CESifo WORKING PAPERS

9037
2021

April 2021

Optimal Targeting in Fund-raising: A Machine-Learning Approach

Tobias Cagala, Ulrich Glogowsky, Johannes Rincke, Anthony Strittmatter

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9037*

Optimal Targeting in Fundraising: A Machine-Learning Approach

Abstract

Ineffective fundraising lowers the resources charities can use for goods provision. We combine a field experiment and a causal machine-learning approach to increase a charity's fundraising effectiveness. The approach optimally targets fundraising to individuals whose expected donations exceed solicitation costs. Among past donors, optimal targeting substantially increases donations (net of fundraising costs) relative to bench-marks that target everybody or no one. Instead, individuals who were previously asked but never donated should not be targeted. Further, the charity requires only publicly available geospatial information to realize the gains from targeting. We conclude that charities not engaging in optimal targeting waste resources.

JEL-Codes: C930, D640, H410, L310, C210.

Keywords: fundraising, charitable giving, gift exchange, targeting, optimal policy learning, individualized treatment rules.

Tobias Cagala
Deutsche Bundesbank
Frankfurt am Main / Germany
tobias.cagala@bundesbank.de

Ulrich Glogowsky
Johannes Kepler University
Linz / Austria
ulrich.glogowsky@jku.at

Johannes Rincke
Friedrich Alexander University of Erlangen-Nuremberg / Germany
johannes.rincke@fau.de

Anthony Strittmatter
CREST-ENSAE, Institut Polytechnique
Paris / France
anthony.strittmatter@ensae.fr

April 19, 2021
We have benefited from comments by Maja Adena, El-liott Ash, Pierre Boyer, Bruno Crépon, Philipp Dörrenberg, Michael Knaus, Andreas Haufler, Steffen Huck, Michael Lechner, Benjamin M. Marx, Michael P. Price, Dominik Sachs, and Benoit Schmutz. We also thank conference and seminar participants at the ETH Zurich 2021, the Causal Data Science Meeting 2020, the Annual Conference of the National Tax Association 2019, the Annual Conference of the German Economic Association 2019, the Annual Conference of the International Institute of Public Finance 2019, and the WZB Workshop on Charitable Giving in 2018. We gratefully acknowledge financial support from the Swiss National Science Foundation (Spark Project 190422, Strittmatter), the French National Research Agency (LabEx Ecodec/ANR-11-LABX-0047, Strittmatter), the Hans-Frisch-Stiftung, and the Emerging Field Initiative at University of Erlangen-Nuremberg (Cagala, Glogowsky, and Rincke). The paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff.

1 Introduction

Fundraising is a costly activity: The 25 largest US charities spend between 5% and 25% of total donations on fundraising expenses (Andreoni and Payne, 2011). These numbers are a matter of concern for two reasons. First, high fundraising costs leave a smaller proportion of overall donations to finance charitable projects. This can lead to an under-provision of the goods and services that charities provide and may, thus, lower welfare if the donors' utility depends on provision levels (Rose-Ackerman, 1982; Name-Correa and Yildirim, 2013). Second, high fundraising costs also matter from the charity's perspective because donors are averse to financing overhead costs (Tinkelman and Mankaney, 2007; Gneezy *et al.*, 2014). Hence, charities with excessive fundraising expenses will be less successful in raising donations. In conclusion, reducing disproportional fundraising costs can be crucial, from both the welfare and charity-management perspectives. However, while there is broad literature studying how different fundraising instruments such as matching grants and unconditional gifts affect donors' behavior (surveyed by Andreoni and Payne, 2013), previous research has paid less attention to how charities can increase the cost-effectiveness of fundraising.

In this paper, we exploit a causal machine-learning-based approach to maximize a charity's fundraising effectiveness: optimal targeting of fundraising activities to potential donors.[1] Machine-learning-based optimal targeting exploits the possibility that, due to heterogeneity in donors' preferences, the effects of any fundraising campaign are likely heterogeneous across individuals with different observable characteristics.[2] Charities that ignore this heterogeneity may engage in loss-leading fundraising, for example, by directing a costly fundraising instrument to notorious nondonors or donors who, in response to the instrument, do not increase their donations enough to cover the instrument's cost. Against this backdrop, our key contribution is to use machine learning to identify a targeting rule (out of all feasible rules) that is optimal in the sense that it maximizes expected profits by avoiding loss-leading solicitations.[3] Consequently, charities that engage in this type of optimal targeting can increase net donations raised (i.e., donations net of fundraising costs) and, hence, provide more goods and services.

We demonstrate the potential of machine-learning-based optimal targeting using a common fundraising instrument as an example: small unconditional gifts that accompany a solicitation letter. In theory, such unconditional gifts should increase donations by triggering a reciprocal reaction in recipients (Falk, 2007). In practice, the evidence on the effectiveness of unconditional gifts is, however, mixed. Some studies suggest that

[1] See Athey and Imbens (2019) for a review of causal machine-learning methods.

[2] For example, donation motives (such as altruism, warm glow, and reciprocity) are heterogeneously distributed across individuals (Falk *et al.*, 2018). Consequently, responses to fundraising activities that leverage (some of) these motives are likely heterogeneous as well.

[3] A targeting rule is feasible if it is a deterministic function of the observable characteristics.

unconditional gifts are an effective fundraising instrument, while others find that they do not affect average donations or even backfire and lower giving (see the literature section for a discussion). This type of effect heterogeneity might very well operate through individual heterogeneities, such as heterogeneous social preferences. Thus, fundraising gifts offer a promising context to study the benefits of machine-learning-based optimal targeting.

The goal of optimal targeting lies in directing a fundraising instrument (a gift in our example) to a subset of individuals such that the fundraising campaign's expected profits are maximized (i.e., the additional expected net donations collected by the campaign). Yet, who should charities target for that purpose? By definition, the *optimal targeting rule* states that the profit-maximizing targets are the so-called net donors: individuals whose expected additional donation is higher than the marginal fundraising cost. In an ideal world, the charity would know each individual's donation with and without the fundraising instrument (i.e., the gift) and could, thus, determine the set of net donors. In reality, however, this set is unknown to the charity. The reason is that donations under both conditions (gift vs. no gift) are unknown before the campaign. Moreover, even after the campaign, the charity could still only learn a donor's behavior under her assigned condition. These complications motivate our approach to identify optimal targets for fundraising activities.

Our approach to optimal targeting relies on two ingredients. First, it exploits random variation in the assignment of an unconditional gift at the donor level.[4] To induce this variation, we teamed up with a charity that sends out solicitation letters to its donor base once a year. In this setting, we randomly assigned almost 20,000 potential donors to a gift treatment (in which individuals received the solicitation letter with a gift) and a control group (in which they received only the letter). Second, as previously highlighted, our approach relies on machine-learning algorithms. Particularly, we let an algorithm learn the relationship between the individuals' observable characteristics and their expected donation behaviors in the experiment's treatment and control groups (i.e., with and without the gift). From this relationship, we then estimate the (out-of-sample) set of predicted net donors who should be targeted.[5] This procedure establishes what we label the *estimated optimal targeting rule*. Importantly, although we cannot trace out the causal effect of the characteristics that drive the heterogeneity, the policy-relevant increase in net donations achieved by the targeting rule is causally identified.

