Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Miaomiao Zhang April 14, 2020

Algorithmic Targeting of Social Welfare Programs: Machine Learning for Prediction Model Design and Causal Effects Estimation

by

Miaomiao Zhang

Stephen O'Connell
Adviser

Economics

Stephen O'Connell

Adviser

Seunghwa Rho

Committee Member

Shomu Banerjee

Committee Member

2020

Algorithmic Targeting of Social Welfare Programs: Machine Learning for Prediction Model Design and Causal Effects Estimation

By

Miaomiao Zhang

Stephen O'Connell

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Economics

2020

Abstract

Algorithmic Targeting of Social Welfare Programs: Machine Learning for Prediction Model Design and Causal Effects Estimation

By Miaomiao Zhang

Governments and aid organizations in developing countries implement algorithmic rules to identify and provide necessary aid for households in underprivileged conditions. Given demographic and background characteristics from administrative data, traditional econometric methods along with regularized linear regressions have been used for targeting social welfare programs. Non-parametric machine learning techniques, however, are less common in these contexts. In this paper, I compare non-parametric forests to parametric linear regression techniques in both prediction and causal treatment effects estimation problem settings. The standard metric of prediction accuracy suggests that random forests perform slightly better than regularized linear regressions, validated across multiple subsets of data; the estimated average treatment effects using both modeling techniques are positive, with only causal forests showing statistically significant results. There is no evidence of significant heterogeneity in individual treatment effects.

Algorithmic Targeting of Social Welfare Programs: Machine Learning for Prediction Model Design and Causal Effects Estimation

By

Miaomiao Zhang

Stephen O'Connell

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Economics

2020

Acknowledgments

Attending Emory and having the opportunity to study abroad at the London School of Economics was one of the greatest experiences of my life. The economics department at both universities, filled with a rich diversity of thoughts and practices, opened up this new possibility for me to explore further into the research field. I have been incredibly fortunate to befriend so many wonderful people that I am not able to mention in this short acknowledgment. Thank you all for sharing your wisdom, excitement, and commitment to this field. I feel truly honored to finish my undergraduate degree with all of your influence along the way.

First, I would like to give my special thanks to Prof. Stephen O'Connell, my advisor, who offered me the dataset for this study, instructed me on my thesis objectives and structure, and provided careful guidance on my content development. His continuous efforts in editing my drafts until the last minute are genuinely appreciated. His unwavering encouragement and support for preparing the next stage of my career will always be remembered. His sincere concern to me goes beyond what I can ever ask for both as a thesis advisor and as a career mentor. His inspiring conversations with me made me think through my potential career choices and start to consider becoming a researcher in academia and a professor at university.

I also want to thank Prof. Seunghwa Rho for her countless hours teaching me causal inference and machine learning, helping me overcome my self-doubt when I encountered obstacles, and pointing me in a new direction when the development of my thesis failed to align with the original objectives. Her knowledge inputs and patient explanations greatly contributed to the technical piece of my work. Her faith in me doing a Ph.D. in the related field motivated me to apply for pre-doctoral fellowships after getting rejected by my dream Master's programs. I am honored to call her a friend who shared with me her Ph.D. journey, her industry experiences as well as her thoughts behind the decision to go back to academia. I see her as my role model for life as a strong and independent woman.

I am also incredibly grateful to Prof. Shomu Banerjee, the professor whom I have known for the longest. He has always been proactively maintaining our relationship by regularly squeezing time out of his busy schedule for us to chat. He always reminds me to be aware of my surroundings, and other things, good or bad, mundane or thought-provoking, happening in this world. He urges me to look at myself outside of my own bubble and to be a better person who thinks, cares and loves. I will always keep his wise words in mind and remember them by heart.

Last but least, to my friends and family – thank you. I could not have studied abroad and completed my studies in the U.S. without the unconditional support from my family. I could not have lived my most fulfilled college career without any of them. Having been able to focus on

my studies, build meaningful friendships, and participate in a variety of extracurricular activities is both a privilege and a blessing to my ordinary life. Thanks for keeping me grounded and optimistic. To quote C.S. Lewis for that, “there are far, far better things ahead than any we leave behind.”

Most importantly, I thank God, for leading our journeys ahead.

Table of Contents

I. Introduction

The concept of nonparametric analysis, estimation, and inference has a long and storied existence in the annals of economic measurement. From distribution-free methods and order statistics to kernel estimators for regression, nonparametric methods with less restrictive density and distributional assumptions have provided ways to complement traditional parametric econometric tools. An emerging literature looks at the intersection of machine learning (ML) and traditional econometric tools to develop more systematic and data-driven approaches to facilitate algorithmic policy design and evaluation. Mckenzie (2018) writes that ML can be used for development interventions and impact assessment in measuring outcomes and targeting treatments, measuring heterogeneity, and addressing confounders. Andini et al. (2015) present two examples regarding policy targeting and illustrate the benefits of using ML techniques when compared to the standard practice of employing coarse policy assignment rules based on a few arbitrarily chosen characteristics.

Athey (2018) and Mullainathan & Spiess (2017) provide an overview of the impact of machine learning on economics and summarize different roles that standard econometrics and ML play in causal inference and prediction problems. In particular, when the goal is semi-parametric estimation or when there are a large number of factors that need to be controlled for, adopting ML techniques is advantageous. Furthermore, ML techniques' ability to choose flexible functional forms is well suited for tasks like prediction. In fact, ML methods are designed to maximize out-of-sample accuracy by uncovering complex model structures that were not specified in advance. The tradition in economic literature has been to identify

causality using one specific model and focus on the parameter estimation for determining treatment effects, whereas ML aims to fit complex and very flexible functional forms to the data without simply overfitting using a wide selection of ML algorithms.

Despite the clear distinction between estimation and prediction problems in economics, in causal inference, there is also a growing interest in using ML algorithms to estimate the average treatment effect (ATE). For example, targeted maximum likelihood estimators of the ATE (van der Laan & Rose, 2011) often utilize an ensemble learner called SuperLearner (van der Laan et al., 2007), which combines different ML algorithms such as the Lasso, K-nearest matching, generalized additive models, generalized linear models, random forests, and multivariate adaptive regression splines, to flexibly estimate the propensity score model and the outcome model in order to estimate the ATE. Also, Künzel, et al. (2019) proposed an ensemble learning method to estimate the conditional average treatment effect (CATE), the ATE among individuals with a given covariate vector. Causal Forests are another popular ML method for estimating the CATE and can account for heterogeneity within each clustered group (Athey & Wager, 2019). Although prediction is often a large part of a resource allocation problem, determining which units benefit the most from treatment is an inferential question, and answering it requires different types of data and assumptions.

With the goal of further improving model prediction accuracy in the targeting of large-scale social welfare programs, this paper intends to compare parametric and nonparametric approaches in the following two ways. First, I investigate whether non-parametric ML techniques trained on subsets of data that exclude noise from the two extrema of the modeled outcome capture more precise signals from the sample; second, I integrate ML with the

traditional econometric methodology to estimate causal heterogeneous treatment effects. The standard metric of prediction accuracy suggests that forests-based approach yields better performance and more reliable estimates for average treatment effects. Although there is no evidence suggesting heterogeneous subsidies effects across the entire population group, this study provides a new algorithmic approach for estimating the causal impact of subsidies on households' living conditions. It also poses a possibility of developing a new algorithmic targeting model for social welfare programs based on treatment effects.

The structure of the study is as follows: in Section II, I provide the context for the *status quo* social welfare programs targeting model with a brief summary of approaches undertaken to address the data scarcity and prediction inaccuracy issues. In Section III, I discuss the methodologies commonly used in econometrics and ML and the reasons why the latter should be taken as a preferable approach for solving prediction problems. In Section IV, I describe the survey dataset and the parametric and nonparametric model design for this study. Section V compares each of the models' performance for prediction and treatment effects estimation: nonparametric models show better outcomes in both contexts. Section VI discusses what has been learned and not learned from the results, the limitation of this study, and possible improvements. Section VII concludes the major contributions of this study as well as machine learning's potential to address prediction and causal problems in a broader context.

II. Context

Social welfare programs provide a viable use-case for testing and evaluating algorithmic targeting policies. The particular context that I use falls under the recent refugee crisis caused

by the Syrian Civil War, which has caused one of the largest episodes of forced displacement since World War II and some of the densest refugee-hosting situations in modern history (Krishnan et al., 2019). Syria's immediate neighbors host the bulk of Syrian refugees, where in many cases local non-governmental organizations, multilateral international organizations, government policymakers aim to provide targeted aid to refugees, but face persistent challenges in identifying households in greatest need of financial assistance (Coady et al., 2004). The most prominent reasons for such challenges are: first, databases maintained by humanitarian agencies for internal programming purposes are not collected for the purpose of program targeting *per se*; second, many impoverished refugees work in the informal sector and records of income in these areas are generally poorly kept, if available at all; and third, the displaced have a high degree of mobility and they are often unwilling to speak to surveyors (Krishnan et al., 2019).

Due to these constraints, an econometric Proxy Means Test (PMT) method has become a very popular approach used to determine what subsets of the population are in the greatest need for financial aid. The PMT is, in the most technical sense, a multiple regression formula that is employed practically to produce a prediction ("score") that is an estimate of a given household's level of wealth, based a variety of features or "proxies" (Altındag et al., 2019). According to an International Labor Organization assessment of PMT methodology in Kidd, Gelders, & Bailey-Athias (2017, p.2), these "proxies" are usually based on demographics (such as age, gender, and number of people in the household); human capital (such as level of education of the household head); type of housing (such as the type of roof, walls, floor, and

toilet); durable goods (such as whether a household has a radio, refrigerator or television); and productive assets (such as whether a household owns animals or land).

