (2019), we also train a pilot causal forest on all covariates. Then, we train our final causal forest using only those covariates that were found important in growing the pilot forest. In selecting the important covariates, we use the 'variable_importance' function of the GRF package which assigns a score for each covariate by taking a simple weighted sum of how many times the covariate was chosen by the algorithm in building trees. We select those covariates whose 'variable_importance' score is above average. This approach improves the precision of our estimation as it enables the forest to make more splits on the most important features (Athey and Wager 2019). These important features identified from the pilot forest have also been used for our classification analysis (CLAN) to characterize the most and least affected households.

As a complement to causal forest estimations, we used four generic machine learning (ML) methods namely Elastic Net, Boosting Tree, Neural Network, and Random Forest method in estimating CATE. Similar to how we trained the Causal Forest model, we also controlled for subdistrict level fixed effects in training these generic ML methods.

IV. Results and Discussion

Following the ML approaches described above, our discussion focuses on two sets of results - degree of heterogeneity (HET) in treatment effect and classification analysis (CLAN) for the heterogeneity in treatment effects on per-capita household wealth and per-capita household expenditures. Among the four generic ML methods, we rely more on the results from the random forest and elastic net methods since these two methods outperformed the other two (boosting tree and neural network) in terms of their ability to detect greater heterogeneity in the treatment effect estimates[3].

A. Average vs. Heterogeneous Treatment Effects

Table 1 presents coefficients of the average treatment effect (ATE) and the degree of heterogeneity (HET) for the two outcomes. The ATE estimates for both outcomes are large, positive, and significant across the causal forest model and the two generic ML methods. The causal forest estimates of ATE coefficients for the log value of per-capita wealth and household expenditures are 2.54 and 0.14 respectively indicating that the per-capita wealth and expenditures increased by 254% and 14% among the ultra-poor women in the treated areas relative to the control areas, both significant at 1% level. Reassuringly, the estimated ATE coefficients from all four ML methods closely match the OLS estimates, both in terms of magnitude and statistical significance.

Table 1. Average Treatment Effects (ATE) and Heterogeneous Treatment Effect (HET)

Method		Per Capita Wealth (log)	Per Capita Expenditure (log)
Causal Forest	ATE	2.54	0.14
		[2.13 2.95]	[0.07 0.21]
		(0.000)	(0.000)
	HET	1.30	0.96
		[1.00 2.00]	[-1.00 3.00]

[3] Following Chernozhukov et a. (2020), we choose the best ML methods that maximize the criterion function $\Lambda = |\beta_2|^2 \text{Var}(S(X_i))$

		(0.000)	(0.123)
Random Forest	ATE	2.41	0.14
		[2.29 2.53]	[0.12 0.17]
		(0.000)	(0.000)
	HET	0.94	0.30
		[0.79, 1.09]	[0.11 0.48]
		(0.000)	(0.003)
Elastic Net	ATE	2.38	0.14
		[2.26 2.50]	[0.12 0.16]
		(0.000)	(0.000)
	HET	0.98	0.18
		[0.82 1.15]	[-0.03 0.39]
		(0.000)	(0.198)
Neural Network	ATE	2.45	0.14
		[2.32 2.57]	[0.12 0.17]
		(0.000)	(0.000)
	HET	0.17	0.06
		[0.12 0.22]	[-0.01 0.14]
		(0.000)	(0.184)
Boosting Tree	ATE	2.42	0.14
		[2.30 2.53]	[0.12 0.16]
		(0.000)	(0.000)
	HET	0.39	0.08
		[0.31 0.48]	[-0.02 0.15]
		(0.000)	(0.229)
OLS	ATE	2.51	0.13
		[2.24 2.79]	[0.10 0.16]
		(0.000)	(0.000)

Note: 95% confidence intervals in brackets, and p-value in parenthesis. In estimating the ATE for the causal forest, we used the built-in function in the GRF package. For the generic ML methods, we applied the BLP test that estimates both ATE and HET using equation 4. The covariates used in the OLS models include respondent's age, respondent's education years, household head's gender, log per-capita expenditure, log per capita food consumption, log per capita wealth, wage income, income from self-employment activities, Gini score for livestock, and distance to the nearest market.

Table 1 also reports the treatment effect estimates from ordinary least square (OLS) regression models fitted on each outcome with the treatment variable and several covariates at respondent, household, and neighborhood levels. The OLS estimates of the average treatment effects for the two outcomes are 251%and 13%, both significant at 1% level, which consistently match with the estimates from machine learning methods. These results basically reproduce the conclusions drawn by Bandiera et al (2017) from this data.

Turning to the heterogeneity in treatment effect, the coefficients for HET are 1.30 for per-capita wealth and 0.96 for per-capita expenditure when we use the causal forest method. The non-zero coefficients indicate that the causal forest estimates of the treatment effects are important relevant predictors.[4] Based on the p-value of the HET coefficient for the wealth outcome, we reject the hypothesis of no heterogeneity at the 1% level, suggesting that there is a significant heterogeneous effect of the TUP intervention on this outcome. The heterogeneity coefficients from the random forest and elastic net also confirm the same high level of heterogeneity in the treatment effect in per-capita wealth. However, the p-value associated with the HET coefficient of the expenditure outcome failed to reject the hypothesis of no heterogeneity at 10% level of significance by causal forest estimate (p value=0.123). Similarly, elastic net, neural network, and boosting tree methods also show weak heterogeneity and are consistent with the causal forest estimation while random forest shows a significant level of treatment effect heterogeneity (at 5% level). In other words, graduation interventions produced a highly diverse level of impact in asset accumulation by beneficiary households while the impact variations for consumption are less robust.

B. Group Average Treatment Effects (GATES):

While the HET coefficients are useful in understanding the existence of heterogeneity, they do not reveal the magnitudes of differences. To check the magnitude, we use group average treatment effects (GAES) that involve dividing the observations into different subgroups according to their effect sizes. Specifically, we divide the sample into 5 groups based on the quintiles of the estimated $\hat{\tau}(X)$ from our causal forest model and the generic ML proxy predictor $S_i(x)$ and estimate the average effect for each group. Next, we compare the GATES between the top and the bottom quantiles, alternatively called the most and the least affected groups.

As shown in Table 2, the causal forest estimates show that the differences of average treatment effects between the most and least affected groups are significantly different from zero at 1% level for both wealth and expenditure outcomes. The average treatment effect among the most affected households on log per-capita wealth is 3.44, which is 79% higher than the average treatment effect among the least affected households. Likewise, the average treatment effect of the most affected groups on log per-capita expenditure is 63% higher compared to that in the least-affected households. For generic ML methods, we present the results from the random forest and elastic net only as the estimations from these two methods have been found more efficient in detecting heterogeneity in our data. The GATES estimates from the random forest and elastic net also report significant differences between the two groups in both outcomes.

In Figure 1, we present the box-plot distribution of GATES scores and confidence bands for the five quantile groups using the causal forest estimates. The figure also shows ATE and associated confidence intervals obtained from the casual forest. This is apparent from the box-plot distribution that the treatment effects are positive on both outcomes for all sub-group of households, and the graduation model did not adversely affect any beneficiary households.[5] The figure also reveals that the top and bottom 20th quantile groups (group 1, and 5) are less symmetric than the ones in the middle, with positive skew on the top quantile group and negative skew in the bottom quantile group.

[4] See Chernozhukov et al. 2020 for more technical details on BLP estimates. However, the coefficient value greater than 1 implies that the random forest predictions are over-shrunk, and the CATE estimates from the forest under-estimate the true treatment heterogeneity. For example, suppose the random forest gives us a CATE estimate $\hat{\tau}(X) \approx \frac{\tau(X)}{2}$. Then calibration would give us a coefficient of roughly 2. (We are grateful to Stefan Wager, Assistant Professor of Statistics in Stanford University, for this explanation).