More specifically, our paper draws on the following machine-learning methods. We implement the optimal-policy-learning algorithm of Athey and Wager (2021), which extends the empirical welfare-maximization approach of Kitagawa and Tetenov (2018) by

[4] In principle, researchers could apply similar targeting methods to observational data.

[5] A benefit of the machine-learning approach is that it does not require a preanalysis plan. Cross-fitting techniques counteract the dissemination of spurious results and ex post "p-hacking."

machine learning. In our main specifications, we then estimate optimal targeting rules with the Exact Policy-Learning Tree of Zhou *et al.* (2018) and show the sensitivity of our results to various alternative estimators (Logit, Logit Lasso, Classification and Regression Trees, and Classification Forest). Regarding data, we feed our algorithm with information from various sources and of different types, including socioeconomic characteristics, past donation data, and geospatial information. The geospatial information consists of publicly available information from Google Maps on economic and cultural facilities close to the potential donor's place of residence.

Our analysis yields two sets of main results. The first set concerns the warm list (i.e., the sample of previous givers). For this sample, we demonstrate that machine-learning-based optimal targeting substantially boosts the charity's net donations compared to benchmarks in which everybody receives the gift (increase relative to this benchmark: 13.8%) or no one receives the gift (increase: 14.3%). We also show that these positive effects on net donations do not merely reflect pull-forward effects (i.e., shifts of donations from a later year to the experimental year). Moreover, we highlight that our fundraiser can reap the full benefits of machine-learning-based optimal targeting by relying on data that should be easily accessible by all charities. For the gains of our approach to materialize, knowledge about who is in the warm list and publicly available geospatial information is sufficient. This result suggests that our approach should be widely applicable. The second set of results concerns the cold list (i.e., the sample of previous nondonors). In contrast to the warm list, machine-learning-based optimal targeting in the cold list does not increase net donations compared to the no-gift benchmark. Furthermore, the estimated optimal targeting rule does not broaden the donor base enough to justify the gift's additional fundraising costs. We conclude that, in our context, the charity should not target the gift to cold-list individuals at all. Whereas the details of our findings may well be specific to the setting, a general conclusion is that charities that do not optimally target their fundraising efforts waste significant resources.

The paper is organized as follows. Section 2 outlines our contributions to the literature, Section 3 explains the institutional background and design of the experiment, and Section 4 describes our empirical strategy. Section 5 discusses our results, and Section 6 concludes. Online Appendices A–G provide supplementary materials.

2 Contributions to the Literature

In the following, we detail how our study relates and contributes to various literature strands in economics and marketing.

Economics literature. The first relevant strand of literature, the theoretical fundraising literature in public economics, provides a theoretical underpinning for why fundraising targeting can be beneficial. The argument is as follows. Many charities have properties similar to privately provided public goods (Andreoni and Payne, 2013): contributions are voluntary and the provided goods are nonexcludable and nonrivalrous. In such a context, fundraising tools can counteract free-riding and, hence, underprovision problems (see e.g., Andreoni, 1988; Morgan, 2000; Vesterlund, 2003; Andreoni and Payne, 2003, 2013).[6] However, if fundraising is costly and charities must compete for donors, competition can push the costs to such high levels that the total service provision falls (Rose-Ackerman, 1982; Aldashev and Verdier, 2010; Aldashev *et al.*, 2014). Targeting of fundraising instruments can then be a tool to maximize donations net of fundraising costs and, thus, provision levels. Specifically, Name-Correa and Yildirim (2013) show that a charity's optimal strategy is to target net donors. However, for charities, it is challenging to follow this theoretical rule, as they cannot easily identify these optimal targets. Along these lines, our contribution to this literature is to explore how a data-driven machine-learning approach can help charities predict the set of net donors.

A second emerging literature strand closely related to our work empirically studies how to target fundraising among heterogeneous donors. We are aware of only two papers that examine this topic. First, Adena and Huck (2019) apply a targeting strategy to *matching gifts,* a fundraising tool where funds collected before a campaign top up donations above a threshold. Their main result is that charities can crowd in donations by conditioning these thresholds on past giving behavior (i.e., the thresholds are targeted). Second, Drouvelis and Marx (2021) focus on belief-based targeting. They conclude that charities can increase net donations by targeting information treatments to potential donors who hold incorrect (low) beliefs about others' donations. Our paper differs from these studies in two dimensions: First, Adena and Huck (2019) and Drouvelis and Marx (2021) both build on conceptual considerations to identify dimensions of heterogeneity used for targeting. Following Athey and Wager (2021), our paper takes a more agnostic and more flexible, data-driven approach. Particularly, our machine-learning algorithms independently identify the most influential dimensions of heterogeneity based on observable donor characteristics. Consequently, the resulting estimated targeting rules can flexibly account for many different sources of heterogeneity (e.g., preferences or income), as long as they correlate with individuals' observable characteristics. Second, instead of considering debiasing or threshold matching, we focus on the unconditional gift as a different fundraising tool.

A third relevant literature strand is the literature on fundraising gifts (see, e.g., the review of List and Price, 2012). While Falk (2007) documents that unconditional gifts are a cost-effective tool to increase donations in the warm list, other studies paint a more

[6] Which tool is suited to increase provision depends on the context. Solicitation letters (perhaps providing return envelopes) counteract the underprovision problem in settings with transaction costs (Andreoni and Payne, 2003). By contrast, leadership gifts oppose the underprovision of threshold public goods (Andreoni, 1988) and public goods under imperfect information (Vesterlund, 2003). Lotteries, in contrast, can increase efficiency in standard public goods settings (Morgan, 2000).

scattered picture. For example, Landry *et al.* (2010) find zero effects of gifts in the warm list, and Alpizar *et al.* (2008) conclude that conditional on giving, gifts even lower contributions. Yin *et al.* (2020) present similar results.[7] We add to the looming discussion on the causes of effect heterogeneity by showing that individual heterogeneity alone is powerful enough to account for gifts' negative and positive effects.[8] In particular, in our setting, some groups of potential donors increase donations in response to gifts, and other groups reduce their donations. This finding persists if we restrict our sample to the warm list only. While these findings are insightful in themselves, our main contribution to this literature is to demonstrate that charities can effectively exploit the effect heterogeneities to target gifts optimally.