Quite often, however, these models incorporate high built-in design errors and struggle with overfitting when applied in practice. For example, Alatas et al. (2012) found that the PMT model used to facilitate the *Program Keluarga Harapan* (PKH) conditional cash transfer scheme in Indonesia resulted in 93 percent of the poorest 5 percent of households being excluded. Another study of the *Oportunidades* (formerly *Progresa*) program in Mexico found that a PMT selection process meant to target the poorest 20 percent of the population had inclusion or false-positive errors of 36% and exclusion or false-negative errors of 70% (Veras et al, 2007). As seen in Figure 1, these error levels tend to increase with more specific coverage levels or to the extent that they aim to target smaller, poorer, subsets of the general population (Kidd & Wylde, 2011). These shortcomings have led to pushback against the PMT model by academics and the development community at large, as critics claim that such inaccurate dispersal of funds is essentially arbitrary and thus could lead to negative stigmatization of aid recipients within their communities as well as mistrust of aid organizations in the future (Kidd et al., 2017).

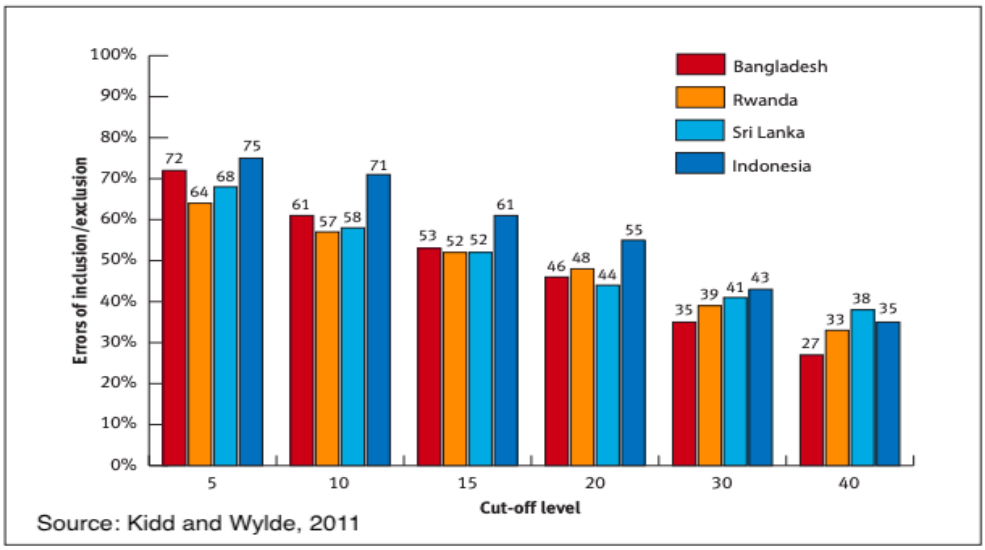


Figure 1: PMT Performance

The informational scarcity that has necessitated the use of the PMT model in the first place is not going to be alleviated soon. Recent research has sought to viable alternative confronting the criticisms of traditional PMT methodology. McBride and Nichols (2018) and Kshirsagar et al. (2017) show that approaches that prioritize out-of-sample accuracy perform substantially better in accurately identifying the poor population compared to a standard PMT approach relying only on in-sample fitting; most relevant to my study, Altindag et al. (2019) conducted a study that combined nationally representative expenditure survey data and routinely collected administrative data in Lebanon to predict refugee households' expenditure using regularized linear regression. The prediction performance of the targeting model has proven to be "at least as accurate using only basic demographic information from administrative records compared to the traditional 'scorecard' PMT that requires a household survey of the entire population". However, most historical targeting literature focuses on improving the prediction model accuracy — whether the subsidies have been allocated properly to households that are poor

now. Emerging literature looks at the effect of targeted benefits (Griffith et al., 2018; Hoynes et al., 2016; Cunha, 2014). This paper poses a new direction for social welfare targeting programs – whether we can target the people who would benefit the most *after* the treatment.

III. Methodology

III.1 Proxy Means Tests, OLS, and Regularized Linear Regressions

The econometric approach to PMTs typically uses consumption or expenditure data from a representative household survey as a proxy for poverty and derives a model, typically using forward stepwise regression, that assigns weights to factors used to predict poverty in the broader eligible population. The predictors in a standard PMT model comprise a set of household assets and demographics; OLS is used in the regression model to choose the measures that predict consumption. These features are usually easily verifiable and thus becomes the key step in targeting the eligible population (Brown et al., 2018). However, OLS is known as the best linear unbiased estimator under Gauss-Markov assumptions but standard empirical techniques like this are not optimized for prediction problems. It struggles with overfitting because the model is built on existing administrative-recorded, in-sample data, and fails to perform well out-of-sample.

While the main goal of targeting is to accurately predict welfare in a population for which the data on the outcome of interest is not available, assessment of in-sample prediction performance does not seem to be meaningful. In fact, the goal of both OLS and ML methods is to find some function that accurately expresses y as a function of x (or, technically speaking, to minimize the sum of squared residuals). The difference is that the assessment of ML models

relies on *new* out-of-sample observations of x, not the observations used to fit the model. More specifically, the within-sample error is of less concern in prediction context, since we care more about the model's performance out-sample. For example, consider the mean squared error at a point x. $MSE(x)$ can be decomposed as:

$$\underbrace{E_D[(\hat{f}(x) - E_D[\hat{y}_0])^2]}_{\text{Variance}} + \underbrace{(E_D[\hat{y}_0] - y)^2}_{\text{Bias}^2}.$$

Because the *f* varies from sample to sample, it produces variance (the first term) and this must be traded off against bias (the second term). By ensuring zero bias, OLS allows no trade-off. Such overfitting the prediction sample yields poor out-of-sample performance, and prediction tools that are designed to minimize out-of-sample error can potentially increase targeting accuracy.

Beyond traditional linear or logistic regression, penalized regression methods such as LASSO linear regressions have proven to be useful (Altındag et al., 2019). First, since simpler models tend to work better for out-of-sample forecasts, there are various ways to penalize models for excessive complexity called "regularization". Second, by dividing the data into training, testing, and validation sets for the purpose of estimating, choosing, and evaluating a model, we can have more accurate and reliable model outputs. Third, if we have an explicit numeric measure of model complexity, we can view it as a parameter that can be "tuned" empirically to produce the best out of sample predictions. Usually, we pick $k = 5, 10,$or $n - 1$to perform "k-fold cross-validation" for choosing the most appropriate tuning parameter upon examining some associated loss function. Even if there is no tuning parameter, we use cross-

validation to report goodness-of-fit measures since it measures out-of-sample performance, which is generally more meaningful than in-sample performance (Varian, 2014).

In particular, picking the best LASSO regularized linear regression function involves two steps. The first step is, conditional on a class of linear functions (over some fixed set of possible variables), to pick the best in-sample quadratic loss function as in OLS. The second step is to choose the optimal regularizer (which is the sum of absolute values of coefficients) using cross-validation. This effectively results in a linear regression in which only a small number of predictors from all possible variables are chosen to have nonzero values: the absolute-value regularizer encourages a coefficient vector where many are exactly zero (Mullainathan & Spiess, 2017).

III.2 Non-parametric Model – Trees Algorithm

Decision trees utilize the two insights of regularization and empirical choice of the regularization penalty as a non-parametric approach. Techniques built on decision trees allow for sparser datasets to predict an outcome of interest and more flexible functional forms to include higher-order interaction terms. Within the prediction-based machine learning literature, regression trees differ from most other methods in that they produce a partition of the population according to covariates, whereby all units in a partition receive the same prediction (Hastie et al., 2009).

Most economists are familiar with decision trees that describe a sequence of decisions that results in some outcome. A tree classifier has the same general form, but the decision at the end of the process is a choice about how to classify the observation. The goal is to construct

(or "grow") a decision tree that leads to good out-of-sample predictions. Trees tend to work well for problems where there are important nonlinearities and interactions. The most common solution to this problem is to "prune" the tree by imposing a cost for complexity, such as the number of terminal nodes. The cost of complexity is, therefore, the tuning parameter chosen to provide the best out-of-sample predictions, which is typically measured using the k-fold cross-validation procedure mentioned earlier (Varian, 2014).

Bagging, random forests and boosting all work similarly by growing many trees on the training data and then combining the predictions of the resulting ensemble of trees. One crucial advantage of nonparametric tree algorithms is that extraneous predictors do not affect too much of their performance and there are no assumptions that the response has a linear (or even smooth) relationship with the predictors. However, although random forests and boosting are among the "state-of-the-art" methods for observing the actual outcomes in a cross-validation sample, the fundamental problem of causal inference is not addressed: no valid confidence intervals can be directly computed and we cannot make inference based on variable selection about the data-generating process without further assumptions such as a true sparse model and the absence of the relationship between irrelevant and relevant covariates (Mullainathan & Spiess, 2017).