[5] Although this appears contradictory to the findings of asset depletion for those below threshold Balboni et al (2021), but the key distinction is in timeline. Their long-term follow-up look at asset dynamics after the endline and asset depletion does not imply non-positive long-term impact.

We looked into whether the differences in treatment effects between the mid-quantile groups (e.g. group 3rd vs 1st, and 4th vs 2nd) are statistically significant or not (see Table A2 in the appendix). The test shows significant differences between the mid-quantile groups on both outcomes.

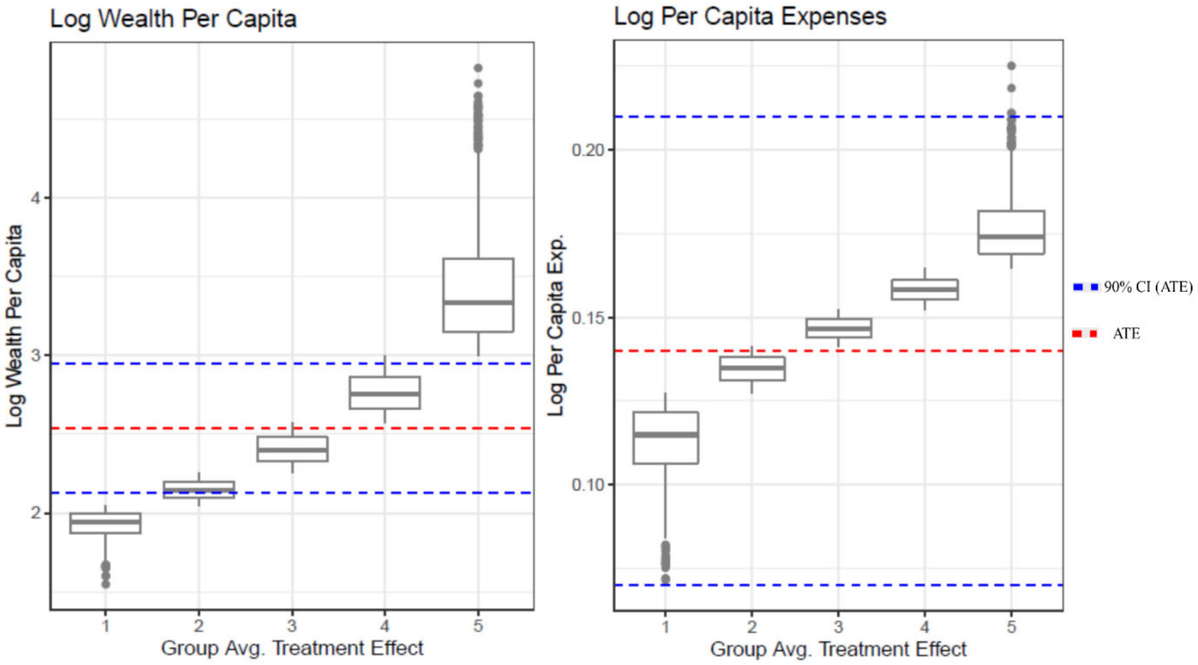
Table-2: Group Average Treatment Affect (Top 20% vs Bottom 20%)

		Per Capita Wealth (log)	**Per Capita Expenditure (log)**
Causal Forest	ATE of Most 20%	3.44	0.18
		[3.32 3.55]	[0.17 0.18]
	ATE of Least 20%	1.92	0.11
		[1.81 2.04]	[0.11 0.12]
	Diff. (Most vs Least)	1.52	0.06
		[1.35 1.68]	[0.05 0.07]
		(0.000)	(0.000)
Random Forest	ATE of Most 20%	3.55	0.17
		[3.29 3.81]	[0.12 0.22]
	ATE of Least 20%	1.71	0.09
		[1.44 1.99]	[0.04 0.14]
	Diff. (Most vs Least)	1.84	0.10
		[1.46 2.22]	[0.03 0.17]
		(0.000)	(0.011)
Elastic Net	ATE of Most 20%	3.35	0.17
		[3.09 3.60]	[0.12 0.22]
	ATE of Least 20%	1.56	0.10
		[1.28 1.84]	[0.05 0.15]
	Diff. (Most vs Least)	1.78	0.09
		[1.40 2.16]	[0.02 0.16]
		(0.000)	(0.037)

Note: 95% confidence intervals in brackets, and p-value in parenthesis. The group average treatment effects have been estimated using equation 5 specified in the Method section.

Overall, our heterogeneity analyses show consistent results for assets and less so for expenditure. Heterogeneity test using the best linear predictor (BLP) model applied on the estimates of the causal forest, and generic ML methods found detectable heterogeneity in per-capita wealth outcome. We did not find strong heterogeneity in treatment effects on per-capita total expenditure from the causal forest and the majority of the generic ML methods. However, the group average treatment effect analysis (GATES) from the causal forest and some of the generic ML methods detected a significant level of heterogeneity in treatment effects in both outcomes.

Figure 1: Box-plot of causal forest estimates of GATES by quantiles



C. Classification analysis:

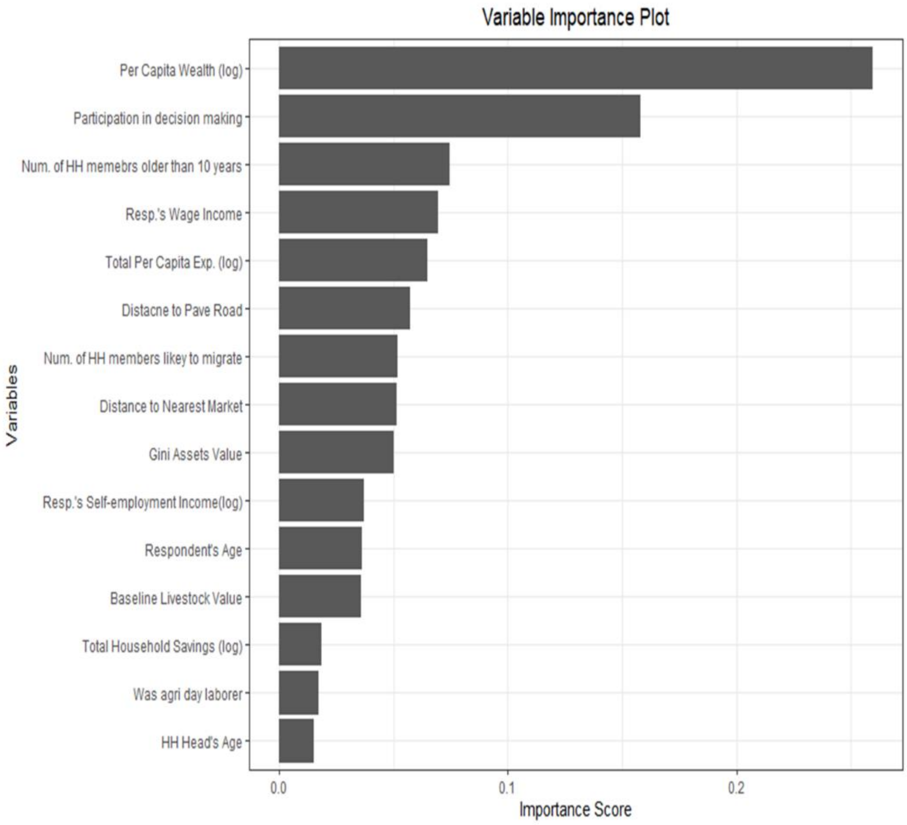
While the analysis presented so far reveal significant heterogeneity in wealth and relatively less consistent heterogeneity in consumption, the HET coefficients do not tell us who are the most and least affected households. To understand the differences in characteristics of the most and least affected groups, we look at the average characteristics of the two groups by Classification Analysis (CLAN) developed by Chernozhukov, Fernández-Val, and Luo (2018). For this, we use the estimations from the causal forest method since it offers a data-driven approach in determining which variables are most important in estimating the treatment effect in the sample. Following the selection criteria for the most important covariates (as explained in the method section), we get a list of a total of 15 variables. Figure 2 below shows the list of these important variables and their corresponding importance score. These variables fall into three groups: individual or primary beneficiary level characteristics, household-level characteristics, and neighborhood or community-level characteristics. The most important baseline characteristic is their wealth level followed by participant woman's voice in household decision making. These two indicators have much larger importance scores compared to the rest. The important factor from household demography is the number of members aged above 10 at baseline. Three community-level variables that are in the middle of this important variable list are – distance to paved roads and market and asset inequality. Interestingly, the age of the participant and household head, which are often considered important dimensions of heterogeneity are ranked lower in this list.