By focusing on effect heterogeneity, we contribute to a fourth literature strand, heterogeneous responses to fundraising that identifies five forms of heterogeneity: (a) characteristics of the charitable organization and the purpose of the charity (e.g., Okten and Weisbrod, 2000; de Vries *et al.*, 2015), (b) characteristics of the donors (e.g., Andreoni *et al.*, 2003; Andreoni and Vesterlund, 2001; Rajan *et al.*, 2009; Wiepking and James, 2013), (c) donation motives or preferences of donors (e.g., Bakshy *et al.*, 2012; Harbaugh *et al.*, 2007; Kizilcec *et al.*, 2018), (d) past donation behavior (Schlegelmilch and Diamantopoulos, 1997; Hassell and Monson, 2014), and (e) crowding out (Meer, 2017). Instead of studying single, selected dimensions of heterogeneity, we combine a range of individual characteristics and past donation behavior. Additionally, our paper adds a rarely used determinant of heterogeneity to the analysis, geospatial characteristics. As this information is easily accessible to charities, it is particularly beneficial for optimal targeting.[9]

Marketing literature. Our paper also relates to two strands of literature on the targeting of marketing interventions. The first strand applies machine-learning techniques to target marketing in contexts other than charitable giving. For example, Guelman *et al.* (2015) and Ascarza (2018) study targeting of retention efforts to lower customer churn in telecommunications, professional memberships, and insurance. Moreover, Fong *et al.* (2019), Ellickson *et al.* (2020), Gubela *et al.* (2020), and Smith *et al.* (2021) study targeting of recommendations, promotions, coupons, and pricing (mainly in online marketing). These studies not only consider contexts very different from ours, but also employ other methods. Particularly, the papers estimate heterogeneous effects of (randomized)

[7] Thank-you gifts that charities hand out after a donation also seem to lower subsequent donations (Newman and Shen, 2012). Eckel *et al.* (2018) directly compare unconditional and thank-you gifts and show that donors are twice as likely to give when they receive a high-quality unconditional gift.

[8] Of course, differences in the overall setting (such as varying causes of charities) might also explain why different studies come to varying conclusions.

[9] Dong *et al.* (2019) and Glaeser *et al.* (2018, 2020) show that geospatial characteristics are good proxies for income and socioeconomic characteristics. One reason is neighborhood segregation (see, e.g., Heblich *et al.*, 2020).

marketing interventions and transform them into a targeting rule by discretizing.[10] Our paper, instead, follows Athey and Wager (2021) and estimates the optimal targeting rule directly (instead of first estimating the effect heterogeneity). To the best of our knowledge, our paper is the first application of this novel, more direct method, explicitly tailored for optimal targeting.[11] The second relevant literature strand is the literature on *charitable marketing* (see, e.g., Winterich *et al.*, 2013; Kizilcec *et al.*, 2018). While this literature has studied targeting in charitable giving, it so far has not used the powerful machine-learning-targeting toolkit. To sum up, the available studies deviate from our paper either because they use machine learning in contexts other than charitable giving or because they study targeting in charitable giving without using causal machine-learning techniques.[12]

Methodological literature. Methodologically, we contribute to the small but rapidly growing literature that applies machine-learning methods to target public and private policies (e.g., Andini *et al.*, 2018; Hitsch and Misra, 2018; Kang *et al.*, 2013; Knaus *et al.*, 2020; Knittel and Stolper, 2019; Rockoff *et al.*, 2011; Kleinberg *et al.*, 2015). While these papers consider such contexts as taxation and labor-market programs, our study is the first that applies a fully fledged optimal policy learning algorithm like that of Athey and Wager (2021) to the context of charitable giving. One of our goals is to provide an intuitive introduction of the used methods to guide their application.

3 Experimental Design and Data

Our approach to derive a machine-learning-based optimal targeting rule proceeds in three steps. First, we conduct a field experiment that randomly allocates our fundraising instrument, an unconditional gift. Second, we use machine-learning algorithms to estimate the optimal targeting rule for this gift in a random subsample of the experimental data while retaining the remaining sample. Third, we extrapolate the estimated optimal targeting rule to the retained sample and apply off-policy-learning techniques to assess the estimated rule's out-of-sample performance. While this section details the experimental design and data, Section 4 outlines the machine-learning and off-policy learning approaches.

[10] The marketing literature labels this procedure "uplift modeling." Economists call similar methods "effect-based targeting."

[11] Related machine-learning-based literature profiles or segments customers solely based on their responses under one condition. For example, Cui *et al.* (2006) and Kim *et al.* (2005) show that neural networks are valuable tools to improve targeting in response models. Moreover, Abe *et al.* (2004) successfully apply reinforcement learning to cross-channel marketing, and Schwartz *et al.* (2017) apply a multiarmed bandit to target display advertising.

[12] See Simester *et al.* (2020) for a discussion of common data challenges of applying machine learning to target customers "in the wild."

3.1 The Natural Field Experiment

In 2014, we implemented a natural field experiment in collaboration with a fundraiser of the Catholic Church that operates in a German urban area.[13] For decades, this fundraiser has organized a large-scale, annual fundraising campaign: Once a year, it has mailed solicitation letters to all resident church members, irrespective of previous donations. This fund drive aims to finance local church-related projects, such as the renovation of clergy houses, parish centers, or churches. Our experiment exploited this campaign by (a) experimentally altering how the fundraiser contacted potential donors in 2014 and (b) analyzing individuals' behavior in 2014 and 2015.

Control group. Individuals in our experiment's control group received the standard solicitation letter, the contents of which remained unchanged from the pre-experiment years. Particularly, the letter highlighted the fundraiser's cause and asked recipients for a donation. To lower transaction costs, the fundraiser distributed the solicitation letter together with a remittance slip, prefilled with the fundraiser's bank account and the donor's name. In the pre-experiment years, potential donors received identical transaction forms.[14]

Gift treatment. Our design of the gift treatment closely follows Falk (2007). In 2014, individuals in this treatment received the solicitation letter together with an unconditional gift. The gift consisted of three envelopes paired with different folded cards, picturing Albrecht Dürer's "immaculate flower studies" (see Figure 1). Further, we added one sentence to the solicitation letter, stating that the fundraiser "would like to provide the included folded cards as a gift." The total per-unit cost for mailing the control-group solicitation letter amounted to 0.43 euro (printing plus postage). In the gift treatment, the per-unit cost increased by 1.16 euro (postcards and envelopes: 0.47 euro; boxing and additional postage: 0.69 euro). Notably, our treatment consisted of a one-time intervention. From 2015 onward, all individuals in the sample received a solicitation letter very similar to the one distributed in the pre-experiment years.

Sample. In 2014, 26% of the urban area's population were members of the Catholic Church. We drew a sample from this population consisting of 2,354 warm-list individuals (individuals who had donated at least once before the experiment) and 17,425 cold-list individuals (individuals who had never donated before). These individuals were then randomly allocated to the control and treatment groups, exploiting a stratified randomization scheme. Particularly, we assigned 1,180 of the warm-list individuals to the gift

[13] In 2015, we implemented a second experiment with two treatments, an unconditional gift treatment and a gift treatment that framed the gift as a reward for past donations. We evaluate and compare the effects of these differently framed gifts in a companion paper.

[14] For years, donations in the context of the fund drive could be made exclusively via bank transfer.

Figure 1: The gift consisting of the three folded cards and envelopes

Notes: The gifts consisted of three different folded cards showing flower motifs from paintings of Albrecht Dürer plus three envelopes.

treatment and 1,174 to the control group.[15] By contrast, 2,283 cold-list individuals received the treatment, while 15,142 were part of the control group.

