Using Causal Forests to Estimate Treatment Heterogeneity: Encouraging Tap Water Adoption in Urban Morocco

Barkat Sikder BSIKDER1@SWARTHMORE.EDU
Parker Snipes PSNIPES1@SWARTHMORE.EDU
Peem Lerdputtipongporn PLERDPU1@SWARTHMORE.EDU

Abstract

This paper uses examines the data from an RCT conducted on the effectiveness of a behavioral and credit-based intervention designed to increase water connection rates in Tangiers, Morocco. The authors used causal forests to derive the heterogeneous treatment effects for a number of salient features, including household income, distance to nearest public tap, and number of small children. Although the treatment was broadly effective, with a 57% increase in connection, it was particularly effective for households with monthly income below \$2500 (59%) and within 71 meters of the nearest water tap 71%. We conclude that the intervention has some degree of targeting benefits, and examine causal forests' potential as a tool for economists and policymakers.

1. Introduction

In this section, we discuss a study in development economics as an example case to motivate the use of causal machine learning techniques, in particular the Causal Forest, for public policy analysis and compare their use to more traditional methods of linear regression and supervised learning.

1.1. Motivating Example Study

For many policymakers, especially those in developing countries, the provision of public goods and services to citizens under a limited budget is among the important aspects of the job. They often face a range of problems and have to allocate scarce resources between them, doing their best to target their interventions to the groups who need them most. As an example, we will consider Devoto et al's 2011 work regarding the welfare effects of tap water adoption in Tangiers, Morocco (Florencia Devoto, 2012). Many citizens do not have their own water tap at home and their daily access to water is via public wells, fountains, or their neighbor's tap, among other sources. Some of the obstacles that prevent private tap adoption include shortage of finances, lack of knowledge regarding how to get a tap, and unawareness of the benefits. However, connecting private households to the water main often cannot be publicly financed since it is expensive.

The controlled randomized trial (RCT) found that willingness to pay for a private connection is high when it can be purchased on credit, not because a connection improves health but because it increases the time available for leisure and reduces conflicts on water matters. This suggests that facilitating access to credit for households to finance lump sum quality-of-life investments can increase welfare even if those investments do not result in any health or income gains (Florencia Devoto, 2012).

Since households cannot be forced to adopt piped connections, they can only be encouraged to do so through targeted advertising and easy-credit access campaigns. Two obstacles arise for policy makers that make it inefficient to target every household with such campaigns. Firstly, running such campaigns is expensive, even if it is less expensive than public financing of the tap itself. Secondly, people respond differentially to campaigns: some people are more likely to be benefit than others, while others are less affected and some may even respond negatively. In this paper, we discuss the application of a novel machine learning technique called Causal Forests to help determine how to target public policy interventions.

1.2. Data Overview

It will help in the subsequent sections if we explain the data we worked with. We use the publicly available data generated by the RCT (Florencia Devoto, 2012). This data consisted of two baseline household surveys, one conducted before treatment was implemented and one after, capturing relevant demographic and household information for both

CPSC 66 Machine Learning Proceedings, Swarthmore College, Spring 2021.

the treatment and control groups. Data was recorded for the 845 households who participated in the baseline survey with 434 in the treatment group and 411 in the control group.

Beyond household characteristics, features consist of answers to survey questions, meaning there are many sparse variables. To conduct our analysis, we extracted relevant features and outcomes from this dataset. We decided to extract a number of key features: household income, employment status, the existence of informal pipeline connection, minimum distance to public tap, children's health, self-reported health, residence type, number of rooms, and number of small children, since these features had the greatest impact on connection rates and reduction of social outcomes. Our main outcome variable is new connection (a binary representing whether the tap was adopted between the first and second surveys). Using these two outcomes allows us to estimate the conditional effects of both the treatment's effect on uptake, and the overall welfare benefit of the treatment, as social tension was estimated by the RCT to be the welfare outcome most affected by the treatment.

Despite the fact that the data was collected in a relatively controlled setting, many features (especially those which were self-reported) had missing values. We cleaned the data by imputing with the variable's mean. Although this is a potential source of bias, given that households which did not report certain information may differ from the general populace in systematic ways, enough features are missing from the dataset that dropping all households missing any feature value would significantly decrease our sample size.

1.3. Research Problem and Candidate Solutions

Given the dataset, we wanted to understand the causal influence of treatment on connection. In particular, we are interested in how these causal influences vary by subgroups of the population. So, we ask two questions.

(i) If we want to target people with ad and information campaigns, how do different groups respond to campaigns by getting a tap connection, and what are the definitive features of these groups?