IV. Data and Model Design

IV.1 Survey Data

In the context of forced displacement, the selection of a representative sample of hosts and the displaced poses a major challenge to drawing credible inferences about Syrian refugees and

the host communities' socio-economic outcomes. I used the results from the *Syrian Refugee and Host Community Surveys* (SRHCS), which were implemented over 2015-2016 in Lebanon, Jordan and the Kurdistan region of Iraq by the World Bank Group. For the prediction model design and subsidies impact estimation, only the Syrian refugees' data in the Lebanon host community are used to avoid potential host-community-specific characteristics. However, given the lack of an updated sample frame and cartographic division of the country into small geographic areas, and with *Circonscription Foncières (CF)* being the finest level of disaggregation available, the surveyors depended on UNHCR data on registered Syrian refugees and combined the estimates of Lebanese population at the CF level (Krishnan et al., 2019). The survey dataset accrued as a result nonetheless provides comprehensive households' information and reveals comparable findings on the lives and livelihoods of Syrian refugees and host communities.

The questionnaire includes detailed questions on demographics, employment, access to public services, health, migration, and perceptions. More specifically, it contains a total of 642 questions, which are broken into 23 sections, including A. Roster, B. Dwellings, C. Services, D. Assets, E. Sources of income, F. Types of assistance, G. Income shocks, H. Prices, I. Food security, J. Health access, K. School access, L. Movements, M.-P. Employment, Q.-T. Retro Employment, U. Norms and Relations, V. Conflicts, and W. Assessment. The Lebanon group contains 12, 523 members in 2, 865 unique households. Since all aids targeting has been done on the household level, the first step I took was to aggregate the individual-level responses into household-level features. For both the income prediction and subsidies effects estimation, the

methodologies and the reasoning for engineering the outcome and other related variables are introduced in Section IV.2.

Table 1 presents descriptive statistics of the representative samples in Lebanon based on the SRHCS survey design and sampling document. Panel A summarizes the statistics of major household characteristics in the survey samples. The average household size is around 4 with the average dependency ratio[1] of 0.68. On average, more than half of the households' income is from wages and only 7% come from humanitarian assistance. Approximately half of the households rent their dwellings. Panel B shows that Syrian refugee households in Lebanon are on average headed by male adults (22 to 47 years). In terms of educational attainment, few (less than 25% of) refugees have completed secondary schooling or more. The situation is better when we restrict our attention to labor market respondents from 20 to 60 years old (Panel C). A large share of refugees does not participate in the labor force and work. Finally, and consistent with the reliance on household income on wages, the large majority of forcibly displaced work for wages if employed.

1 The dependency ratio is a measure showing the ratio of the number of dependents aged zero to 14 and over the age of 65 to the total population aged 15 to 64.

	N	Mean	SD	Min.	Max.
Panel A: Households					
Size	2865	4.37	2.05	1	20
Dependency Ratio	2743[2]	0.68	0.76	0	7
% income from wages	2865	54%	45%	0	1
% income from business earnings	2865	27%	42%	0	1
% income from assistance	2865	7%	19%	0	1
Any new member since 2010 or since household formation	2865	0.14	0.35	0	1
Old members left since 2010 or since household formation	2734[3]	0.24	0.45	0	1
Rents dwelling currently	2865	0.51	0.50	0	1
Panel B: Household head					
Male	2865	0.89	0.31	0	1
Age	2865	44.74	14.24	15	95
Never attended school, illiterate	2865	0.11	0.32	0	1
Secondary schooling or more	2865	0.23	0.42	0	1
Panel C: Labor market respondents (ages 20-60)					
Male	6400	0.48	0.50	0	1
Age	6400	36.08	11.53	20	60
Never attended school, illiterate	6400	0.08	0.28	0	1
Secondary schooling or more	6400	0.33	0.47	0	1
Participated in the labor force	6400	0.48	0.50	0	1
Employed	6400	0.47	0.50	0	1
Wage worker (if employed)	3006	0.71	0.45	0	1

Table 1: SRHCS (Lebanon) - Refugee household and household head's characteristics, per refugee status

Note: The summary statistics are based on office calculation.

IV.2 Variables

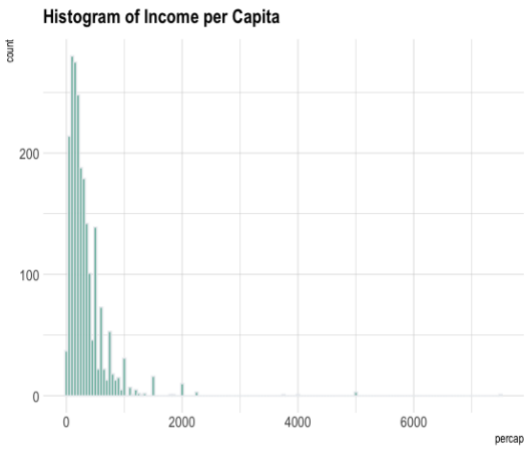
(i) Income

According to the survey mandates, households can choose not to disclose their income information and the surveyors mark "98" on the record. I first excluded those households from the dataset and 2,169 households remained. Dividing the aggregate household income by the

2 Infinity values are excluded from the calculation.

3 Missing values are excluded from the calculation.

household size to generate income per capita as the second step is conventional. As shown in Figure 2a, the distribution of income per capita is highly skewed to the right. After a log transformation of the variable, the distribution looks more normal, as shown in Figure 2b. This variable serves as the outcome of interest for the prediction models.



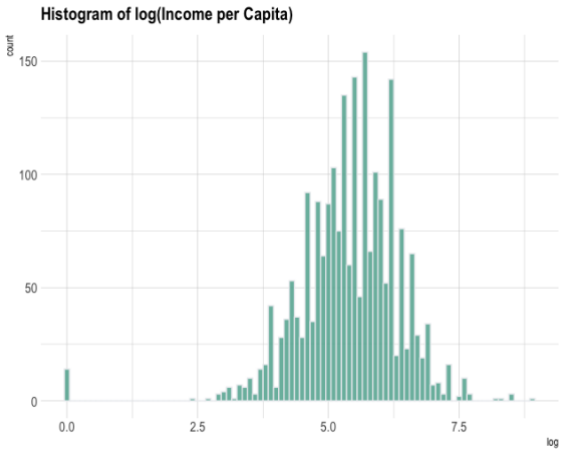


Figure 2a. (binwidth = 50) Figure 2b. (binwidth = 0.1)
Source: SRHCS, Lebanon 2015-2016, Section E. Sources of Income

(ii) Administrative Variables

A large degree of a model's predictive power is attributed to the management of data and the methodology employed to formulate the outcomes (Mullainathan & Spiess, 2017). Since the survey contains mostly multiple choices or closed questions with many not applicable to some respondents, a sparse data set with a considerable number of features being categorical or dummy is given. However, as the main focus of this paper is not about selecting and encoding the most optimal variables, I circumvent such concern by only extracting information which governments and aid organizations are typically able to obtain from administrative

surveys. From all the survey questions asked for households in Lebanon, I divided them into six categories: demographics, education, occupation, years of arrival, protection measures, and assistance received and constructed the candidate variables list in Table 2. This feature set served as the independent variables for income prediction modeling and the confounding variables for treatment effects estimation, except for the category "Assistance Received" being the treatment dummy (constructed as 1 if any type of the four types of assistance was received and 0 otherwise). Table 3 presents the summary statistics of all the variables used in the study.

Demographic Variables	*Type*	*Year of Arrival*	*Type*
Household Size	Positive integer	Not displaced by the Syrian Crisis	Indicator [0/1]
Household Size Squared	Positive integer	2010	Indicator [0/1]
Frac. of Dependent HH memberes	Continuous [0,1]	2011	Indicator [0/1]
Age of the HoH	Continuous [0,100]	2012	Indicator [0/1]
Female HoH	Indicator [0/1]	2013	Indicator [0/1]
Frac. of HH members aged 0-5	Continuous [0,1]	2014	Indicator [0/1]
Frac. of HH members aged 6-10	Continuous [0,1]	2015	Indicator [0/1]
Frac. of HH members aged 11-17	Continuous [0,1]	2016	Indicator [0/1]
Frac. of male members aged 18-50	Continuous [0,1]		
Frac. of female members aged 18-50	Continuous [0,1]	*Protection Measures*	
Frac. of members older than 60	Continuous [0,1]	Frac. of HH members with a disability	Continuous [0,1]
Education		Disabled HoH	Indicator [0/1]
Frac. of HH members education unknown	Continuous [0,1]	Existence of a disabled dependent member	Indicator [0/1]
Frac. of HH members no education	Continuous [0,1]	Single Parent	Indicator [0/1]
Frac. of HH members some education below primary	Continuous [0,1]		
Frac. of HH members with primary education	Continuous [0,1]		
Frac. of HH members with intermediate education	Continuous [0,1]		
Frac. of HH members with secondary education	Continuous [0,1]		
Frac. of HH members above secondary education	Continuous [0,1]		
Occupation			
Frac. of HH members Laborer	Continuous [0,1]	*Assistance Received**	
Frac. of HH members Student	Continuous [0,1]	WFP food voucher	Indicator [0/1]
Frac. of HH members Housekeeper	Continuous [0,1]	Cash assistance	Indicator [0/1]
Frac. of HH members Unemployed	Continuous [0,1]	Food in-kind	Indicator [0/1]
Frac. of HH members Not looking for job	Continuous [0,1]	Other types of assistance	Indicator [0/1]
Frac. of HH members Not eligible for work	Continuous [0,1]		
Frac. of HH members with unknown occupation	Continuous [0,1]		

Table 2: SRHCS (Lebanon) Section A – Administrative variables list

Note: There are three types of variables we use: integer, continuous, and indicator. Integer variables take positive integer values only. Continuous variables can (but do not necessarily) take any value on the closed interval indicator. Indicator variables capture whether the household exhibits the characteristic indicated.