The setting's benefits. Our setting serves as a suitable testing ground for machine-learning-based optimal targeting. First, as the fundraiser did not employ any targeting strategies before the experiment, the setting offers a clean environment to study our machine-learning approach's potential. Second, it provides rich data that not only allow us to estimate powerful targeting rules, but also enable us to test which type of data are especially beneficial for machine-learning-based optimal targeting (see the following description of the data). Third, because the fundraiser contacts all church members exhaustively, the setting offers the possibility to study the cold and warm lists separately. We, hence, can not only examine the optimal targeting of gifts among past donors, but also explore whom to target in the process of acquiring new donors. Fourth, because we were able to gather data for two postexperiment years, the setting allows us to study if the estimated optimal targeting rule increases total donations or simply pulls forward donations from 2015 to 2014. Fifth, because religious giving dominates the charitable giving landscape (List, 2011), targeting is particularly relevant in this context.[16]

[15]The strata were defined based on list (warm vs. cold), gender, household type indicators, quintiles of individuals' predicted baseline willingness to give, and quintiles of age. To construct a proxy for the baseline willingness to give in the treatment year, we first regressed an indicator variable for giving in the year before the experiment on indicator variables for further lags of the giving indicator. We then used the estimated model to predict the probability of giving in the treatment year (out of sample).

[16]For example, in Germany, church-related causes benefit the most from private giving: They receive approximately 35% of total private giving (Deutscher Spendenrat, 2016). No other type of cause benefits from a similarly high share of total donations. The numbers for the United States are very similar (Andreoni and Payne, 2013).

3.2 The Data

Data sources. Our study draws on two separate, comprehensive data sets. The first set of data includes administrative records provided by the Catholic Church. The records hold a number of socioeconomic characteristics, such as gender, marital status, and age. Furthermore, they contain individual-specific information on donations for the years 2006–2015. Accordingly, we observe all potential donors' donation histories for eight pre-experiment years and their donations in the first two years after the experiment. Our second data source is Google Maps. Specifically, we used the Google Maps API to collect geospatial information on economic and cultural facilities near each individual's residence. We then merged this data with the administrative records based on postal addresses. In particular, we collected the number of restaurants, supermarkets, medical facilities, cultural facilities, and churches within 300 meters of each the home address. We also web-scraped the distance from the home address to the central train station, city hall, main church, and airport. Furthermore, we retrieved the elevation of the home address. The main reason for using these geospatial characteristics is that they are readily available to charities and are powerful proxies for income and other socioeconomic characteristics (Dong *et al.*, 2019; Glaeser *et al.*, 2018, 2020) that likely explain response heterogeneity. Taken together, our algorithms for optimal targeting rely on three types of input data: (a) socioeconomic information, (b) information on past donation behavior, and (c) publicly available geospatial information. Note that charities that manage to collect even more comprehensive datasets could further improve machine-learning-based optimal targeting.

Descriptive statistics. Table 1 reports the descriptive statistics for the donation amount and the donation probability. Furthermore, supplementary tables either study the balance of observable characteristics across the warm and cold lists (see Table A.1 in Online Appendix A) or the control and treatment groups (see Tables A.2 and A.3). Several distinctive features of the data stand out. First, unsurprisingly, cold-list individuals donated much less in the first year after the experiment (average donation: 0.18 euro) compared to warm-list individuals (average: 16.02 euro). Their donation probability is also much lower. Similar results emerge when summing up the donations made in the first two years after the experiment. Second, donations are highly right-skewed and have excess kurtosis, highlighting that few donors give extraordinarily large gifts. The following analysis highlights that, despite this data feature complicating our prediction task, we are nevertheless able to estimate effective optimal targeting rules. Third, cold list and warm list individuals differ in socioeconomic characteristics (see Table A.1).[17]

[17] On average, cold list individuals are younger, have a higher likelihood of being single, and tend to have a shorter residency duration in the urban area. Individuals in the warm list donated an average of four times, with a total donation amount of 126 euro over the eight years before the experiment. By construction, cold list individuals have a donation history of zero donations. Individuals in the cold list

Table 1: Descriptive statistics of donation amount and donation probability

	Mean	Std. dev.	Skewn.	Kurt.	Min.	Max.
	(1)	(2)	(3)	(4	(5)	(6)
Panel A: Warm list						
1st year after the experiment						
A1. Donation amount (euro)	16.02	30.38	4.70	39.05	0	450
A2. Donation dummy	0.49				0	1
1st and 2nd year after the experiment						
A3. Donation amount (euro)	30.48	53.19	4.96	49.23	0	900
A4. Donation dummy	0.57				0	1
Panel B: Cold list						
1st year after the experiment						
B1. Donation amount (euro)	0.18	3.00	38.39	2,049.4	0	200
B2. Donation dummy	0.009				0	1
1st and 2nd year after the experiment						
B3. Donation amount (euro)	0.43	4.85	23.12	779.6	0	240
B4. Donation dummy	0.017				0	1

Notes: This table shows descriptive statistics for our primary outcomes, (a) the donation amount and (b) variables indicating whether a person donated or not. We consider the warm list (Panel A) and cold list (Panel B) separately. Further, we track our outcomes over two periods, the first year after and the first two years after the experiment. For dummy variables, the first moment is sufficient to infer the entire distribution.

This observation suggests that individual characteristics might explain heterogeneous giving behavior. Fourth, the observable characteristics are very well balanced across the treatment and control groups (see Tables A.2 and A.3).

4 Empirical Strategy

This section describes the estimation and identification strategy of machine-learning-based optimal targeting rules and how to assess the rules' out-of-sample performance. We start by introducing conditional average treatment effects (CATEs) in Subsection 4.1. The CATEs formally describe heterogeneous treatment effects as a function of observable characteristics. In Subsection 4.2, we then introduce binary optimal targeting rules that are related to the continuous CATEs. These rules maximize the charity's net donations by assigning individuals to either the targeted group or the untargeted group. Recall that our approach directly estimates the optimal targeting rule. This feature distinguishes it from approaches that first estimate the continuous CATEs and then derive the rules as

live, on average, closer to the city center (closer to city hall and the central station) than individuals on the warm list. Close to the home address (within 300 meters), cold-list individuals have, on average, more access to restaurants, supermarkets, medical and cultural facilities, and churches than warm-list individuals.

nonlinear transformations of these effects. In the final step, we discuss our classification approach in Subsection 4.3. In Subsection 4.3, we further discuss how we measure our targeting rules' out-of-sample performance compared to benchmark rules.

4.1 Conditional Average Treatment Effects

Notation. We use the potential-outcome framework (Rubin, 1974) to describe the parameters of interest. The potential-outcome framework is useful when studying targeting because it allows us to describe an individual's reaction under different (counterfactual) treatment conditions. The treatment variable D_i indicates whether a fundraising gift was sent to individual i (for $i = 1, \dots, N$), with

$$D_i = \begin{cases} 1 & \text{when a gift was sent, and} \\ -1 & \text{otherwise.} \end{cases}$$

$Y_i(1)$ denotes the potential donations in response to the solicitation letter with a fundraising gift. $Y_i(-1)$ denotes the potential donations in response to the letter unaccompanied by fundraising gifts.