The fundamental problem of causal inference is that we do not observe true intervention effects. It is common place in many fields of research to use Ordinary Least Squares (OLS) regression to study these relationships and make arguments about the average treatment effects (ATE) given certain sets of assumptions, such as the nature of an RCT (Angrist, 2004). The differential impact of a treatment on an outcome, i.e. impact conditioning on certain values of the other control variables is known as a heterogenous or conditional treatment effect (CATE) and it is estimated by interaction terms. However, the conditional impact does not usually simply depend on one or two covariates, and it becomes difficult to estimate multiple interaction effects. Although this disincentives finding spurious results, OLS is conservative and often does not uncover unexpected but important results which makes us consider alternatives for answering our research questions.

Supervised learning has been used for predicting likelihoods of responses given features for many year but recently, they have been combined with counterfactual modelling techniques for causal inference, such as Generalized Random Forests based on the idea of the Random Forest (Athey & Imbens, 2015). A Causal Forest is one example that allows us to efficiently navigate multiple interaction effects without expending the effort needed to regress every pair of features across the dataset while providing the same set of estimates relevant to statistical inference. Hence, in this paper, we present our experiments with the Causal Forest to answer question (i).

2. Methods

In this section, we broadly introduce the Causal Forest before going into the details of the algorithm and the various ways a user interacts with an implementation of the algorithm in order to provide the theoretical background necessary to understand the work we present later in our paper.

2.1. Causal Forest Overview

Let us recall some properties of the more well-known Random Forest algorithm (Brieman, 2001). The method uses an ensemble of decision trees to predict an outcome variable Y and it creates independence by bootstrapping and feature subsampling.

The first important difference between a Random Forest and a Causal Forest is the splitting criterion. For Random Forests, the variable and value to split at each node are chosen such that the greatest reduction in the error, such as mean squared, with regard to the outcomes Y is achieved. With Causal Forests, the prediction of a treatment effect is given by the difference in the average outcomes Y between the treated and the untreated observations in a terminal node. The splitting criterion is adapted such that it searches for a partitioning where the treatment effects differ the most including a correction that accounts for how the splits affect the variance of the parameter estimates (Athey & Imbens, 2015). Although we can't directly observe a given observation both treated and untreated, the fact that our treatment and control groups were randomized allows us to compute the treatment effect.

The second major difference is the ”honesty” criterion of the trees. For each tree, the training data is split into two subsamples: a splitting subsample and an estimating subsample. The former is used to perform the splits and thus grow the tree and latter is then used to make the predictions. The prediction of the treatment effects is then given by the difference in the average outcomes between the treated and the untreated observations of the estimating subsample in the terminal nodes. This estimator is consistent and asymptotically normal which giving valid confidence intervals (CIs), which make it usable for inference tasks (Athey & Imbens, 2015).

2.2. Algorithm Steps

The following is a breakdown of the steps of the Causal Forest algorithm (Davis & Heller, 2017).

(i) Draw a subsample b without replacement containing $n_b = k_1 N$ observations from the N observations where k_1 is the subsample size.

(ii) Randomly split the b into equally sized sets tr and e for training and prediction respectively.

(iii) For each value of each covariate, $X_j = x$, nominate candidate splits of the observations into two groups based on whether $X_j \leq x$. Then, consider only splits where there are at least k_2 treatment and k_2 control observations in both new leaves and choose the single split that maximizes an objective function O capturing how much the treatment effect estimates vary across the two resulting subgroups, with a penalty for within-leaf variance. If this split increases O relative to no split, implement it and repeat this step in both new leaves. If no split increases O, this is a terminal leaf.

(iv) Once the tree is fully grown for b, make predictions on e using the features.

(v) Using only e, calculate $\hat{\tau}(x)_l = \bar{y}_{Tl} - \bar{y}_{Cl}$ within each terminal leaf l. This step makes the tree honest, since treatment effect estimates are made using different observations than the ones that determined the splits.

(vi) Save the prediction $\hat{\tau}(x)_{l,b}$ for b.

(vii) Repeat steps (i) through (vi) for B trees.

(viii) Define observation i's predicted CATE as $\hat{\tau}(x)^{CF} = \frac{1}{B}\sum_{b=1}^{B} \hat{\tau}(x)_{l,b}$ which is the average across trees.