While this list shows us the important factors behind impact heterogeneity, the next step is to understand their direction of relationship with impact estimates. In Figure 3, we visualize the classification analysis by plotting the differences between the most and least affected groups on these fifteen baseline variables.[6] The figure shows whether the difference between the two groups is

[6] see Table A3 in Appendix for detail estimates.

positive or negative, where a positive difference means the average of the most affected group is higher than that of the least affected group, and vice versa.

Figure 2: Covariates used most often in building trees

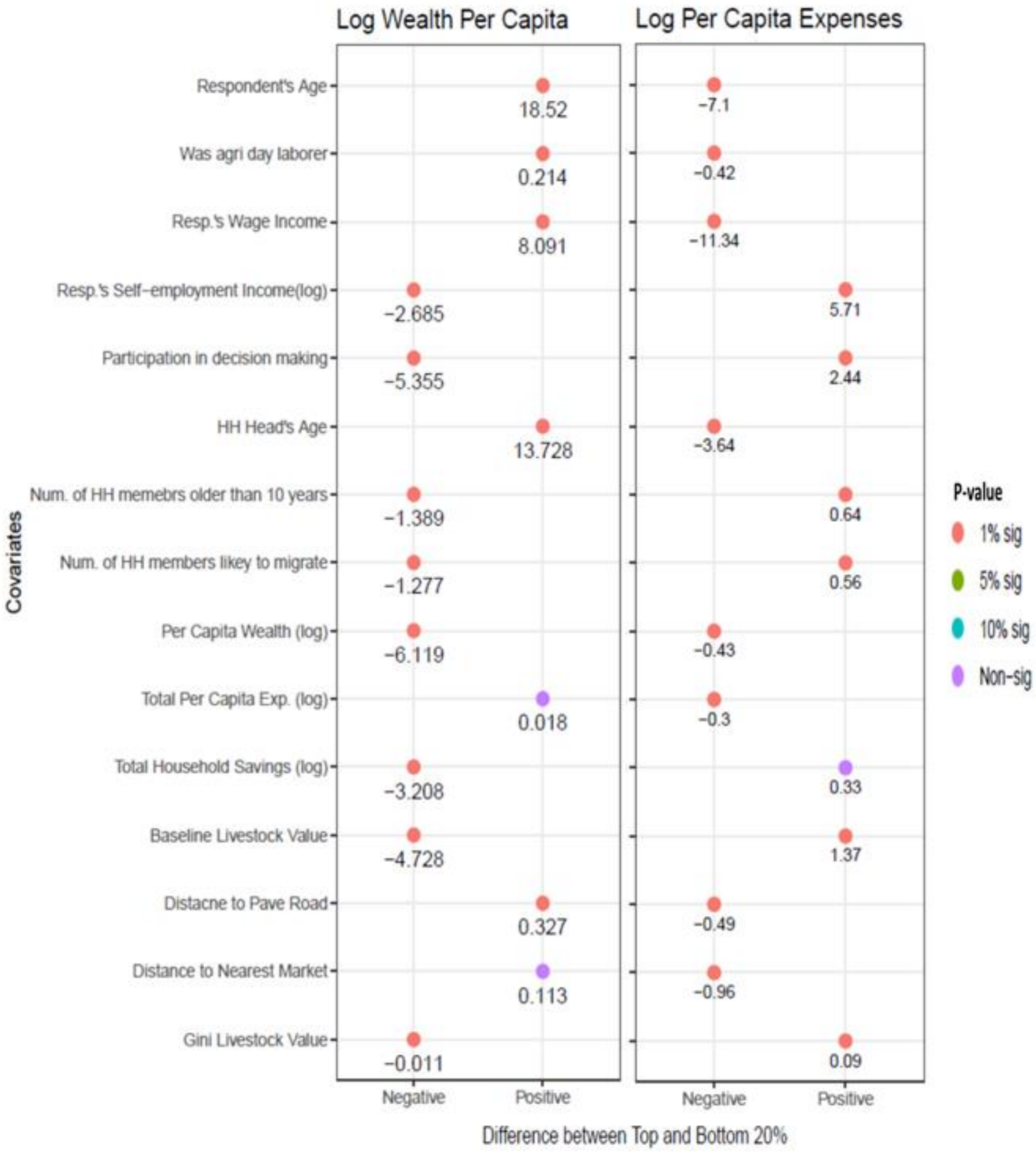


The primary beneficiaries (respondents) among the most benefitted households in per-capita wealth outcome were relatively older, were more dependent on wage income (mostly from agricultural labor works), had less involvement in self-employment activities, and had lower participation in household decisions making at baseline. More specifically, they were 19 years older, and 21 percent more likely to work as agriculture labor than the respondents from the least affected households on this outcome. Log of their income from daily wage activities was 8.09 greater, and from self-employment activities was 2.69 lower compared to the respondents from the least affected households. Their score on a matrix measuring their participation in household decision-making was 5.36 points less; their empowerment score was 1.91 compared to 7.27 of the other group.

Looking at the household level characteristic, the top gainers in wealth have household heads who are on average 14 years older than the heads in the least affected group. These households also have fewer members above 10 years old (by 1.34 members), who were less likely (by 1.28 percent) to migrate out of villages for work. This group of top gainers had higher per-capita household expenditures, and lower savings at the baseline. Compared to their counterparts in the least-gainers group, these households' log value of per-capita expenditures was higher by 0.18 (though not statistically significant), and the log value of savings was lower by 3.21. Finally, the log value of livestock of the top gainers was lower by 4.73.

Regarding the community-level characteristics, the households that benefited the most in wealth accumulation are more likely to live in communities far away from paved road and markets; their communities are 0.11 km farther away from the nearest markets (not significant at conventional level), and 0.33 km farther away from paved roads, compared to the communities of the least affected households. The Gini coefficient, measuring the distribution of livestock value in neighborhoods, shows that the most affected households for the wealth outcome live in communities with lower inequality (by 0.01 points).

Figure 3: Classification analysis of most and least affected groups



When it comes to gains made in per-capita household expenditure, we find that the characteristics of the most affected households are generally in opposite direction (for 14 out of the 15 variables) of what we found in the wealth accumulation outcome. The primary respondents of the most affected households in this outcome are younger by 7 years compared to those in the least affected households. They also differ from the top gainers in the wealth outcome in terms of having less wage income (by 11.34 points), more self-employment income (by 5.71 points), and higher participation in household decision making (by 2.44 points).