Causal effects. Using the previous notation, the individual causal effects are

$$\delta_i = Y_i(1) - Y_i(-1).$$

In an ideal world, the charity would know δ_i. It could then (exclusively) assign the gift to individuals for whom δ_i exceeds the gift's cost. However, because $Y_i(1)$ and $Y_i(-1)$ cannot be observed simultaneously, the fundamental problem of causal analysis is that δ_i is unobservable. Nevertheless, it is possible to identify and estimate group averages of δ_i. For example, the average treatment effect (ATE), $\delta = E[\delta_i] = E[Y_i(1) - Y_i(-1)]$, is the expected average effect of the gift on donations. Moreover, there might be effect heterogeneity with regard to observable characteristics X_i, which allows researchers to identify even finer-grained subgroup-specific effects. For example, Andreoni and Vesterlund (2001) and Andreoni *et al.* (2003) show that men and women differ in their donation behavior. In this vein, the CATE describes the influence of an individual's characteristics x on the expected average effect of sending the fundraising gift to the individual:

$$\delta(x) = E[\delta_i | X_i = x] = E[Y_i(1) - Y_i(-1) | X_i = x].$$

Identification of causal effects. The ATEs and CATEs are identified from observable data under the stratified experimental design and the stable unit treatment value assumption (SUTVA) (see proof in Appendix B):

$$Y_i = Y_i(-1) + \frac{1+D_i}{2}\left(Y_i(1) - Y_i(-1)\right),$$

The strata characteristics, which we call Z_i, are relevant to achieve identification. In contrast, the exogenous characteristics, X_i, are potentially associated with heterogeneous effects of the gift, but we do not need them for identification.

4.2 Targeting Rules

Optimal targeting rule. A targeting rule, $\pi(X_i) \in \{-1, 1\}$, is a deterministic function that assigns the gift to prospective donors based on their observable characteristics, X_i. Under the rule, individuals with $\pi(X_i) = 1$ receive the solicitation letter with the gift, and individuals with $\pi(X_i) = -1$ receive the solicitation letter without the gift. The purpose of the optimal targeting rule is to maximize the expected net donation $P(\cdot)$ of the fundraising campaign, defined as the expected donation minus the gift's variable costs. Formally, the expected net donation is

$$P(\pi(X_i)) := E\left[Y_i(\pi(X_i)) - \frac{1+\pi(X_i)}{2}c\right], \tag{1}$$

where $Y_i(\pi(X_i))$ is the donation amount of individual i under the rule $\pi(X_i)$ and c are the variable costs of the gift. We ignore fixed costs, as they do not alter the targeting rule.

Benchmarks rules. To evaluate the gains of optimal targeting, we compare the expected net donations under the optimized rule $P(\pi(X_i))$ to the expected net donations under three benchmarks: a rule that assigns the gift to everybody (all-gift benchmark), a rule that assigns the gift to no one (no-gift benchmark), and a rule with random allocation (random-gift benchmark). First, we can contrast the optimal targeting rule to the all-gift benchmark $\pi(X_i) = \pi_1 = 1$. For this benchmark, the expected net donation is $P(\pi_1) = E[Y_i(1)] - c$. Consequently, the excess net donation of the optimal targeting rule compared to this benchmark is

$$Q_1(\pi(X_i)) := P(\pi(X_i)) - P(\pi_1) = E\left[\frac{\pi(X_i) - 1}{2}(\delta_i - c)\right].$$

Second, equivalently, we compare the optimal rule to the no-gift benchmark $\pi(X_i) = \pi_{-1} = -1$, under which the expected donations are $P(\pi_{-1}) = E[Y_i(-1)]$. Thus, relative to this benchmark, optimal targeting increases net donations by

$$Q_{-1}(\pi(X_i)) := P(\pi(X_i)) - P(\pi_{-1}) = E\left[\frac{1+\pi(X_i)}{2}(\delta_i - c)\right].$$

Third, we consider the random-gift benchmark, π_R, under which each individual has a 50% probability of receiving the gift. Given that π_R triggers the expected net donation of $P(\pi_R) = 1/2 \cdot (E[Y_i(1) + Y_i(-1)] - c)$, the excess net donation of the optimal rule is

$$\begin{aligned} Q_R(\pi(X_i)) := & \quad P(\pi(X_i)) - P(\pi_R) = \frac{1}{2} E\left[\pi(X_i)(\delta_i - c)\right] \\ = & \quad [Q_1(\pi(X_i)) + Q_{-1}(\pi(X_i))]/2. \end{aligned} \tag{2}$$

The random rule can be viewed as a default option when no information about the effectiveness of the fundraising instrument is available, and the fundraiser has no preferences about the allocation of the instrument.

Two further points are of note. First, the optimal targeting rule which maximizes the net donations $P(\cdot)$, also maximizes $Q_1(\cdot)$, $Q_{-1}(\cdot)$, and $Q_R(\cdot)$. The reason is that $P(\pi_1)$, $P(\pi_{-1})$, and $P(\pi_R)$ are constant. Second, for the estimation of the optimal targeting rule, we maximize the sample analog of (2). Because we do not observe the individual causal effects, δ_i, which we need to determine (2), we first discuss how to approximate these parameters.

4.3 Estimation

To introduce our estimation strategy of the optimal targeting rule, we proceed in two steps. In the first step, we discuss augmented inverse probability weighting (AIPW) to estimate an approximation of the individual causal effects. In the second step, we show how to use the AIPW score to estimate the optimal targeting rule.

Augmented inverse probability weighting. An essential ingredient for the optimal targeting rule is δ_i. As we mentioned before, δ_i is unobservable and cannot be estimated directly. However, an approximation score of δ_i can be sufficient to estimate the optimal targeting rule. The AIPW score,

$$\Gamma_i = \mu_1(Z_i) - \mu_{-1}(Z_i) + \frac{1+D_i}{2} \cdot \frac{Y_i - \mu_1(Z_i)}{p(Z_i)} + \frac{D_i - 1}{2} \cdot \frac{Y_i - \mu_{-1}(Z_i)}{1 - p(Z_i)},$$

is an example of such an approximation score.[18] The so-called nuisance parameters are the conditional expectations of the donations, $\mu_1(z) = E[Y_i|D_i = 1, Z_i = z]$ and $\mu_{-1}(z) = E[Y_i|D_i = -1, Z_i = z]$, and the conditional probability that the gift was sent $p(z) = Pr(D_i = 1|Z_i = z)$. The latter is often called the propensity score. Under the SUTVA and the experimental design, the expected value of the AIPW score identifies the ATE $\delta = E[\Gamma_i]$.[19] The conditional expectations of the AIPW score identify the CATEs,

[18]Alternatively, Kitagawa and Tetenov (2018) suggest inverse probability weighting scores, and Beygelzimer and Langford (2009) propose offset weighting scores.