(iii) Treatment Effect of Interest

In the C. Services section of SRHCS, four questions asked respondents to evaluate the overall access to essential house services – water, sewerage, solid waste disposal, and

electricity— as compared to five years ago before they were allocated subsidies. I assigned a score of “-1”, “0”, “1” to their answers of “Better Today”, “Same”, and “Worse Today”, respectively. Moreover, in the H. Prices section, the survey respondents were asked to compare the prices of food, shoes, clothes, and rents today to those of five years ago. In a similar sense, I assigned a score of “-1”, “0”, “1”, to their answers of “Less affordable”, “About the same”, and “More affordable”, respectively. In the end, I aggregated them together and constructed a new variable, denoted “Y” in Table 3, as the proxy for households’ perception about their current living conditions compared to the past. This variable served as the treatment outcome of interest upon which I evaluated the effectiveness of subsidies received on households’ well-being. It is worth noting that the income variable here has been adjusted to be *pre-subsidies* income to avoid the issue of having “bad controls”; covariates that are directly affected by receiving subsidies should be excluded from regressions.

	Mean	Stdev	Min	25%q	Med	75%q	Max
Household Size	4.33	2.08	1	3	4	6	20
Frac. of HH members aged 0-5	0.12	0.18	0	0	0	0.25	0.8
Frac. of HH members aged 6-10	0.09	0.14	0	0	0	0.17	0.75
Frac. of HH members aged 11-17	0.12	0.17	0	0	0	0.25	0.67
Frac. of male members aged 18-50	0.24	0.21	0	0.13	0.2	0.33	1
Frac. of female members aged 18-50	0.25	0.18	0	0.17	0.25	0.33	1
Frac. of members older than 60	0.11	0.26	0	0	0	0	1
Frac. of members no education	0	0.03	0	0	0	0	0.75
Frac. of members some education below primary	0.21	0.29	0	0	0	0.33	1
Frac. of members with primary education	0.21	0.29	0	0	0	0.33	1
Frac. of members with intermediate education	0.26	0.28	0	0	0.2	0.43	1
Frac. of members with secondary education	0.12	0.2	0	0	0	0.2	1
Frac. of members above secondary education	0.11	0.21	0	0	0	0.17	1
Frac. of members Laborer	0.3	0.24	0	0.17	0.25	0.4	1
Frac. of members Student	0.22	0.24	0	0	0.17	0.4	1
Frac. of members Housekeeper	0.22	0.16	0	0.14	0.2	0.33	1
Frac. of members Unemployed	0.03	0.1	0	0	0	0	1
Frac. of members Not looking for job	0.04	0.1	0	0	0	0	1
Frac. of members Not eligible for work	0.18	0.24	0	0	0	0.33	1
Frac. of HH members with a disability	0.02	0.09	0	0	0	0	1
Single Parent	0.15	0.36	0	0	0	0	1
Existence of a disabled dependent member	0.01	0.1	0	0	0	0	1
Household Size Squared	23.1	23.44	1	9	16	36	400
Frac. of Dependent HH members	0.35	0.28	0	0	0.33	0.5	1
Age of the HoH	44.48	14.62	15	33	42	54	95
Female HoH	0.12	0.32	0	0	0	0	1
Not displaced by the Syrian Crisis	0.59	0.49	0	0	1	1	1
Year of Arrival = 2010	0.02	0.13	0	0	0	0	1
Year of Arrival = 2011	0.09	0.28	0	0	0	0	1
Year of Arrival = 2012	0.13	0.34	0	0	0	0	1
Year of Arrival = 2013	0.11	0.31	0	0	0	0	0
Year of Arrival = 2014	0.06	0.23	0	0	0	0	1
Year of Arrival = 2015	0.01	0.12	0	0	0	0	1
Year of Arrival = 2016	0	0.02	0	0	0	0	1
Disabled HoH	0.03	0.18	0	0	0	0	1
log(income per capita)	5.26	1.27	0	4.8	5.46	5.99	8.92
W: treatment of receiving subsidies	0.41	0.49	0	0	0	1	1
Y: score for living conditions improvement	-3.69	2.14	-7	-5	-4	-2	4

Table 3: Summary statistics of all variables used

Note: Summary statistics based on SRHCS data.

IV.3 Model Design

(i) Prediction

For training and validation purposes, the data were split into five folds – four folds were used for training and the remaining one for testing. All the models were built on the training set with cross-validation for parameter tuning; the testing set was used to assess the models' predictive performance. Then I repeated the steps above using a different fold for testing and training on the other four remaining folds. Within each iteration, the testing fold was held out first to prevent overestimating model accuracy. In the end, I obtained the average root mean squared error (rMSE) across the five testing folds and compared the prediction accuracy to that of a PMT approach. Parametric LASSO regularized linear regressions and non-parametric random forests were used for the algorithmic targeting model design. Figure 3 provides an illustration of the five-fold cross-validation procedure.

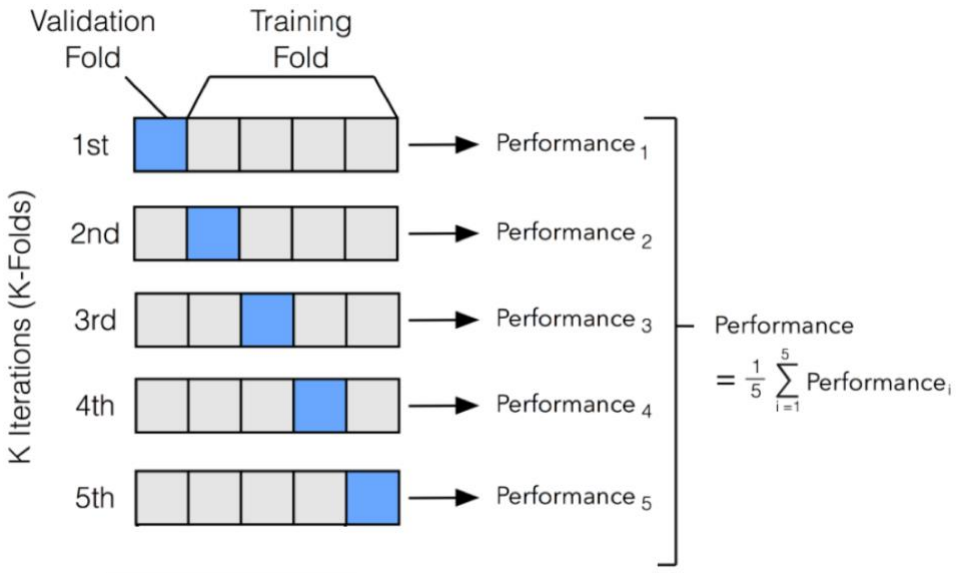


Figure 3: Five-fold cross-validation

Furthermore, an innovative step I took here is that I built models on partial training sets by leaving out the top and bottom percentiles step by step – 0.5%, 1%, and 3%. The hypothesis is that, by training models based on subsets of data in absence of noises from "outliers" on the

two extrema, we would be able to capture more precise signals for income prediction and thus obtain more accurate prediction results. With different partition points for selecting the training sets, I experimented with the two ML algorithms – one parametric and the other non-parametric – to perform feature selection and assessed their prediction accuracy. Specifically, within each training group (100%, 99%, 98%, 97% of the full training set), I obtained the average rMSE of the training and testing sets for LASSO regularized linear regression and random forests across the five iterations. The results are reported in Section V.1.

(ii) Treatment Effects Estimation

As shown in Table 2, the survey data also contain information about whether or not households have received any type of assistance in the past. To estimate the *ex post* benefits received by targeted households – whether or not receiving subsidies has positively impacted refugees' well-being, we need to compare the *status quo* to the counterfactual scenario if they had not been allocated the assistance. The histograms of pre-subsidies income and the assigned scores which indicate households' ratings of their current living conditions compared to the past are shown in Figure 4. In general, households feel less satisfied with their current living conditions than before with the majority of scores less than 0, as suggested in Figure 4b. This means that the majority of survey respondents consider their access to house services to be worse and the daily necessities to be less affordable. Indeed, households that received subsidies (indexed by one and color-coded in orange) are on average under worse economic and living conditions: the average for the untreated group is -3.16 whereas the average for the treated group is -4.46. A Welch Two Sample t-test in Table 4 shows that the difference in means

of Y between the households who received subsidies and those who did not is statistically significant at all conventional levels of significance.