Turning to the household level variables, most affected households for this outcome are headed by relatively younger (by 4 years) people, had more members older than 10 years, and had members who were more likely (by 56%) to migrate out. These households had lower per-capita expenditures at the baseline: their baseline expenditures were lower by 0.3 points than that among the least affected households. They also had higher savings and livestock assets. Their savings and livestock assets value, respectively, was higher by 0.33 points (not significant at conventional level), and by 1.37

points (p <0.01). At the community level characteristics, unlike the differences between most and least gainers in the wealth outcome, households that increased their expenditure the most were living in communities closer to markets and paved road, by 1 kilometer and 0.49 kilometers respectively. Finally, unlike the top gainers in wealth outcome, the most affected households in expenditures outcome live in communities with greater inequality in the distribution of livestock assets (by 0.09 points).

Finally, top gainers of both asset and consumption had less per-capita wealth at the baseline compared to the least affected group. This variable was scored as the most important (Figure 2) and the only variable that has the same direction of relationship with impacts on both asset and consumption. This highlights the possibility of maximizing impact by targeting explicitly asset ownership.

One might wonder whether the top gainers on these two outcomes also differ from each other on other welfare dimensions. This might happen that households that increased their consumptions by most had been able to generate more income or accumulate more savings than those who gained most in assets, and vice versa. To answer these questions, we also investigated the treatment effect heterogeneity on the log of household savings and self-employment income (i.e., income from livestock and small businesses) and conducted the CLAN for these two outcomes using the same variables we used for the CLAN of assets and consumption. Figure A1 in Appendix shows the CLAN for household savings and self-employment income. The top gainers in the savings outcome closely resemble the top gainers in assets accumulation in almost all of the baseline characteristics, implying that households that gained most in assets also accumulated savings. On the other hand, the CLAN for the self-employment income failed to show significant differences between the top and the least gainers in half of the baseline characteristics (7 out of 15). On the remaining eight characteristics, the top gainers in income do not distinctively match with the top gainers in either assets or consumption outcomes. Our results also showed weak heterogeneity in the treatment effects for the self-employment income: the estimates of the HET coefficient from the causal forest, random forest, and elastic net are non-significant at the conventional levels (see Table A4 in Appendix). Consequently, we conclude that most of the beneficiary households experienced a uniform increase in their income from self-employment activities while a group of them focused on assets accumulation and the others on smoothing their consumptions.

V. Conclusion

Our results from the causal forest, as well as generic ML methods, report a positive and significant effect of the graduation model on wealth accumulation and household consumption. These findings are consistent with the existing studies that also report a strong positive effect of the graduation model on asset accumulation, and household expenditure outcomes. We find significant heterogeneity in impact on assets while the results for heterogeneity in impact on consumption are less robust. Our classification analysis that seeks to characterize the households which benefited most from the graduation model shows that there is a trade-off between accumulating assets and increasing household expenditures.

Characteristics that are associated with higher gain in asset accumulation show the opposite direction of association with consumption gain. Households that benefited most in asset accumulation were relatively poorer at the baseline compared to the most affected households on the expenditure outcome. The most affected households for the asset outcome had primary beneficiaries who were older, were more dependent on wage income and had less self-employment income at the baseline.

On the contrary, the most affected households for the expenditure outcomes had younger beneficiaries with higher income from self-employment activities and less income from daily wage activities at the baseline. In terms of community-level characteristics, proximity to roads and markets helps in consumption gain over asset accumulation. The lower level of baseline wealth (combining all productive and non-productive assets and savings) is the only variable that shows a higher impact on both asset and consumption.

Besides demonstrating a trade-off in the impact between asset and consumption, these results identify at least a couple of ways the graduation program can improve long-term effectiveness in the context of Bangladesh. Firstly, keeping overall asset ownership of the household as a stricter targeting criterion can help in improving the impact on both outcomes. Secondly, the coaching component can be customized to mitigate the asset-consumption trade-off, e.g. by targeting more intensive support for younger beneficiaries with more decision-making power and closer to markets, to improve on asset accumulation.

References

Ahmed, Akhter U., Mehnaz Rabbani, Munshi Sulaiman, and Narayan C. Das. 2009. *The Impact of Asset Transfer on Livelihoods of the Ultra Poor in Bangladesh. BRAC Research Monograph Series (29)*. Vol. 7.

Andrews, Colin, Aude de Montesquiou, Inés Arévalo Sánchez, Puja Vasudeva Dutta, Boban Varghese Paul, Sadna Samaranayake, Janet Heisey, Timothy Clay and Sarang Chaudhary. 2021. "The State of Economic Inclusion Report 2021: The Potential to Scale". *World Bank Publications.*

Athey, Susan, and Guido W. Imbens. 2017. "The Econometrics of Randomized Experiments." *Handbook of economic field experiments* 1:73–140.

Athey, Susan and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences of the United States of America* 113(27):7353–60.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47(2):1179–1203.

Athey, Susan and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application." *ArXiv* 1–15.

Balboni, Clare, Oriana Bandiera, Robin Burgess, Maitreesh Ghatak, and Anton Heil. 2020. "Why Do People Stay Poor?" *Working Paper*.

Bandiera, Oriana, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. 2017. "Labor markets and poverty in village economies." *The Quarterly Journal of Economics* 132, no. 2: 811-870.

Bandiera, Oriana, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul, Munshi Sulaiman, , Francisco Buera, Bronwen Burgess, Anne Case, Arun Chandrasekhar, Angus Deaton, Greg Fischer, Guy Michaels, Ted Miguel, Mush Mobarak, Benjamin Olken, Steve Pischke, and Mark Rosenzweig. 2013. "Can basic entrepreneurship transform the economic lives of the poor?". *Working Paper.*

Banerjee, Abhijit, Esther Duflo, Raghabendra Chattopadhyay, and Jeremy Shapiro. 2016. "The Long Term Impacts of a 'Graduation' Program: Evidence from West Bengal." Working Paper *Massachusetts Institute of Technology* (September):1–25.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A Multifaceted Program Causes Lasting Progress for the Very Poor: Evidence from Six Countries." *Science* 348(6236).

Basu, Sumanta, Karl Kumbier, James B. Brown, and Bin Yu. 2018. "Iterative Random Forests to Discover Predictive and Stable High-Order Interactions." *Proceedings of the National Academy of Sciences of the United States of America* 115(8):1943–48.

Blattman, Christopher, Eric P. Green, Julian Jamison, M. Christian Lehmann, and Jeannie Annan. 2016. "The Returns to Microenterprise Support among the Ultrapoor: A Field Experiment in Postwar Uganda." *American Economic Journal: Applied Economics* 8(2):35–64.

Carter, Michael R. and Christopher B. Barrett. 2006. "The Economics of Poverty Traps and Persistent Poverty: An Asset-Based Approach." *Journal of Development Studies* 42(2):178–99.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2020. "Generic Machine Learning Inference On Heterogenous Treatment Effects In Randomized Experiments ". arXiv:1712.04802v5

Chernozhukov, Victor, Iván Fernández-Val, and Ye Luo. 2018. “The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages.” *Econometrica* 86(6):1911–38.

Duflo, Esther. 2020. “Long-Term Effects of the Targeting the Ultra Poor Program.” *NBER Working Paper Series* 17.

Emran, M. Shahe, Virginia Robano, and Stephen C. Smith. 2014. “Assessing the Frontiers of Ultrapoverty Reduction: Evidence from Challenging the Frontiers of Poverty Reduction/Targeting the Ultra-Poor, an Innovative Program in Bangladesh.” *Economic Development and Cultural Change* 62(2):339–80.

Foster, Jared C., Jeremy M. G. Taylor, and Stephen J. Ruberg. 2011. “Subgroup Identification from Randomized Clinical Trial Data.” *Statistics in Medicine* 30(24):2867–80.

Gobin, Vilas J., Paulo Santos, and Russell Toth. 2016. "Poverty graduation with cash transfers: a randomized evaluation." *Department of Economics Discussion Paper* 23: 16.