[19]Note that the nuisance parameters, $\mu_1(z) = E[Y_i|D_i = 1, Z_i = z] = E[Y_i(1)|Z_i = z]$ and $\mu_{-1}(z) = E[Y_i|D_i = -1, Z_i = z] = E[Y_i(-1)|Z_i = z]$, equal conditional expectations of the potential donations with

$\delta(x) = E[\Gamma_i | X_i = x]$. For completeness, we sketch the identification proofs for the AIPW score in Online Appendix C (see Knaus *et al.*, 2021, for a detailed discussion).

We can estimate the AIPW score as follows. First, we estimate the nuisance parameters. In this step, we obtain the estimated conditional expectations of the potential donations with and without the gift by $\hat{\mu}_1(z)$ and $\hat{\mu}_{-1}(z)$ and the estimated propensity score by $\hat{p}(z)$. Second, we plug the estimated nuisance parameters into the estimator of the AIPW score,

$$\hat{\Gamma}_i = \hat{\Gamma}_i(1) - \hat{\Gamma}_i(-1),$$

with

$$\hat{\Gamma}_i(1) = \hat{\mu}_1(Z_i) + \frac{1 + D_i}{2} \cdot \frac{Y_i - \hat{\mu}_1(Z_i)}{\hat{p}(Z_i)},$$

and

$$\hat{\Gamma}_i(-1) = \hat{\mu}_{-1}(Z_i) - \frac{D_i - 1}{2} \cdot \frac{Y_i - \hat{\mu}_{-1}(Z_i)}{1 - \hat{p}(Z_i)}.$$

The corresponding average-treatment-effect estimator

$$\hat{\delta} = \frac{1}{N} \sum_{i=1}^{N} \hat{\Gamma}_i \tag{3}$$

is consistent, asymptotically normal, and semiparametrically efficient under the requirement that the nuisance parameter estimators are consistent and converge sufficiently fast (e.g., Chernozhukov *et al.*, 2017; Robins *et al.*, 1994). In our application, we have precise information about the stratification process. Therefore, we use parametric nuisance parameter estimators which satisfy the requirements. In particular, we use a Logit to estimate the propensity score and OLS to estimate the conditional expectations of the potential donations with and without the gift (Table D.1 in Online Appendix D reports the estimated coefficients of the different models). As for the CATE estimator, Semenova and Chernozhukov (2020), Fan *et al.* (2019), and Zimmert and Lechner (2019) show that the AIPW scores can also be used to estimate the CATEs, $\hat{\delta}(x)$. This requires additional restrictions on the parameter space of X_i (see, e.g., Knaus, 2020, for a comprehensive review).

Note that we could alternatively estimate the ATEs and CATEs with a multivariate OLS regression model. To obtain the CATEs, we would have to interact the treatment dummy, D_i, with the characteristics, X_i. However, in contrast to multivariate OLS, the AIPW does not impose any restrictions on the (variable-specific) effect heterogeneity, which is particularly relevant for the targeting approach. Having said this, we show in Section 5 that the OLS and AIPW estimates of the ATEs are similar.

and without the gift under the SUTVA and the experimental design.

Estimating the optimal targeting rule. For the estimation of the optimal targeting rule, Athey and Wager (2021) propose replacing the unobservable individual causal effect, δ_i, in (2) with the estimated AIPW score, $\hat{\Gamma}_i$:

$$\pi^* = \text{argmax}_\pi \left\{ \frac{1}{2N} \sum_{i=1}^{N} \pi(X_i)(\hat{\Gamma}_i - c) \right\}. \tag{4}$$

Alternatively, the objective function (4) can be formulated as the weighted classification estimator

$$\pi^* = \text{argmax}_\pi \left\{ \frac{1}{2N} \sum_{i=1}^{N} \pi(X_i) \cdot \text{sign}(\hat{\Gamma}_i - c) \cdot |\hat{\Gamma}_i - c| \right\}, \tag{5}$$

where $(\hat{\Gamma}_i - c) = \text{sign}(\hat{\Gamma}_i - c) \cdot |\hat{\Gamma}_i - c|$ (see, e.g., Beygelzimer and Langford, 2009; Zadrozny, 2003; Zhao *et al.*, 2012). This estimator aims to classify the sign of the net donation effects and weigh each observation by $|\hat{\Gamma}_i - c|$. The objective function is maximized when the signs of $\pi(X_i)$ and $(\hat{\Gamma}_i - c)$ are equal. If some signs differ, misclassifications of individuals who respond strongly (i.e., individuals with large weights) reduce the net donations more than misclassification of individuals who do not respond strongly to the gift (i.e., individuals with small weights). Accordingly, the optimal targeting estimator should prioritize individuals with large weights. Because the estimated optimal targeting rule is a deterministic function of the observable characteristics X_i, there is an implicit connection between optimal targeting and the CATEs, even though we estimate the optimal targeting rule directly (without estimating the CATEs first).[20]

The main result of Athey and Wager (2021) that enables estimation of (5) is that, when the complexity of the estimator is restricted, the optimal targeting rule π^* achieves asymptotically minimax-optimal regret (Manski, 2004). Along these lines, in principle, any restricted weighted classification estimator could be used to estimate (5). We, however, follow Athey and Wager (2021) and use shallow decision trees to estimate the optimal rule. Trees partition the sample into mutually exclusive strata based on the heterogeneity characteristics, X_i. Furthermore, the tree depth restricts the complexity of the estimated optimal targeting rule, which makes trees suitable estimators in our context. In our main specifications, we follow Zhou *et al.* (2018) and use Exact Policy-Learning Trees, with a search depth of two, to estimate the optimal targeting rule.[21]

Advantages of decision trees. It is possible to use standard estimators, such as a weighted Logit regression, to estimate (5). However, decision trees have several ad-

[20] Furthermore, the CATEs do not account for the gift's costs.

[21] For implementation, we use the `R` package `policytree` (Sverdrup *et al.*, 2020). Alternatively, we could have used classification and regression trees (CARTs). CARTs select the partition with a greedy algorithm by adding recursive sample splits to the tree without anticipating later splits (e.g., Breiman *et al.*, 1984). In contrast, Exact Policy-Learning Trees search for a fixed tree depth over all possible targeting rules. In contrast to CARTs, Exact Policy-Learning Trees estimate the global optimum of (5).

vantages compared to Logit regressions for the estimation of optimal targeting rules. They select the relevant heterogeneity characteristics in a data-driven way. This feature is particularly useful when we have no *a priori* domain knowledge about the relevant characteristics. Even if we know the relevant characteristics, there might be several highly correlated measures of these characteristics, and it may be *a priori* unclear which are the most relevant. For example, in our application, several geospatial characteristics are highly correlated, and there is little *a priori* guidance on which should be used. In the extreme case, including too many highly correlated characteristics in a Logit regression could cause multicollinearity problems. Furthermore, it is typically unclear how flexible the empirical model should be with regard to nonlinear and interaction terms. Trees can automatically incorporate nonlinear and interaction terms of the different characteristics without precoding. This feature minimizes the risk of overlooking important heterogeneities. Having said this, we study the sensitivity of our results to different estimation methods for the targeting rule in Section 5.6. In particular, we consider Logit, Logit Lasso, CART, and classification-forest estimators.