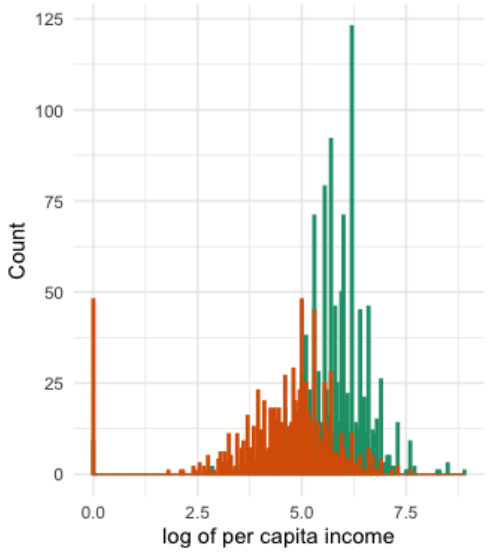


Figure 4a: Histogram of household income

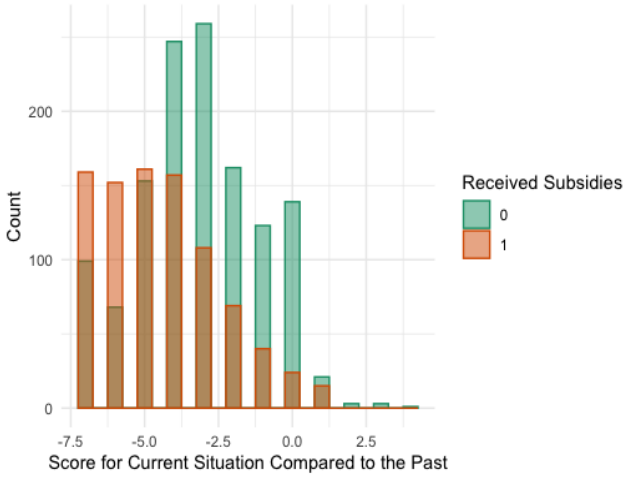


Figure 4b: Histogram of household living conditions

Welch Two-Sample t-test					
H_0: True difference in means is equal to zero	Mean0	Mean1	t	p-value	95% CI
0: No Subsidies; 1: Received Subsidies	-3.159624	-4.457627	14.637	2.2e-16	[1.124091, 1.471915]

Table 4: t-test for mean-difference in Y between treatment and control groups

Because subsidies were not allocated randomly in the first place, the households in the two groups here cannot be valid counterfactuals for us to estimate average treatment effects directly. A key insight here is that the allocation of subsidies for households was based on the administrative variables listed in Table 2, which, in another word, suggests that treatment could be as good as random conditionally on this set of features. Therefore, after systematically controlling for factors not possibly affected by receiving subsidies, we can potentially establish unconfounded-ness where estimating heterogeneous treatment effects using regression

approaches and random forests is possible. In this study, I used the propensity score to reduce the bias in the estimation of treatment effects. Briefly, a propensity score is a study unit's conditional probability of receiving treatment given observed pre-treatment covariates (Rosenbaum & Rubin, 1983). Each unit in the dataset has a propensity score ranging from 0 to 1. Matching typically follows after obtaining the propensity scores: treated and control units with similar propensity scores are matched in pairs. The underlying notion is that if the balancing hypothesis is satisfied, for a given propensity score, exposure to treatment is random and therefore households who received subsidies and who did not should be on average observationally identical. To examine the pre-treatment covariates, a pairwise correlation plot for the dataset used to estimate treatment effects is shown in Figure 5.

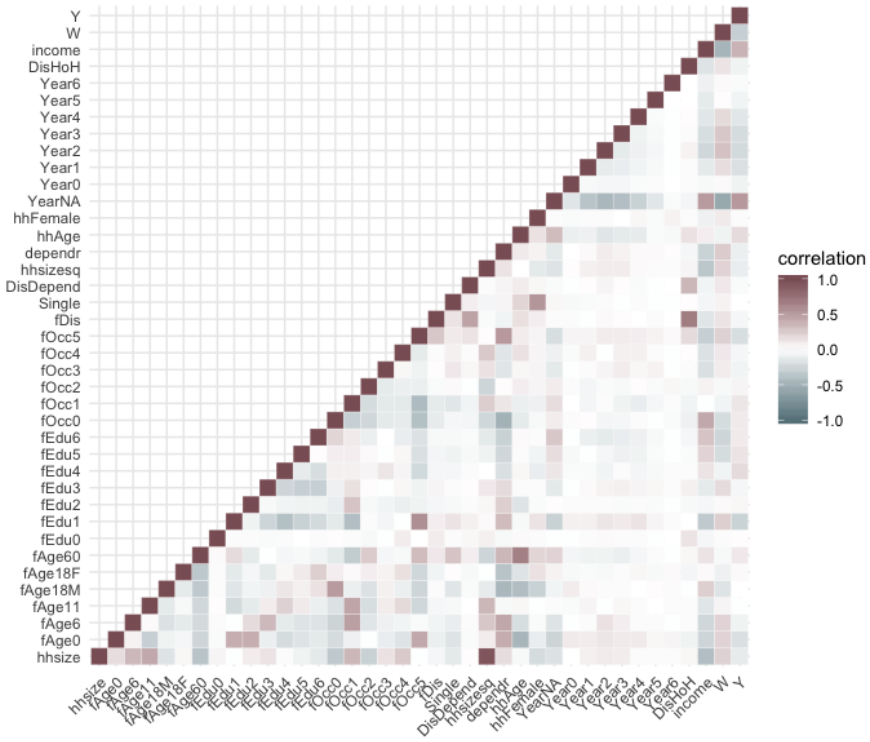


Figure 5: Pairwise correlation plot of all variables used in causal forests

However, a major limitation of traditional approaches is that propensity score is often estimated using parametric models, specifically a logistic regression model where the probability of receiving treatment is modeled as a logistic function of the pre-treatment covariates. If the propensity score model was to be incorrect, which is often the norm in observational data, estimates of the average treatment effect (ATE) might be biased. Figure 6 compares matched observations that have the same propensity scores but differ in the treatment status after implementing standard logistic regression. However, just by visual inspection of the six selected covariates, we can already tell that the matching has not done well because the treatment and control groups have different means at each value of the propensity score. Formally, t-tests results showed that there exist significant differences across the two groups for many of the covariates. Proceeding with the matched samples, I used OLS with covariates to estimate the ATE. Section V.2 outlines the findings.

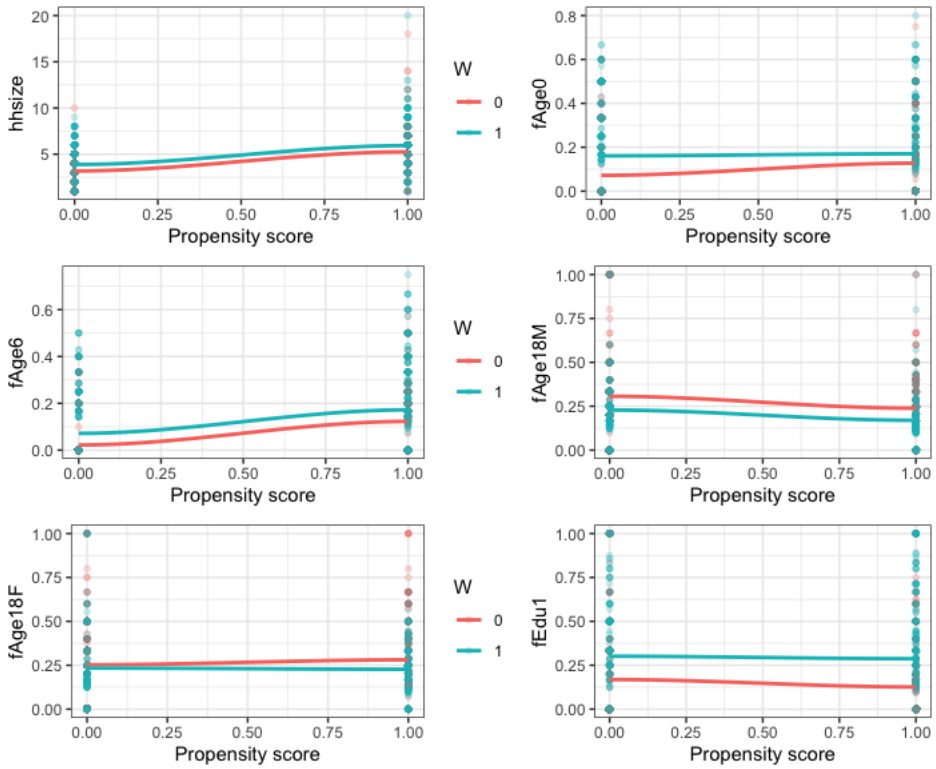


Figure 6: Mean of selected covariates after matching

A further step taken here was using a non-parametric random forests ML method. Trees algorithm does not require a parametric model for the propensity score estimation in the first stage nor for the treatment effects estimation in the second stage. The motivation is that, by allowing algorithms to flexibly choose functional forms, it would potentially lessen the risk of model misspecification. Specifically, after using random forests to estimate propensity scores, under default settings of the *grf* package[4] (Tibshirani et al., 2019) in R (R Core Team, 2017), I implemented causal forests to estimate ATE with the forests-based propensity scores and assessed heterogeneities among each individual household. The *grf* implementation of causal

4 A pluggable package for forest-based statistical estimation and inference. GRF currently provides methods for non-parametric least-squares regression, quantile regression, and treatment effect estimation (optionally using instrumental variables). Reference: https://github.com/grf-labs/grf

forests starts by fitting two separate regression forests to estimate $m(x) = E[Y|X = x]$ and $e(x) = P[W = 1|X = x]$. It then makes out-of-bag predictions: predictions outputs are averaged across trees whose training data did not include the i^{th} observation. In the end, causal forests are grown using these two first-stage forests and the out-of-bag prediction errors (residual from this non-parametric prediction).