Green, Donald P. and Holger L. Kern. 2012. “Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees.” *Public Opinion Quarterly* 76(3):491–511.

Ikegami, Munenobu, Michael R. Carter, Christopher B. Barrett, and Sarah Janzen. 2016. “Poverty Traps and the Social Protecion Paradox.” *The Economics of Poverty Traps* (December):223–56.

Imai, Kosuke and Marc Ratkovic. 2013. “Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation.” *Annals of Applied Statistics* 7(1):443–70.

Lee, Myoung Jae. 2009. “Non-Parametric Tests for Distributional Treatment Effect for Randomly Censored Responses.” *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 71(1):243–64.

Lin, Yi and Yongho Jeon. 2006. “Random Forests and Adaptive Nearest Neighbors.” *Journal of the American Statistical Association* 101(474):578–90.

Mallick, Debdulal. 2013. “How Effective Is a Big Push to the Small? Evidence from a Quasi-Experiment.” *World Development* 41(1):168–82.

Matin, Imran and David Hulme. 2003. “Programs for the Poorest: Learning from the IGVGD Program in Bangladesh.” *World Development* 31(3):647–65.

Matin, Imran, Munshi Sulaiman, and Evaluation Division. 2008. *Working Paper Crafting a Graduation Pathway for the Ultra Poor : Lessons and Evidence from a BRAC Programme*.

Morel, Ricardo, and Reajul Chowdhury. 2015. "Reaching the Ultra-Poor: Adapting Targeting Strategy in the Context of South Sudan." *Journal of International Development* 27, no. 7: 987-1011.

Quisumbing, Agnes R. and Bob Baulch. 2013. “Assets and Poverty Traps in Rural Bangladesh.” *Journal of Development Studies* 49(7):898–916.

Raza, Wameq A., Narayan C. Das, and Farzana A. Misha. 2012. “Can Ultra-Poverty Be Sustainably Improved? Evidence from BRAC in Bangladesh.” *Journal of Development Effectiveness* 4(2):257–76.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2008. "Nonparametric tests for treatment effect heterogeneity." *The Review of Economics and Statistics* 90, no. 3:389-405.

Schiltz, Fritz, Chiara Masci, Tommaso Agasisti, and Daniel Horn. 2018. “Using Regression Tree Ensembles to Model Interaction Effects: A Graphical Approach.” *Applied Economics* 50(58):6341–54.

Sedlmayr, Richard, Anuj Shah, and Munshi Sulaiman. 2020. “Cash-plus: Poverty Impacts of Alternative Transfer-Based Approaches.” *Journal of Development Economics* 144(November 2019):102418.

Sulaiman, Munshi. 2018. “Graduation Approaches : How Do They Compare in Terms Of.” Pp. 102–20 in *In Boosting growth to end hunger by 2025: The role of social protection*, edited by F. S. W. Taffesse and A. Seyoum. Washington, DC: International Food Policy Research Institute (IFPRI).

Thomas, Marius, Björn Bornkamp, and Heidi Seibold. 2018. “Subgroup Identification in Dose-Finding Trials via Model-Based Recursive Partitioning.” *Statistics in Medicine* 37(10):1608–24.

Wager, Stefan and Susan Athey. 2018. “Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests.” *Journal of the American Statistical Association* 113(523):1228–42.

Willke, Richard J., Zhiyuan Zheng, Prasun Subedi, Rikard Althin, and C. Daniel Mullins. 2012. “From Concepts, Theory, and Evidence of Heterogeneity of Treatment Effects to Methodological Approaches: A Primer.” *BMC Medical Research Methodology* 12.

Appendix A:

Figure A1: Classification analysis for most and least affected groups for savings and self-employment income