Estimating the gains of targeting. Once we have estimated the optimal targeting rule, π^*, we can apply the sample analogy principle to estimate the gains of targeting relative to the benchmarks:

$$\hat{P}(\pi^*(X_i)) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\Gamma}_i(\pi^*(X_i)) - \frac{1+\pi^*(X_i)}{2}c\right),$$

$$\hat{Q}_1(\pi^*(X_i)) = \frac{1}{N}\sum_{i=1}^{N}\frac{\pi^*(X_i)-1}{2}\left(\hat{\Gamma}_i - c\right),$$

$$\hat{Q}_{-1}(\pi^*(X_i)) = \frac{1}{N}\sum_{i=1}^{N}\frac{1+\pi^*(X_i)}{2}\left(\hat{\Gamma}_i - c\right), \text{ and}$$

$$\hat{Q}_R(\pi^*(X_i)) = \frac{1}{2N}\sum_{i=1}^{N}\pi^*(X_i)\left(\hat{\Gamma}_i - c\right).$$

These estimators are consistent, asymptotically normal, and semiparametrically efficient (see, Chernozhukov *et al.*, 2018a).

We estimate the gains of targeting using a cross-fitting procedure. Our procedure randomly partitions our data into $K = 20$ equally sized samples. We then use $K-1$ partitions to estimate the targeting rule π^* and calculate $\hat{P}(\pi^*(X_i))$, $\hat{Q}_1(\pi^*(X_i))$, $\hat{Q}_{-1}(\pi^*(X_i))$, and $\hat{Q}_R(\pi^*(X_i))$ in the retained partition. We repeat this procedure, discarding each of the K partitions once. In this way, we use the entire dataset efficiently. Finally, we report the average values of $\hat{P}(\pi^*(X_i))$, $\hat{Q}_1(\pi^*(X_i))$, $\hat{Q}_{-1}(\pi^*(X_i))$, and $\hat{Q}_R(\pi^*(X_i))$ over all 20 partitions. The cross-fitting approach allows us to assess the estimated targeting rules' out-of-sample performance. It also addresses the concern that the targeting rule reflects spurious relationships and overstates the success of targeting due to overfitting.

Table 2: Average treatment effects of the gift on donations

	Warm list			Cold list		
	OLS	OLS	AIPW	OLS	OLS	AIPW
	(1)	(2)	(3)	(4)	(5)	(6)
A. Average treatment effects	1.24	1.21	1.22	0.19***	0.19***	0.19*
	(1.25)	(1.16)	(1.15)	(0.07)	(0.07)	(0.10)
B. Average treatment effects net of costs	0.08	0.05	0.06	-0.97***	-0.97***	-0.97***
	(1.25)	(1.16)	(1.15)	(0.07)	(0.07)	(0.10)
Strata controls	No	Yes	Yes	No	Yes	Yes

Notes: This table shows the estimated ATEs of the gift treatment on donations. The first set of estimates uses the amount donated in the first year after the gift as an outcome variable (euro). The second set of estimates additionally subtracts the gift's cost from the donation amount. We report results for the following specifications: unconditional OLS (Columns 1 and 4), OLS with strata control variables (Columns 2 and 5), and AIPW (Columns 3 and 6). Because the AIPW model allows for heterogeneous treatment effects, this model represents our preferred specification. Standard errors are in parenthesis. ***/**/* indicate statistical significance at the 1%/5%/10% level.

5 Results

This section presents our results. Subsection 5.1 discusses the ATEs of the gift on donations, and Subsection 5.2 explores the heterogeneity of the effects. The section proceeds by discussing the effectiveness of our optimal targeting approach in Subsection 5.3 before describing several properties of the estimated optimal targeting rule in Subsection 5.4. Finally, Subsection 5.5 outlines which characteristics are sufficient to increase profits significantly through machine-learning-based optimal targeting, and Subsection 5.6 explores the robustness of our results to alternative estimators.

5.1 Average Effects of Gifts on Donations

To facilitate comparison to the literature studying the effects of unconditional gifts on donations, our first step is to estimate the ATE of the gift on donations. Table 2 reports three different estimates, focusing on behavior in the first year after the experiment: estimates from unconditional OLS regressions (Columns 1 and 4), estimates from conditional OLS regressions (Columns 2 and 5), and estimates from the previously introduced AIPW estimator (Columns 3 and 6).[22] Columns 1–3 cover the warm list and Columns 4–6 the cold list.

Average treatment effects for the warm list. Two observations characterize the responses in the warm list. First, the gift increased average donations by 1.21–1.24 euro, though the effects are not statistically significant (see Row A in Table 2).[23] Notably, these

[22] In contrast to the OLS regressions, the AIPW estimator relies on fewer functional-form assumptions, allows for heterogeneous treatment effects, and is more robust to misspecification.

[23] This result is in line with Landry *et al.* (2010), who also report insignificant effects for the warm list.

estimates do not account for the cost of the gift (which were 1.16 euro). Second, when accounting for the costs, we find small and insignificant net-of-cost effects between 0.05 euro and 0.08 euro, depending on the chosen estimator (see Row B in Table 2). The small values imply that we neither find evidence for the hypothesis that the gift treatment was, on average, profitable (i.e., increased average donations by more than the costs) nor that it resulted in a net loss for the fundraiser. The two observations are insightful from a targeting perspective. To see why, note that we can think of the treatment effects as driven by a change from the benchmark targeting rule where no one receives the gift (control group) to the one where everybody receives the gift (treatment group).[24] Along these lines, the insignificant net-of-cost effects speak against the hypothesis that the all-gift benchmark outperforms the no-gift benchmark in terms of available funds.

Average treatment effects for the cold list. Very different results emerge for the cold list. The ATEs are significant (note the larger sample size), but much smaller, and amount to just 0.19 euro (Row A). As a result, when accounting for costs, the all-gift benchmark rule would result in a significant loss of 0.97 euro per donor, compared to the no-gift benchmark rule (Row B). Accordingly, considering only these two benchmark rules, we find that the more profitable strategy is to send the gift to no one in the cold list.[25] Taken together, we conclude that naive targeting of gifts to all individuals in the warm and cold list does not increase our fundraiser's net donations. Such a strategy would even likely result in a net loss as there are many more cold-list than warm-list individuals. Based on these insights, one might be tempted to conclude that our fundraiser can never increase net donations with the gift. In the following, we, instead, show that a more targeted gift campaign can be very successful.

5.2 Heterogeneous Treatment Effects

This subsection provides a descriptive analysis suggesting that the charity can likely increase raised net donations by deviating from the no-gift and all-gift benchmarks. For that purpose, we explore effect heterogeneity. To motivate the analysis, note that in the absence of heterogeneous treatment effects, the charity cannot benefit from targeting rules that are more flexible than the benchmarks. Intuitively, as all individuals respond similarly to the gift, the optimal targeting rule would either correspond to the no-gift or all-gift benchmark. However, with effect heterogeneity, some individuals may increase their donations by more and others by less than the gift's cost. If this is the case, the charity's optimal strategy would be to deviate from the benchmarks and target a subset of individuals only.