V. Results

V.1 Prediction

I built and fine-tuned LASSO regularized linear regression and random forests models using the full, 99%, 98%, and 97% of the training data and predicted income per capita outcome for the households in Lebanon. Table 5 shows that forward stepwise regression, LASSO regularized linear regression, and random forests perform similarly in terms of their predictive results. However, contrary to my hypothesis, by leaving out a small set of observations on the two extrema of the income distribution, the standard metric of prediction accuracy (rMSE) suggests that models trained on subsets of the population perform no better than the commonly used forward stepwise regression model trained on the entire population. Random forests, on the other hand, perform monotonically better than LASSO in each of the selected datasets, even though the difference in testing rMSE is less than 0.1. This is not a huge difference to be considered as a significant improvement, given the outcome of interest (log of income per capita) ranges from 0 to 8.92. All the results were averaged across separate blind validation

tests after model derivation.

Predictive Models	Training rMSE	Testing rMSE	# of Variables Selected	Shrinking Parameter
Forward Stepwise Regression (full)	0.5885	0.6612	22	
LASSO Regularized Regression (full)	0.5815	0.6523	38	0.0018
Random Forests (full)	0.3152	0.6287	36	
LASSO Regularized Regression (99%)	0.5917	0.6633	26	0.0060
Random Forests (99%)	0.3868	0.6336	28	
LASSO Regularized Regression (98%)	0.6349	0.6673	25	0.0100
Random Forests (98%)	0.3963	0.6386	27	
LASSO Regularized Regression (97%)	0.6173	0.6701	37	0.00005
Random Forests (97%)	0.4262	0.6457	33	

Table 5: Prediction Performance Results

One of the advantages of random forests is that it is an ensemble method with each individual tree grown by greedy recursive partitioning. The prediction for a particular observation is determined by averaging predictions across an ensemble of different trees. The trees are randomized using bootstrap aggregation, whereby each tree is grown on a different random subset of the training data. Additionally, the random split selection also restricts the variables available at each step of the algorithm. Lastly, random forests take care of any non-linear relationship between feature sets and predicted the outcome of interest. If there are interactions between features, we do not have to specify them in advance. Figure 7 provides a demonstration of what a decision tree in this prediction model could look like and how features can potentially be used for splits at each node.

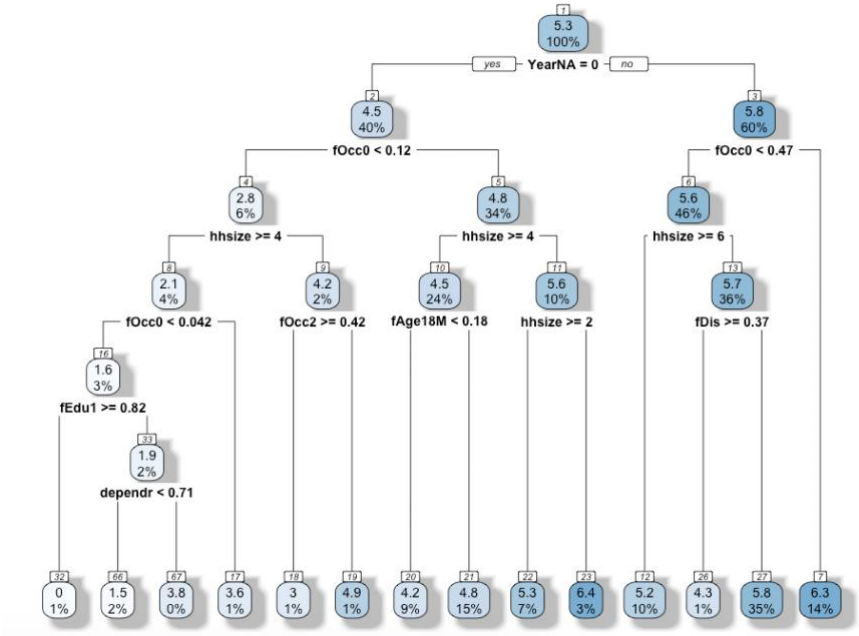


Figure 7: Example of a regression tree with complexity parameter = 0.01

V.2 Heterogeneous Treatment Effects Estimation

As a benchmark, I first implemented parametric logit models to estimate propensity scores, then performed matching by finding nearest neighbor based on Euclidean distance, and in the end paired observations that have similar propensity scores but differ in treatment status. The new dataset contains 1, 770 observations, meaning that 885 pairs of treated and control observations were matched. Finally, I ran an OLS regression with all covariates to estimate the average treatment effects; the result is shown in Table 6. Holding all covariates fixed, receiving subsidies leads to a statistically *insignificant* 0.111 increase in refugee households' living conditions, as measured by the scores assigned to their ratings of house services and living conditions. Though most of the administrative variables are not significant, pre-subsidies income per capita has a statistically significant positive coefficient estimate. This is not

surprising because wealthier households are more likely able to access better house services and afford daily goods and supplies, but this does not provide reliable guidance for studying the treatment effect of interest.

Coefficients:	Estimate (Std. Error)
(Intercept)	-3.672 (4.104)
Household size	0.0500 (0.081)
Frac. of HH members aged 0-5	-0.211 (0.6256)
Frac. of HH members aged 6-10	-0.459 (0.649)
Frac. of HH members aged 11-17	-0.492 (0.516)
Frac. of male members aged 18-50	0.371 (0.373)
Frac. of female members aged 18-50	0.285 (0.364)
Frac. of members older than 60	0.542 (0.399)
Frac. of members no education	3.377* (1.645)
Frac. of members some education below primary	0.245 (0.326)
Frac. of members with primary education	1.286* (0.630)
Frac. of members with intermediate education	0.900** (0.298)
Frac. of members with secondary education	1.464*** (0.297)
Frac. of members above secondary education	1.023** (0.365)
Frac. of members Laborer	-4.037 (3.662)
Frac. of members Student	-3.489 (3.672)
Frac. of members Housekeeper	-4.301 (3.667)
Frac. of members Unemployed	-3.949 (3.669)
Frac. of members Not looking for job	-3.812 (3.675)
Frac. of members Not eligible for work	-4.372 (3.670)

Coefficients:	Estimate (Std. Error)
Frac. of HH members with a disability	0.467 (0.706)
Single Parent	-0.045 (0.144)
Existence of a disabled dependent member	0.098 (0.418)
Household Size Squared	-0.003 (0.006)
Frac. of Dependent HH members	0.420 (0.338)
Age of the HoH	-0.004 (0.006)
Female HoH	0.071 (0.171)
Not displaced by the Syrian Crisis	2.468 (1.792)
Year of Arrival = 2010	0.610 (1.808)
Year of Arrival = 2011	0.521 (1.789)
Year of Arrival = 2012	0.601 (1.787)
Year of Arrival = 2013	0.542 (1.787)
Year of Arrival = 2014	0.676 (1.789)
Year of Arrival = 2015	0.766 (1.814)
Disabled HoH	-0.448 (0.304)
log(income per capita)	0.235*** (0.046)
W: treatment of receiving subsidies	**0.111 (0.104)**
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	
Residual standard error: 1.761 on 1733 df	
Multiple R-squared: 0.338,	
Adjusted R-squared: 0.3243	
F-statistic: 24.58 on 36 and 1733 df,	
p-value: < 2.2e-16	

Table 6: Treatment effects estimation with OLS

To overcome the shortcomings of unknown treatment propensities and poor matching mechanisms with traditional parametric techniques, I applied causal forests to examine treatment effects. The reason for using causal forests is to account for potential heterogeneities across households. As described in Section IV.1, the households in the SRHCS study were not independently sampled due to various challenges; rather, they were all drawn

from segmented CFs and it is reasonable to assume the existence of heterogeneous treatment effects. The average subsidies' effect estimated by causal forests is 0.192 with a standard error of 0.036, indicating a statistically *significant* 0.192 improvement on average when households received subsidies. Figure 8 plots the distribution of the estimated CATEs using causal forests. Figure 9 illustrates the CATE estimates and the corresponding confidence intervals of all observations.