2.3. Implementation

The implementation of the algorithm we will use is Microsoft Research's EconML package (citation). The model must be instantiated with certain default hyperparameters. But in particular, the three we need to choose are the number of trees B, the subsample size k_1, and the minimum number in each leaf k_2. More trees reduce the Monte Carlo error introduced by subsampling. Increasing the minimum number of observations in each leaf trades increases bias and reduces variance; bigger leaves make results more consistent across different samples but predict less heterogeneity, hence underfit. Smaller subsamples reduce dependence across trees but increase the variance of each estimate (Research, 2019).

Once we fit the model to our data, we can get statistical summaries of our data for inference. The importance of each feature is given by how much of the total normalized heterogeneity it creates. Shapley (SHAP) values, a method borrowed from cooperative game theory, enable us to visualize feature importance, magnitude of impact of particular feature value, and the direction of correlation.

In addition, the trained Causal Forest can then be converted and interpreted as a decision tree, the nodes of which contain the information regarding the splitting criterion, number of examples at node, mean of the CATE and its standard error and confidence interval. The decision tree is created by considering the predicted heterogeneous treatment effect of the trained model as the example label and fitting a decision tree to it.

3. Experiments and Results

We apply the Causal Forest method described in the previous section on the RCT data described in Section 1.2 to explore questions two questions posed in Section 1.3. We will walk through our experimental approach with regards to the connection and treatment, present some results, and discuss some of the limits of inference.

3.1. Model Selection

As mentioned in Section 1, our data includes the X, T, and Y, and for this subsection, T is assignment to the treatment group and Y is tap connection to water. In order to make valid inferences, we need to initialize and select a good model. As we initialized our model, some of the design choices we made include the following. The number of trees of the forest were set at $B = 10,000$. Given the small nature of the social experiment, we found this sufficiently large for tight estimates without compromising speed of computation. The splitting and estimation sets of each honest tree were created through 3-fold cross validation to get each tree prediction. We also tuned the hyperparameters k_1 and k_2 of the final stage causal forest by training small forests of 100 trees on a grid of parameters and tested the out of sample R-score. The values $k_1 = 0.5$ and $k_2 = 10$ were chosen with no maximum depth restriction. Training this model predicts the pathway in which groups become most different in regards to the response to the treatment.

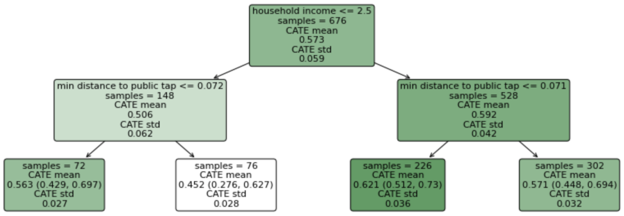

Figure 1. Interpretation of CATE of treatment on connection in train set with $depth = 2$

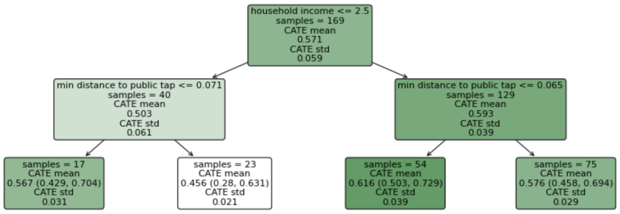

Figure 2. Interpretation of CATE of treatment on connection in test set with $depth = 2$

3.2. Model Interpretation

To see whether the ATE were similar between training and testing sets, we run hold-out testing on a $80/20$ train-test split and compare them to the ATE estimate of OLS regression. The non-parametrically calculated ATE is 0.57 with p-value 0.03 while the parametric OLS estimate on the whole dataset is 0.58 with p-value 0.02. A t-test of these estimates does not reject the null hypothesis that these two are the same. This suggests that being targeted by the campaign increases the average probability of any household installing a tap by 57%.

Next, we want to see the subgroups identified to have differential effects on the adoption of a water connection due to treatment in the training set mentioned above. We do this by plotting the interpretation tree described in Section 2.3 in Figure 1 with $maxdepth = 2$ for visual parsimony. Multiple heterogeneous treatment effects can be inferred: consider the second leaf for example, households whose income was less than the second quintile and live more than 72 meters away from a public tap are least likely to respond to the campaign with CATE being 45.6% which is noticeably lower than the average. But we want check whether the heterogeneity identified in the training data holds for the test set or if it was erroneous. We get a similar results with the testing set as seen in Figure 2; the splits are also at the second quintile and at 71m. Figure 3 shows the Shapley values calculated on the training set also visually supplement the decision tree interpreter plots. Although we do not speculate the social-scientific meaning of these results, why such findings are the case in the data could be the subject of further speculation by experts.