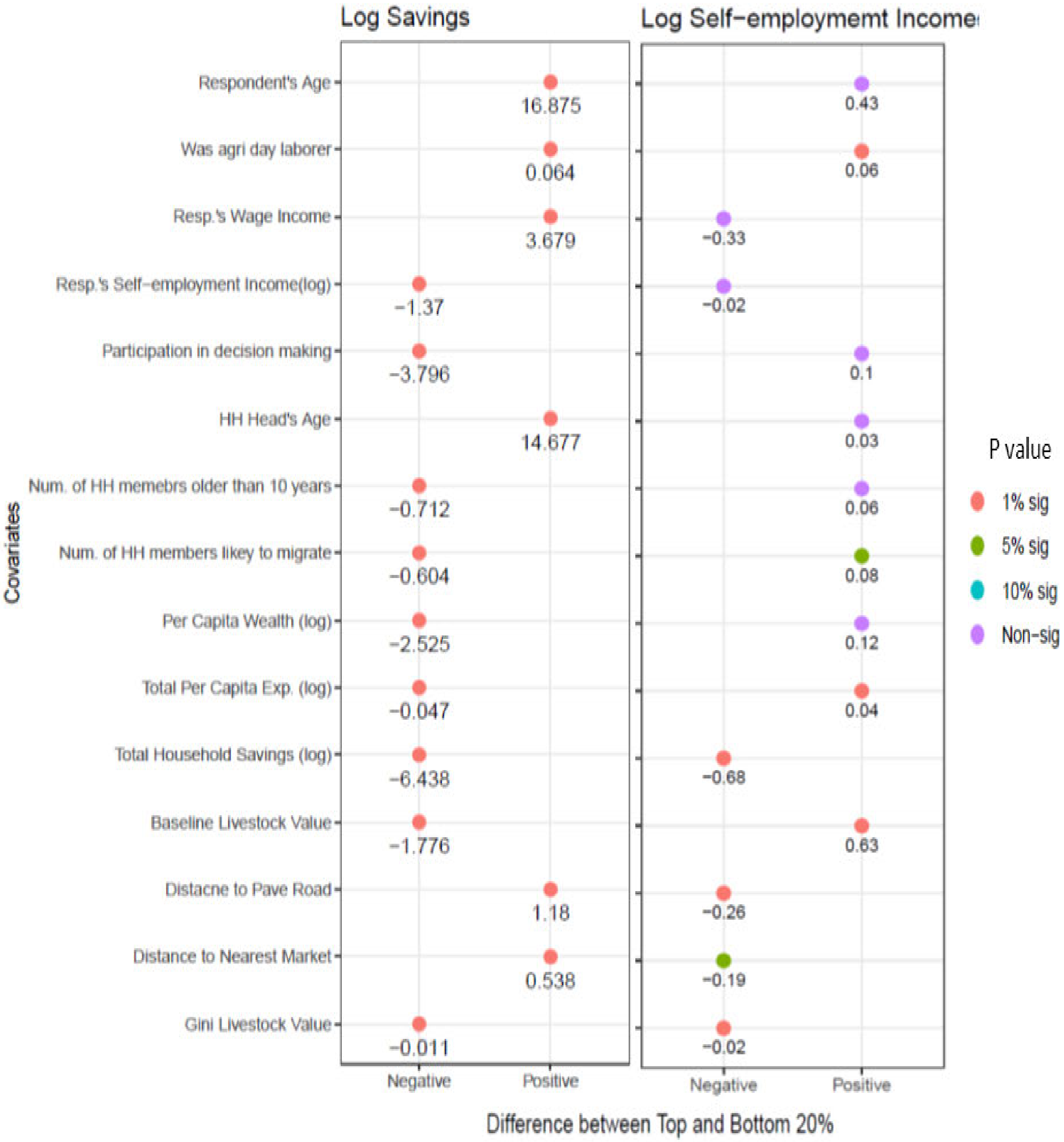


Table A1: Descriptive Statistics of the Baseline Covariates

Baseline Characteristics	N	Mean	SD	P25	P50	P75
Primary Beneficiary Level Characteristics						
Beneficiary's Age	5315	39.61	13.36	28.00	38.00	50.00
Beneficiary never married	5315	0.01	0.11	0.00	0.00	0.00
Beneficiary divorced	5315	0.02	0.12	0.00	0.00	0.00
Beneficiary married	5315	0.61	0.49	0.00	1.00	1.00
Beneficiary widow	5315	0.29	0.46	0.00	0.00	1.00
Beneficiary NGO participation none	5315	0.87	0.34	1.00	1.00	1.00
Beneficiary used to participate in NGO	5315	0.02	0.15	0.00	0.00	0.00
Beneficiary's education years	5315	0.56	1.61	0.00	0.00	0.00
Any future plan for self-employment	5315	0.46	0.50	0.00	0.00	1.00
Empowerment Score (decision making)	5315	5.43	3.20	3.75	6.00	7.75
Empowerment Score (mobility)	5315	5.22	3.70	0.00	8.00	8.00
Any past business activity	5315	0.13	0.34	0.00	0.00	0.00
Log wage income	5315	0.74	8.85	-9.21	7.50	8.85
Had a small business	5315	0.00	0.02	0.00	0.00	0.00
Log income from small business	5315	2.37	104.41	0.00	0.00	0.00
Worked as agricultural day labor	5315	0.27	0.44	0.00	0.00	1.00
Log self-employment income	5315	-2.65	8.00	-9.21	-9.21	6.21
Undernourished	5315	0.56	0.5	0	1	1
Household (HH) Level Characteristic						
HH head is a male	5315	0.61	0.49	0.00	1.00	1.00
Wealth ranked as bottom	5315	0.91	0.29	1.00	1.00	1.00
Wealth ranked medium	5315	0.09	0.29	0.00	0.00	0.00
Number of HH members below 10 years	5315	0.89	1.01	0.00	1.00	2.00
Number of HH members above 10 years	5315	2.48	1.19	2.00	2.00	3.00
HH Head's Age	5315	44.99	13.77	35.00	45.00	55.00
Fraction Muslim	5315	0.83	0.38	1.00	1.00	1.00
Fraction Hindu	5315	0.16	0.37	0.00	0.00	0.00
Any HH member participating in NGO	5315	0.08	0.27	0.00	0.00	0.00
HH members never participated in NGO	5315	0.96	0.21	1.00	1.00	1.00
HH receive any govt. benefits	5315	0.19	0.40	0.00	0.00	0.00
Any HH members migrated out for work	5315	0.24	0.43	0.00	0.00	0.00
HH Head migrated out for work	5315	0.17	0.38	0.00	0.00	0.00
Total HH members migrated out for work	5315	0.27	0.51	0.00	0.00	0.00
HH members likely to migrate for work	5315	2.33	1.08	2.00	2.00	3.00
HH head's education	5315	0.63	1.78	0.00	0.00	0.00
Number of first-degree family members living in same neighborhood	5315	8.69	3.69	6.00	9.00	11.00
Number of first-degree family members living in same village	5315	2.71	2.06	1.00	2.00	4.00
Per-capita annual expenditure (log)	5315	9.29	0.36	9.06	9.27	9.49
Log per-capita food expenditure	5315	9.00	0.38	8.78	8.99	9.22
Log per-capita non-food annual expenditure	5315	7.78	0.50	7.48	7.77	8.07
Log per-capita wealth	5315	5.12	3.73	4.83	5.70	6.61
Had any livestock	5315	0.26	0.44	0.00	0.00	1.00
Cultivable land size (in decimals)	5315	3.11	17.99	0.00	0.00	0.00
Log livestock value	5315	3.04	3.27	0.00	0.00	5.71
Log per-capita education expenditures	5315	-4.81	6.50	-9.21	-9.21	3.76
Log total savings	5315	-6.6	6.44	-11.51	-11.5	1
Community Level Characteristics						
Gini landside	5315	0.79	0.07	0.74	0.79	0.83
Gini livestock	5315	0.74	0.09	0.68	0.75	0.81
Gini total assets	5315	0.76	0.07	0.71	0.76	0.81

Distance to the nearest market	5315	2.63	2.32	1	2	3
Distance to pave road	5315	1.96	2.15	0	1	3

Table A2: GATES by Different Quantiles (Using Causal Forest Estimation)

	Per Capita Wealth (log)	**Per Capita Expenditure (log)**
ATE of 3rd Quantile Group	2.41 [2.27 2.54]	0.15 [0.14 0.15]
ATE of 1st Quantile Group	1.85 [1.85 1.86]	0.12 [0.11 0.12]
Diff. (3rd vs 1st)	0.482 [0.30 0.67] (0.000)	0.034 [0.02 0.04] (0.000)
ATE of 4th Quantile Group	2.77 [2.641 2.89]	0.16 [0.15 0.17]
ATE of 2nd Quantile Group	2.15 [2.02 2.27]	0.14 [0.13 0.14]
Diff. (4th vs 2nd)	0.62 [0.44 0.79] (0.000)	0.02 [0.01 0.03] (0.000)

Table A3: Classification Analysis

Covariates	Stats	Per Capita Wealth [log]			Per Capita Expenditure [log]		
		Top 20%	Bottom 20%	Difference	Top 20%	Bottom 20%	Difference
Primary Beneficiary Level Characteristics							
Respondent's Age	Mean	50.21	31.69	18.52***	37.39	44.49	-7.1***
	95% CI	[49.43 51.00]	[31.15 32.24]	[17.57 19.47]	[36.59 38.20]	[43.71 45.28]	[-8.22 -5.98]
Was agri. day laborer	Mean	0.372	0.158	0.214***	0.101	0.524	-0.423***
	95% CI	[0.342 0.40]	[0.136 0.18]	[0.177 0.25]	0.083 0.119	0.494 0.554	-0.458 -0.388
Resp.'s Wage Income	Mean	4.407	-3.684	8.091***	-4.173	7.166	-11.339***
	95% CI	[3.94 4.87]	[-4.174 -3.193]	[7.416 8.765]	[-4.648 -3.699]	[6.841 7.491]	[-11.914 -10.764]
Resp.'s Self-employment Income(log)	Mean	-3.684	-0.999	-2.685***	0.295	-5.412	5.707***
	95% CI	[-4.17 -3.20]	[-1.489 -0.51]	[-3.371 -1.999]	[-0.202 0.792]	[-5.817 -5.008]	[5.066 6.347]
Participation in decision making	Mean	1.909	7.265	-5.355***	6.163	3.723	2.44***
	95% CI	[1.73 2.09]	[7.155 7.375]	[-5.564 -5.147]	[6.001 6.325]	[3.521 3.925]	[2.181 2.698]
Household Level Characteristics							
HH Head's Age	Mean	52.304	38.576	13.728	43.941	47.582	-3.642
	95% CI	[51.53 53.08]	37.948 39.203	12.733 14.723	43.092 44.79	46.792 48.372	-4.801 -2.483
Num. of HH members older than 10 years	Mean	1.548	2.937	-1.389***	2.653	2.011	0.642***
	95% CI	[1.50 1.60]	[2.873 3.001]	[-1.472 -1.305]	[2.587 2.719]	[1.943 2.079]	[0.547 0.736]
Num. of HH members likely to migrate	Mean	1.486	2.763	-1.277***	2.486	1.923	0.563***
	95% CI	[1.44 1.53]	[2.705 2.821]	[-1.352 -1.201]	[2.426 2.546]	[1.859 1.987]	[0.476 0.651]
Per Capita Wealth (log)	Mean	1.262	7.381	-6.119***	4.739	5.169	-0.43**
	95% CI	[0.88 1.64]	[7.3 7.462]	[-6.51 -5.728]	[4.482 4.996]	[4.964 5.373]	[-0.758 -0.102]
Total Per Cápita Exp. (log)	Mean	9.349	9.331	0.018	9.124	9.427	-0.303***
	95% CI	[9.32 9.37]	[9.311 9.35]	[-0.013 0.049]	[9.104 9.143]	[9.406 9.449]	[-0.332 -0.274]
Total Household Savings (log)	Mean	-8.137	-4.929	-3.208***	-6.314	-6.64	0.326
	95% CI	[-8.47 -7.81]	[-5.349 -4.51]	[-3.742 -2.673]	[-6.71 -5.918]	[-7.022 -6.257]	[-0.224 0.876]
Baseline Livestock Value	Mean	0.78	5.508	-4.728***	3.544	2.172	1.372***
	95% CI	[0.67 0.89]	[5.311 5.705]	[-4.953 -4.502]	[3.344 3.745]	[1.995 2.35]	[1.104 1.64]
Community/Spot Level Characteristics							
Distance to Pave Road	Mean	2.034	1.706	0.327***	1.739	2.225	-0.486***
	95% CI	[1.90 2.17]	[1.594 1.818]	[0.155 0.5]	[1.617 1.861]	[2.094 2.356]	[-0.665 -0.307]
Distance to Nearest Market	Mean	2.644	2.532	0.113	2.174	3.131	-0.957***
	95% CI	[2.52 2.77]	[2.387 2.676]	[-0.077 0.302]	[2.033 2.315]	[3.001 3.261]	[-1.149 -0.765]
Gini Livestock Value	Mean	0.729	0.739	-0.011***	0.781	0.696	0.085***
	95% CI	[0.72 0.73]	[0.734 0.745]	[-0.018 -0.003]	[0.776 0.786]	[0.691 0.702]	[0.077 0.092]