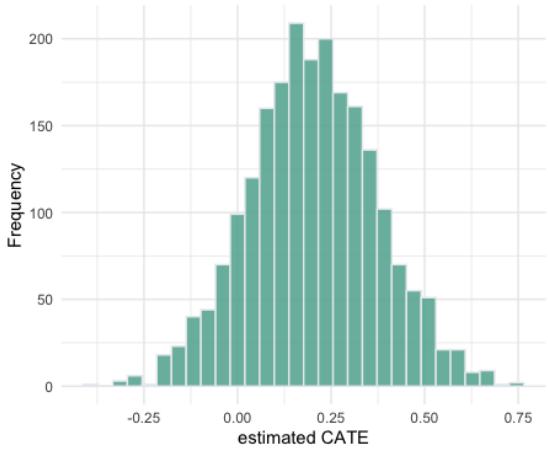


Figure 8: Histogram of out-of-bag CATE estimates from causal forests

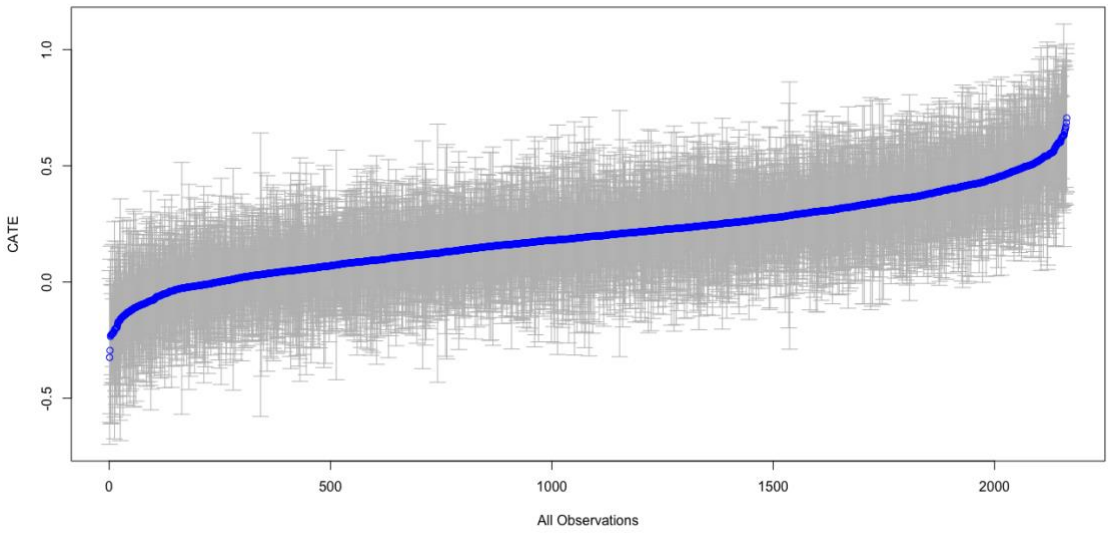


Figure 9: CATEs and the associated 95% confidence intervals for all observations

Non-parametric forests do not produce coefficients estimates, but a measure of variable importance that indicates how often a variable was used in a tree split for maximizing heterogeneity of the treatment effects is available. In Table 7, pre-subsidies income per capita once again stands out as one of the top used features along with other features such as household head being female, the fraction of household members having some intermediate education, and the fraction of male adults in the household.

Names	imp
Female HoH	0.127
Frac. of members with intermediate education	0.090
Frac. of male members aged 18-50	0.070
log(income per capita)	0.069
Frac. of female members aged 18-50	0.064
Year of Arrival = 2014	0.063
Frac. of members Laborer	0.053
Frac. of Dependent HH members	0.050
Frac. of members Housekeeper	0.044
Frac. of members with primary education	0.043
Frac. of members older than 60	0.038
Frac. of members Not eligible for work	0.036
Frac. of members with no education	0.026
Year of Arrival = 2011	0.023
Frac. of HH members aged 0-5	0.020
Frac. of HH members aged 11-17	0.020
Frac. of members with secondary education	0.018
Household Size Squared	0.017

Names	imp
Frac. of members Not looking for job	0.016
Frac. of members Student	0.015
Age of the HoH	0.014
Frac. of members with some education below primary	0.013
Household Size	0.012
Year of Arrival = 2012	0.012
Frac. of members above secondary education	0.010
Single Parent	0.010
Frac. of HH members aged 6-10	0.006
Frac. of members Unemployed	0.005
Not displaced by the Syrian Crisis	0.005
Year of Arrival = 2010	0.005
Year of Arrival = 2013	0.004
Frac. of HH members with a disability	0.002
Disabled HoH	0.000
Existence of a disabled dependent member	0.000
Year of Arrival = 2015	0.000
Year of Arrival = 2016	0.000

Table 7: Variable importance with causal forests

Note: The variable importance measure is a depth-weighted average of the number of splits on maximizing the heterogeneities of households' ratings of their living conditions change upon receipt of subsidies. The variables are ordered by importance, with larger values indicating greater importance.

VI. Discussion

In this study, non-parametric ML models have proved to have more accurate income prediction performance and more reliable estimates of subsidies' causal effects. However, the stability in prediction quality does not imply stability in estimated coefficients. More specifically for this study, because the features of the households might be highly correlated with each

other (e.g. the fraction of household members who have received higher education and the fraction of them who have a decent occupation), such variables can become substitutes in predicting income and similar predictions can be produced using very different variables. Therefore, model selection consistency is not guaranteed. This is the same for causal forests in estimating heterogeneous treatment effects: conditional on a tree, the estimated coefficients are consistent, but we cannot over-interpret variable importance, nor can it be compared across features. For example, the low importance of the variable "fraction of household members being unemployed" should not be interpreted as indicating that it is not related to heterogeneity. In fact, if it is highly correlated with education, then the trees might just split on education but not the other; on other draws of the data, the same procedure could have chosen a tree that split on the fraction of household members being unemployed instead. In general, a feature would be less likely to be chosen if the tree has previously split on another feature highly related to it, but this does not indicate that it is not useful. Moving forward with improving prediction model accuracy, recent literature has also mentioned the use of an ensemble learning method: we can potentially run several models based on different ML algorithms and then average their prediction results with weights chosen by cross-validation (Brownlee, 2018).

Moreover, for the subsidies effects estimation problem using causal forests, we can also test if there indeed exist significant heterogeneities among individual households. Even though it is true that some households have benefited less or more from receiving the subsidies than others, Figure 8 shows that the histogram is roughly concentrated at a point instead of being widely spreading out, which suggests that the treatment effects might not differ significantly

across households. To test the hypothesis about heterogeneity, I first summarized the output of causal forests and created subpopulations based on predicted treatment strength. After splitting the data into groups based on four tiles of predicted treatment effects, I computed the average treatment effect within each subgroup by taking the average difference between raw outcomes for treated and control groups. Lastly, I ran a linear regression of "Y" (assigned living conditions improvement score) on the computed average treatment effects interacted with "W" (indicator variable for treatment) in each of the four tiles. Linear hypothesis testing result is shown in Table 8: there is not enough evidence to reject the null. To better visualize the ATEs across four quantiles, Figure 10 plots the confidence intervals for all observations in the four tiles – no clear heterogeneities can be inspected. Further investigation could be around testing heterogeneities across covariates. Comparing all covariates across n-tiles of treatment effects could present a fuller picture of how high-treatment-effect individuals differ from lowe-treatment-effect individuals.

Linear Hypothesis Test

H_0: Average treatment effects are the same across 4 tiles

ntile1 × W - ntile2 × W = 0

ntile1 × W - ntile3 × W = 0

ntile1 × W - ntile4 × W = 0

	Res.Df	RSS	Df	F	Pr(>F)
Model 1: restricted model	2158	8992			
Model 2: unrestricted model	2155	8989	3	0.3	0.83

Table 8: F-test of heterogeneity in average treatment effects across 4tiles

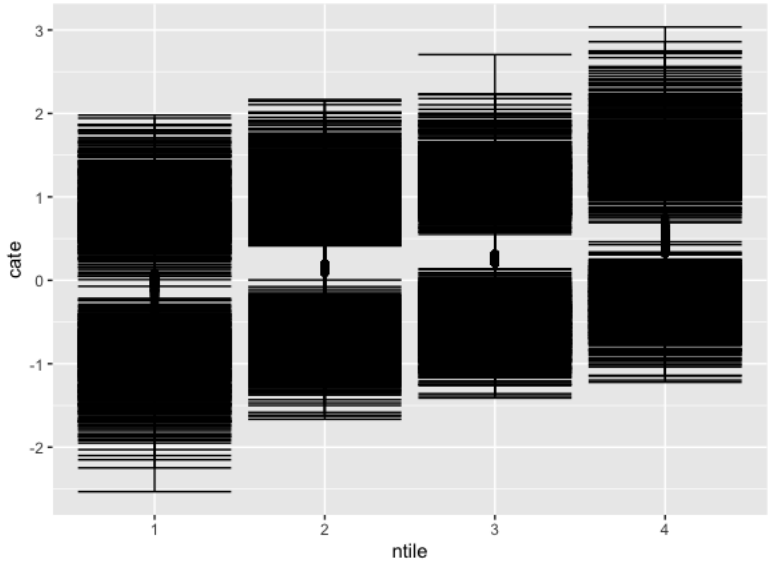


Figure 10: Confidence intervals (95%) for all observations in 4-tiles

It is worth noting that in contrast with traditional matching and OLS regression results, the ATE of receiving subsidies using causal forests is statistically significant. This could be attributed to the non-parametric modeling of propensity score and treatment effects estimation. Poor matching mechanisms and the non-selective, linear relationship enforced by linear regressions failed to work effectively in high-dimensional settings. Even though only an average of 0.192 causal improvements of households receiving the subsidies has been shown, the impact of allocating aid to designated households is still positive. Given that approximately 6% of the households rely entirely on theses assistance to survive, this study suggests that social welfare programs offer a small but still statistically significant remedy for households in extremely vulnerable conditions with the average effect being successfully quantified.

A major limitation of the study is that I assume unconfounded-ness in order to identify the *causal* subsidies effects. It is only with no substantial variation between treatment and control group that I can conduct hypothesis tests about the magnitude of differences in

treatment effects across subsets of the population using causal forests. To relax this assumption, further research work could be dedicated to finding an instrument for treatment assignment (Angrist and Pischke, 2008), or conducting a sensitivity analysis for hidden confounding (Rosenbaum, 2002). Generalized random forests (Athey et al, 2018) also enable treatment effects estimation with instrument variable (IV) applications. However, this does not affect the conclusion of this paper that subsides have been effectively improving beneficiaries' living conditions on average. This is because even though bias exists due to unobservable confounders like subjective standards upon which they rate their conditions and spending, the bias is more likely to be downward. Households receiving the subsidies in the first place are under underprivileged situations, so it is likely that they usually have a lower expectation on house services and living standards and thus higher ratings on the post-subsidies' conditions. Therefore, negative bias could be expected. The true causal treatment effect would be even larger than what has been estimated in this study.