3.3. Limitations

Despite careful model selection, we should interpret the results with caution since our validation method has been simply weak hold-out validation. Although we re-ran our models with different random seeds and got similar results as described, in the ideal scenario, we should be able to perform k-fold cross validation to check that the subgroups we identify, i.e. nodes of the interpretation tree, are in fact statistically the same between the training and test sets by conducting hypothesis testing so that we can say our trained forests can generalize. Then to check that the subgroups are in fact statistically different within a level of a tree and redundant splits are not made, pairwise hypothesis testing regarding the CATE of each subgroup should be performed. We note these limitations of our method since we do not address them here.

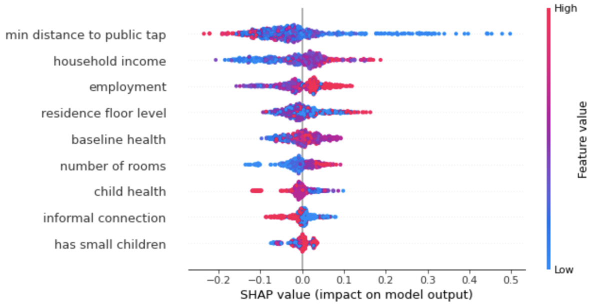

Figure 3. SHAP values of effect of treatment on connection

4. Stakeholders and Ethical Implications

Particularly for socially-oriented projects informed by machine learning algorithms, it's important to consider the impact of using such a system to distribute treatment. Given the limited capacity of governments to provide services for citizens at a municipal level, optimal policy targeting becomes an important problem for local officials. Given this, we think that the use of this system to help target policies would have a number of relevant stakeholders. First, the citizens of the city are necessarily stakeholders, as they feel the effects of the policies put into place. As citizens may be differentially targeted, transparency surrounding the methodology (even for a relatively small-scale program) are key. Further, the public nature of the projects puts some additional constraints on policymakers: targeting a policy on immutable characteristics, for example, would likely lead to a public outcry. Second, the water companies are a second set of stakeholders. Since water companies have a direct profit incentive to boost connection rates, they will likely collaborate with the government on this front. This changes the relationship between customers and the company to some degree, and government officials should take care to ensure that the provider's pricing does not meaningfully change in response to artificially-induced demand. Finally, as the party setting the policy, government officials are stakeholders in such a program. Although most policies are unlikely to be undertaken as a result of randomized controlled trials and machine learning analysis, officials should take special care ensuring that outcomes are fairly equitable and should monitor outcomes to ensure that the approach outperforms more conventional allocation methods.

5. Conclusions

In this paper, we set out to determine the degree to which each feature value collected in the RCT affected connection and social tension. After examining the data, we make a number of important findings. First, we confirm the paper's result that targeting households with the campaign meaningfully increases connection uptake. The estimate is relatively close, with the paper finding a 60% increase in connection resulting from the treatment and our method finding a 57% increase. Secondly, we find that household income and distance to nearest public tap are two of the most important determinants of connection following the treatment. Households within 72 meters of the nearest public tap are only 50% more likely to respond to the treatment, meaningfully lower than the sample average. Similarly, in the tension case, a household having 1.3 or more small children increased the treatment's reduction of social tension from 15% to 21%. These findings suggest that treatment should be focused toward lower-income households closer to the city center, and that potentially larger family homes should also be targeted. Finally, and perhaps most importantly, our findings reinforce the notion that causal forests are an important causal inference tool for economists and policmymakers. Although machine learning does not have a deep foothold in economics, causal tools drawn from computer science have taken on increasing importance in the empirical literature. We hope to contribute one such case to help develop an additional, useful tool for evaluation of causal interaction.

Acknowledgments

We would like to thank Professor Ameet Soni and Professor Benjamin Mitchell for supporting us through this project and an incredible semester.

References

Angrist, Joshua. Treatment effect heterogeneity in theory and practice. *The Economic Journal*, 114(494):52–83, 2004.

Athey, Susan and Imbens, Guido. Recursive partitioning for heterogeneous causal effects. *Annual Review of Economics*, 11:685–725, 2015.

Brieman, Leo. Random forests. *Machine Learning*, 45: 5–32, 2001.

Davis, Jonathan and Heller, Sara. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107:546–550, 2017.

Florencia Devoto, Pascaline Dupas, William Parienté & Vincent Pons. Happiness on tap: Piped water adoption in urban morocco. *American Economic Journal: Economic Policy*, 12(4):68–99, 2012.

Langley, P. Crafting papers on machine learning. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Research, Microsoft. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML, 2019. Version 0.x.