Table A4: Average Treatment Effects (ATE) and Heterogeneous Treatment Effect (HET) for Savings and Self-employment Income

Method		Savings (log)	Self-employment Income (log)
Causal Forest	ATE	9.57	5.66
		[8.18 10.96]	[4.17 7.15]
		(0.000)	(0.000)
	HET	0.99	0.99
		[-2.00 4.00]	[-3.00 5.00]
		(0.241)	(0.33)
Random Forest	ATE	9.03	5.23
		[8.75 9.32]	[4.77 5.69]
		(0.000)	(0.000)
	HET	0.74	0.22
		[0.55 0.93]	[-0.03 0.46]
		(0.000)	(0.179)
Elastic Net	ATE	9.14	5.00
		[8.85 9.43]	[4.54 5.47]
		(0.000)	(0.000)
	HET	1.09	0.43
		[0.64 1.63]	[-0.28 1.02]
		(0.000)	(0.212)

Note: 95% confidence intervals in brackets, and p-value in parenthesis. In estimating the ATE for the causal forest, we used the built-in function in the GRF package. For the generic ML methods, we applied the BLP test that estimates both ATE and HET using equation 4.

Appendix B: Methods – Causal Forest and the Agnostic Approach

There is a small but growing literature on using Machine Learning methods for exploring treatment effect heterogeneity. The literature includes parametric and non-parametric approaches. One strand of this literature constructs parametric estimators of causal inference and treatment effect heterogeneity using regularized regression methods, also known as shrinkage methods (e.g. LASSO, Ridge, and Elastic Net). Like other linear methods, shrinkage methods fit a model by minimizing the reduced sum of squares (RSS); however, they expand the optimization equation by introducing an additional parameter called the shrinkage penalty. In doing so, the estimated association of each variable with the response is reduced (and so is the variance of the coefficient). In the extreme cases of LASSO and Elastic Net, this shrinkage could reduce some coefficients to zero, thus allowing for a data-driven selection of parameters. While these methods allow for high-dimensional setting, they, however, heavily rely on the assumption of sparsity – a setting where only a handful of the covariates explain a large proportion of the variation in the data. Valid inference from these regularized regression-based methods can only be derived if the sample size is sufficiently large relative to the number of selected covariates and their interaction terms (Athey and Imbens 2017; Chernozhukov et al. 2018).

Another stream of the literature builds on nonparametric approaches such as kernels and nearest-neighbor matching methods to explore treatment effect heterogeneity (Lee 2009; Richard K. Crump et al. 2008; Willke et al. 2012). Methods under this approach use Euclidean distance measure to match treated observations with control observations on the covariate space and construct an estimate of the treatment effect by comparing the mean outcomes of the matched observations. These methods perform well with a small number of covariates, and their performance falls quickly as the number of covariates increases (Athey and Imbens 2017; Wager and Athey 2018). Another problem with these methods is that they do not provide information on the relative importance of the covariates in matching observations (Wager and Athey 2018). Tree-based machine learning methods have been widely used to address these limitations of classical non-parametric estimations. Unlike classical matching methods, random forests use a data-driven way in determining which nearby observations should receive more weight (Lin and Jeon 2006; Thomas, Bornkamp, and Seibold 2018; Wager and Athey 2018). The data-driven approach of random forest algorithms allows for flexible modeling of interactions in high dimensional settings, which partitions the data space into a number of subgroups and obtain an accurate prediction of the outcome for those subgroups (Foster et al. 2011; Lin and Jeon 2006; Schiltz et al. 2018). The random forest algorithm has gained popularity in exploring heterogeneity in treatment effects because of its two distinctive advantages; it includes interactions of covariates by construction and thus relieves researchers from prespecifying the form of the heterogeneity, and it accommodates the nonlinear interaction effects and discontinuous relationships (Thomas et al. 2018). However, the random forest method also has its own limitations. Like other nonparametric ML methods, estimates or predictions from random forests cannot be used for constructing confidence intervals that would allow for hypothesis testing. Also, estimates derived from this method eventually become bias-dominated when the number of covariates increases.

Causal forest: Wager and Athey (2018) proposed a modification of the standard random forest algorithm that directly estimates heterogeneity on causal effects and allows for a tractable asymptotic theory and valid statistical inference. This method, referred as casual forest, builds on the causal tree algorithm proposed by Susan Athey and Imbens (2016), which partitions the data into a set of subgroups such that treatment effect heterogeneity across subgroups is maximized (Athey and Imbens 2016). In estimating the treatment effect, the causal tree algorithm follows an 'honest' approach, whereby one sample is used to construct the partition (i.e. building the tree) and another to estimate treatment effects for the subgroups. This honest approach in tree building and estimation of treatment effect claims to eliminate bias and enable centered confidence intervals.

Considering its asymptotic properties and direct estimation of treatment effect heterogeneity, we apply the causal forest method in our paper. In this section, we briefly explain the causal forest algorithm for a randomized trial study setup. The causal forest algorithm starts by drawing a random subsample of training data. Then, it splits the training data into two halves I and J. The algorithm then grows a tree by using the J-sample data to partition the data space, while holding out the I-sample data for within-leaf estimation. When choosing a split, the algorithm seeks to maximize the difference in treatment effect between the two child leaves. This goal of having maximum heterogeneity in treatment effects across the resulting leaves is achieved by using the following so-called honest target function that seeks to minimize mean squared error in estimated treatment effect within each leaf:

$$\mathrm{MS}E_{\tau}(I,J,L) = \frac{1}{\#I} + \sum\nolimits_{i \in J} \{((\tau_i - \hat{\tau}(\mathrm{X}_i; J, L))^2 - \tau_i{}^2)\} \qquad (1)$$