In addition to applying more off-the-shelf techniques in ML, from an econometric standpoint, further improvements can come from data collection mechanisms and other data quality considerations. The ideal would always be having a (nearly) randomized controlled trials for subsidies allocation. In reality, econometric tools like IVs are typically challenging to find in a retrospective way. However, if, say, a social welfare program had clear cut-offs for determining the eligibility of receiving the aid based on multiple metrics and the households were not aware of these standards, we would be able to implement regression discontinuity designs by only looking at observations around the cut-offs. However, sufficient data for model training purposes are needed, and if we want to have more reliable estimates of the heterogeneous

treatment effects across different subgroups of populations around the cut-offs, we would require more intense data collection work. The idea is that if we want to target for households who would benefit the most *after* the subsidies, we would want to find out the counterfactual outcomes of them not receiving the aid. We nonetheless never observe the true treatment effects. And even with unconfounded-ness assumption satisfied, the results obtained using causal forests cannot be generalized further if the variance of each individual CATE is large. In general, heterogeneity across different subgroups of population cannot be guaranteed.

VII. Conclusion

Overall, the major contributions of this study are threefold. The first one is aligned with the continuous efforts for optimizing algorithmic prediction model design: the proposed nonparametric forests approach has proven to perform slightly better than parametric modeling techniques. The second is the initiative of allocating subsidies based on households' estimated treatment effects. Since the tradition has been targeting based on expenditure or income, this study provides new insights for adjusting the objectives of social welfare programs and reorienting the targets for future beneficiaries. The third contribution lies in the possibility of creating a replicable ML methodology for assessment of program effectiveness, which could contribute to the development of scalable targeting systems for social welfare programs in a similar context worldwide.

More broadly, Athey & Imbens (2019) discuss the role of which machine learning methods can play in economics and econometrics research and how the integration of these disciplines can achieve better performance for problems related to prediction and causal

inference. Indeed, the emergence of abundant datasets made the application of machine learning in empirical research appealing. Among all, prediction problems present the most ideal scenario to employ statistical learning – its protection against overfitting, estimation of valid confidence intervals, and insurance of fairness and non-manipulability provides a powerful and flexible way of making quality predictions. The prediction results can be carried further as a crucial component to better identify causal links, more accurately estimate treatment effects, and develop new approaches to cross-validation optimized for causal inference and optimal policy design.

Another advantage of employing machine learning in settings with high-quality granular datasets is that we can have models with better prediction performance and more stable treatment effects estimation over time, as more data come in and the algorithms self-adjust according to predetermined penalty matrices. Moreover, beyond optimizing model design to generate more accurate prediction models, future research work with a focus on the interpretability of coefficient estimates and robust measures for causal parameters is needed. For the social targeting programs specifically, there is also rapidly growing literature using machine learning together with images from satellites and street maps to predict poverty, safety, and home values. For example, imagery from satellites or Google's street view can be used in combination with survey data to train models that can be used to produce estimates of economic outcomes at the level of individual home in developing countries (e.g. Jean et al. (2016), Engstrom et al. (2017), Naik et al. (2014)). Results from these findings will potentially offer improvement for the current selection process by enabling stakeholders to determine the

impact metrics and consider who their intended beneficiaries should be. The goal, eventually, is to identify strategies for helping the poor in both the immediate and the longer term.

References

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., and Tobias, J. (2012). Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review*, 102(4):1206–40.

Altindag, O., O'Connell, S., Sasmaz, A., Balcioglu, Z. Jerneck, M., and Kunze Foong, A. (2019). Improving scalable poverty targeting: Design and validation of an econometric targeting model for basic needs cash assistance to Syrian refugees in Lebanon. *Working Paper https://www.stephenoconnell.org/publication/aosb2019/*

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., and Salvestrini, V. (2018). Targeting with machine learning: An application to a tax rebate program in Italy. *Journal of Economic Behavior & Organization*, 156:86 – 102.

Angrist, J. D., and Pischke, J.-S. (2015). *Mastering 'Metrics: The Path from Cause to Effect*, Princeton University Press, 1: pp. 301-305.

Athey, S., Tibshirani J., and Wager, S. (2018). Generalized Random Forests. *The Annals of Statistics*. arXiv:1610.01271v4.

Athey, S. and Imbens, G. (2019). Machine Learning Methods Economists Should Know About. *arXivpreprint arXiv:1903.10075.*

Athey, S. (2015). "Machine Learning and Causal Inference for Policy Evaluation." In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 5–6. ACM.

Athey, S, and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America* vol. 113,27: 7353-60.

Athey, S. (2018). The Impact of Machine Learning on Economics. *The Economics of Artificial Intelligence: An Agenda,* National Bureau of Economic Research, pp. 507-547.

Brown, C., Ravallion, M., and van de Walle, D. (2018). A poor means test: Econometric targeting in Africa. *Journal of Development Economics*, 134:109-124.

Brownlee, J. (2018). Ensemble Learning Methods for Deep Learning Neural Networks. *Machine Learning Mastery*: Deep Learning Performance. https://machinelearningmastery.com/ensemble-methods-for-deep-learning-neural-networks/

Coady, D., Grosh, M., and Hoddinott, J. (2004). *Targeting of transfers in developing countries: Review of lessons and experience*. The World Bank.

Cunha, J.M. (2014). Testing Paternalism: Cash versus In-Kind Transfers. *American Economic Journal: Applied Economics*, 6(2):195-230.

Engstrom, T., Hersh, J., and Newhouse, D. (2017). *Poverty from space: using high-resolution satellite imagery for estimating economic well-being.* The World Bank.

Griffith, R., von Hinke, S., and Smith, S. (2018). Getting a healthy start: The effectiveness of targeted benefits for improving dietary choices. *Journal of health economics*, *58*, 176–187.

Hoynes H., Schanzenbach W.D., and Almond D. (2016). Long-run impacts of childhood access to the safety net. *American Economic Review*, 106(4):903–934.

Hastie, T., Friedman, J., and Tibshirani, R. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Kidd, S. and Wylde, E. (2011). *Targeting the Poorest: An assessment of the proxy means test methodology*. Australian Agency for International Development.

Kidd, S., Gelders, B., and D. Bailey-Athias (2017). Exclusion by design: an assessment of the effectiveness of the proxy means test poverty targeting mechanism. ILO working papers, International Labour Organization.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review,* 105(5): 491–495.

Krishnan, N. et al (2019). *Survey Design and Sampling: A methodology note for the 2015-16 surveys of Syrian refugees and host communities in Jordan, Lebanon, and Kurdistan, Iraq.* The World Bank Group.
https://microdata.worldbank.org/index.php/catalog/3471/download/46669

Kshirsagar, V., Wieczorek, J., Ramanathan, S., and Wells, R. (2017). Household poverty classification in data-scarce environments: a machine learning approach. *arXiv preprint arXiv:1711.06813*.

Künzel, S.R., Sekhon, J.S., Bickel, P.J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*. 116 (10) 4156-4165; DOI:10.1073/pnas.1804597116.

McBride, L. and Nichols, A. (2018). "Retooling poverty targeting using out-of-sample validation and machine learning." *World Bank Economic Review,* 32(3):531–550.

Mckenzie, D. (2018). “How can machine learning and artificial intelligence be used in development interventions and impact evaluations?” Retrieved from https://blogs.worldbank.org/impactevaluations/how-can-machine-learning-and-artificial-intelligence-be-used-development-interventions-and-impact?utm_source=GDPR&utm_campaign=1e284b7a0f-EMAIL_CAMPAIGN_2018_11_23_02_39&utm_medium=email&utm_term=0_7c51e322b7-1e284b7a0f-278644353

Mullainathan, S. and Obermeyer, Z. (2019). Who is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error. NBER Working Paper No. w26168. https://ssrn.com/abstract=3439178

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

Naik, N., Jade P., Ramesh R., and César H. (2014). Streetscore: Predicting the Perceived Safety of One Million Streetscapes. *IEEE CVPR Workshops*, 793–99. Washington, DC: IEEE Computer Society.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of propensity score in observational studies for causal effects. *Biometrilca* (1083), 70, 1, pp. 41-55.

Rosenbaum P.R. (2002). *Overt Bias in Observational Studies.* In: Observational Studies. Springer Series in Statistics. Springer, pp. 71-104.

UNHCR (1967). Convention and Protocol Relating to the Status of Refugees. Geneva, Switzerland: Office of the United Nations High Commissioner for Refugees (UNHCR), Communications and Public Information Service, pp. 2-3.

Varian, H.R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2): 3-28.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner. *U.C. Berkeley Division of Biostatistics Working Paper Series.* Working Paper 222. https://biostats.bepress.com/ucbbiostat/paper222

van der Laan, M. J., and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. (Springer Series in Statistics)*. Springer.

Veras, F., Peres, R. and Guerreiro, R. (2007). Evaluating the Impact of Brazil’s Bolsa Família: Cash Transfer Programmes. *Comparative Perspective*, IPC Evaluation Note No. 1, International Poverty Centre.