In equation 1, $\hat{\tau}$ is the conditional average treatment (CATE) estimated using J-sample data by simply taking the difference between the outcomes of the treated and control observations within a leaf:

$$\hat{\tau}(\mathrm{X}) = \frac{1}{|\{i: \mathrm{W}_i = 1, \mathrm{X}_i \in L\}|} \sum_{i:\mathrm{W}_i=1, \mathrm{X}_i \in L} \mathrm{Y}_i \quad - \quad \frac{1}{|\{i: \mathrm{W}_i = 0, \mathrm{X}_i \in L\}|} \sum_{i:\mathrm{W}_i=0, \mathrm{X}_i \in L} \mathrm{Y}_i \qquad (2)$$

In equation 2, W is the treatment indicator taking value 1 for treated observation, X is the covariate space, Y is the outcome variable, and L is the leaf within a tree. Following the honest approach, the τ_i in equation 1 is estimated using the I-sample data. However, τ_i, treatment effect for each observation in the I-sample, cannot be estimated as we do not observe the two potential outcomes for the same observation unit (the fundamental challenge of causal inference). Assuming that the potential outcomes are independent of treatment assignment (unconfoundedness) and the fact that τ is constant within each leaf, Susan Athey and Imbens (2016) argued that the leaf average treatment effect can be used as a proxy for τ_i, that is:

$$E[\tau_i | i \in I; i \in L] = E[\hat{\tau}(X; I, L)] \qquad (3)$$

Following this assumption, equation 1 can be simplified as the following:

$$\mathrm{MS}E_{\tau}(I,J,L) = \frac{1}{\#I} + \sum_{i \in I} \{(\hat{\tau}(\mathrm{X}_i; I, L) - \hat{\tau}(\mathrm{X}_i; J, L))^2 - \tau_i{}^2)\} \qquad (4)$$

Computationally, the $\hat{\tau}(\mathrm{X}_i; I, L)$ in equation 4 is estimated using the following algorithm. For each partition, each observation in I-sample is ‘pushed down’ to determine what leaf it falls in. Given this information, a list of weighted neighboring observations is generated by counting how many times the observations fell in the same leaf. Then, the treatment effect is calculated using the outcomes and treatment status of the neighboring observations (for details see Susan Athey and Wager 2019).

The same process is repeated for each leaf to grow a tree following some split balancing parameters such as whether further splitting will result in an improved fit, or each resulting leaf contains at least a certain number of treated and untreated observations. Finally, following the above-explained procedure for generating a single causal tree, the causal forest algorithm generates an ensemble of B such trees, each having an estimate of $\tau_b(x)$. The forest then aggregates their predictions by averaging them: $\hat{\tau}(X) = \frac{1}{B} \sum_{b \in B} \tau_b(X)$.

Though causal forest resolves the fundamental impossibilities of non-parametric inference of causal effects using regression tree-based algorithm and offers asymptotic properties, it also has its own limitations. The method is criticized for exhausting half of the training data at each step of the estimation procedure. Athey and Imbens (2016) addressed this criticism by arguing that randomizing the $I//J$ data splits over each training subsample will make sure that each data point will participate in both I and J samples of some trees, though no data point can be used for both split selection and leaf estimation in a single tree. Therefore, no data will be wasted in achieving honesty in estimation. Yet, this honest estimation requires a large sample size relative to the number of covariates in order to choose high-quality splits and tree building (Athey and Imbens 2017; Chernozhukov et al. 2020). Another limitation of this approach is it is limited to the tree-based methods and does not account for splitting uncertainty (Schiltz et al. 2018).

A more generic approach:

While the causal forests approach uses only one specific ML tool i.e. tree-based algorithm and relies on an honest approach to produce consistent estimates of CATE and explore heterogeneity, Chernozhukov et al. (2020) proposed a different approach that allows applying generic ML methods to estimate causal effects and draw a statistical inference. The empirical strategy of this approach starts by building a proxy predictor of CATE using generic ML methods, and then develop valid inference on some key features of the CATE based on this proxy predictor. Instead of obtaining consistent estimation and uniformly valid inference on the CATE itself, this approach focuses on providing valid estimation and inference on certain features of CATE. Referring to it as an agnostic approach, Chernozhukov et al. (2020) argued that by focusing on key features of CATE rather than CATE itself, this approach avoids making strong assumptions about the properties of ML estimators and yet obtain uniformly valid inference on some features of the estimators. These features of CATE directly focus on detecting heterogeneity in treatment effects, measuring the effects for different bins (e.g. least and most affected units), and looking at the average characteristics of these bins.

Since this approach uses generic ML methods in estimating CATE without imposing any restrictive assumptions on them, and yet allows for constructing valid inference about some key features of the estimators that allow for exploring heterogeneity in treatment effects, we also apply this approach in our paper. The empirical strategy of this approach involves repeatedly splitting the data into two random subsamples, training ML proxy predictors of CATE on one subsample, and developing valid inferences on some features of CATE based on the proxy predictors on the other subsample. The algorithm of this approach is summarized below.

Assume that we have data with a sample size of N, and let the outcome variable, treatment variable, and the other covariates be denoted by Y, W, and X respectively. The data is randomly split into the main sample, denoted by $\mathrm{Data}_M = (\mathrm{Y}_i, \mathrm{W}_i, \mathrm{X}_i)_{i \in M}$, and an auxiliary sample denoted by $\mathrm{Data}_A = (\mathrm{Y}_i, \mathrm{W}_i, \mathrm{X}_i)_{i \in A}$, 100 times. For each split, this approach then trains ML methods and predict the outcome variable on the treated and untreated observations (i.e. E[Y|X, Z=0,1]) in Data_A separately. Applying the trained ML models on Data_A, the algorithm then estimates two potential outcomes [Y(0), Y(1)], and obtain treatment effect estimates, S(X), and baseline effect estimates, B(X), for each observation in Data_M. The baseline effect estimate and the treatment effect estimate are estimated using the following equation:

$$B(X) = E[Y|W=0, X] \qquad (5)$$

$$S(X) = E[Y|W=1, X] - E[Y|W=0, X] \qquad (6)$$

While noting that B(X) and S(X) can be biased and inconsistent, Chernozhukov et al. (2020) argue that their approach does not require them to be consistent, and simply uses them as proxies which are post-processed to make an estimate and make inference on the key features of CATE. Particularly, this approach targets to develop valid inference on three features namely – Best Linear Predictor (BLP), Sorted Group Average Treatment Effects (GATES), and Classification Analysis (CLAN). The BLP, used for testing heterogeneity in treatment effects, is estimated by the following weighted OLS:

$$Y_i = \alpha X_i + \alpha_1 B(X) + \beta_1 (W_i - p(X_i)) + \beta_2 (W_i - p(X_i)) (S_i - ES) + \varepsilon \qquad (7)$$

Where the weights are $\{P(X)(1-(p(X))\}^{-1}$, ES = 1/M $\Sigma S(X_i)$, and P(X) = 1/N Σ W_i=1 for randomized trail study.

β_1 in equation 7 (BLP) indicates the average treatment effect. β_2, the main coefficient of our interest, indicates the degree to which the estimated treatment effect, $S_i(X)$, serves as a proxy for the true treatment effect or CATE. Rejecting the null hypothesis $\beta_2 = 0$ means that there is heterogeneity and $S_i(X)$ is a relevant predictor. Simulation of this process over 100 samples, will result in 100 sets of parameter estimates for β_1, and β_2. To develop valid inference for these parameters, Chernozhukov et al. (2020) proposed a method called 'variational estimation and inference' (VEIN), which keeps the estimates β_1, β_2, and upper and lower bounds of 95% confidence intervals for each of the 100 data split, and take the median values of the parameters as the final estimates of β_1, β_2, and corresponding confidence intervals. More specifically, the lower median of the 100 upper bounds is considered as the final upper bound, the upper median of the 100 lower bounds is taken as the final lower bound. The final estimates of β_1 and β_2 are mid-points of lower and upper medians of β_1, and β_2, (Chernozhukov et al. 2020).