[24] More formally, the expected difference in net donations between the benchmark rules π_1 and π_{-1} is $P(\pi_1) - P(\pi_{-1}) = E[\delta_i] - c$.

[25] The previous literature has reported similar results in the past. For example, Alpizar *et al.* (2008) show that gifts increase donations, but the increase is insufficient to cover the gift's costs.

Estimating effect heterogeneity. We use the CATE-based sorted-effects approach of Chernozhukov *et al.* (2018b) to study effect heterogeneity. This method allows us to visualize the distribution of the effects of the gift on donations while reporting confidence intervals that account for multiple testing. Particularly, we specify a linear OLS regression including all observable characteristics plus interactions with the treatment dummy. Using this model, we estimate the CATE for each individual and report the percentiles of the estimated CATEs (labeled *sorted effects*).[26] We then examine if the sorted-effects model shows heterogeneous effects below and above the gift's costs to investigate if deviations from the two benchmarks are likely beneficial. Note, however, that we only use the sorted-effects approach to assess the potential of optimal targeting descriptively (i.e., we do not use it to identify targets). Instead, we estimate the optimal targets using a machine-learning approach explicitly developed for this purpose.

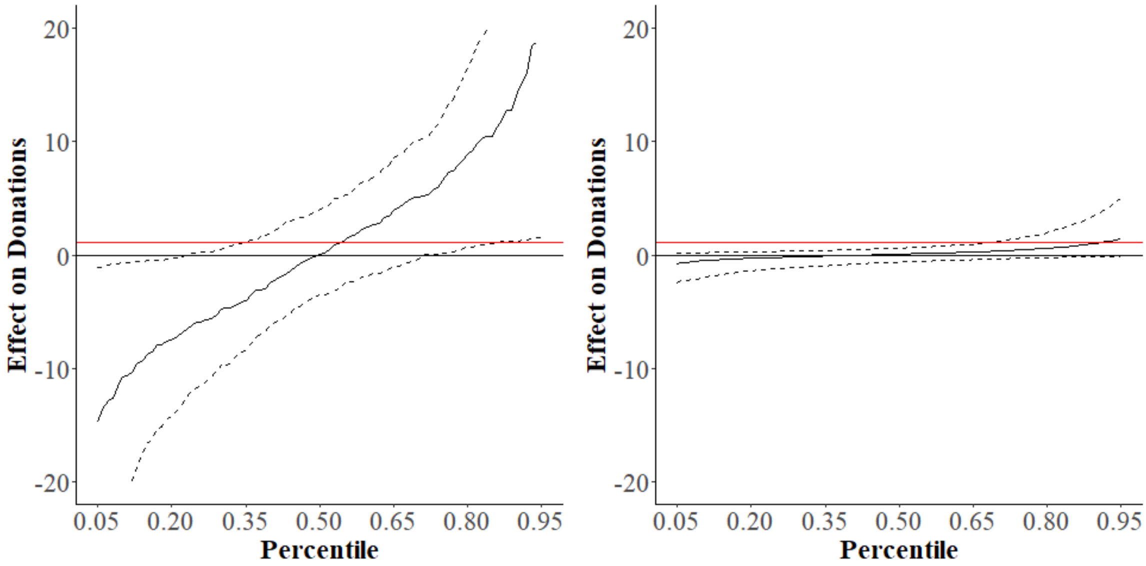
Heterogeneity in the warm list. Panel (a) in Figure 2 depicts the heterogeneity of the treatment effect on the donation amount for the warm list (solid line). It sorts the estimated CATEs by size and plots the size of the treatment effect in euro (vertical axis) against the percentiles of the effect size (horizontal axis). The red horizontal line represents the cost of the gift (1.16 euro).

The figure reveals substantial treatment-effect heterogeneity.[27] For some individuals, the treatment effects are positive, which is in line with the sequential-reciprocity hypothesis (Dufwenberg and Kirchsteiger, 2004; Falk, 2007). For example, the 5% most responsive individuals increase donations by more than 20.62 euro in response to the gift. By contrast, at the fifth percentile, donations decrease by 14.76 euro. Such an adverse effect points to the possibility that even a "warm" gift (folded cards) may change the donors' perception of the relationship with the fundraiser from a communal to an exchange norm (Yin *et al.*, 2020). The pronounced heterogeneity is interesting for at least two reasons. First, and most importantly, the heterogeneity indicates that machine-learning-based optimal targeting can be highly beneficial in the warm list: For only 62% of all individuals, the estimated effects exceed the cost of providing the gift. Thus, by targeting these individuals and not targeting donors with effects lower than the cost, the charity could substantially increase net donations. Second, the figure also reveals that the individual heterogeneity alone is powerful enough to account for the negative and positive effects reported in the literature.

[26] Statistical inference is based on a multiplier bootstrap (see Chernozhukov *et al.*, 2018b, for details).

[27] One might be interested in whether the size of the treatment effects correlates with observable characteristics. Tables E.1 and E.2 in Online Appendix E report the mean values of all characteristics for the groups with the 10% largest and the 10% smallest sorted effects. In the warm list, the individuals with the 10% largest effects tend to have donated less before the experiment and to live at a lower altitude than individuals with the 10% smallest effects, although the effects are insignificant. In the cold list, the individuals with the 10% largest effects tend to live significantly closer to the city center than individuals with the 10% smallest effects.

Figure 2: Sorted effects



(a) Warm list (b) Cold list

Notes: This figure shows the heterogeneity of the effect of the gift on the donation amount. To that end, it sorts the estimated conditional average treatment effects by size and plots the size of the treatment effect in euro (vertical axis) against the percentiles of the effect size (horizontal axis). The red horizontal line represents the cost of the gift (1.16 euro). The solid line depicts the sorted effects. We report results between the 5 and 95 percentiles. The dashed lines report uniformly valid 95% confidence intervals, which build on a multiplier bootstrap and 500 replications.

Heterogeneity in the cold list. Panel (b) in Figure 2 highlights that the effect heterogeneity in the cold list is much smaller than in the warm list. For example, the donation amount decreases by 0.75 euro at the fifth percentile and increases by 1.39 euro at the 95 percentile. Moreover, we find that the gift-induced increase in donations exceeds the costs only for individuals above the 92 percentile in the cold list, and even for these individuals the size of the treatment effects is relatively small. We conclude that the potential to increase net donations by subgroup-specific targeting is much lower in the cold list than in the warm list.

5.3 Effectiveness of Machine-Learning-Based Optimal Targeting

This subsection evaluates if machine-learning-based optimal targeting allows us to exploit the documented response heterogeneity to increase net donations in the first year after the experiment. We first estimate optimal targeting rules (Athey and Wager, 2021). We then evaluate the effectiveness of the estimated rules in raising net donations by comparing their out-of-sample performance to several benchmarks. For that purpose, we use the cross-fitting approach described in Section 4.

The estimated optimal targeting rule in the warm list. Table 3 focuses on the warm list and documents how the estimated optimal targeting rule performs out of sample

Table 3: Out-of-sample performance of targeting rule in the warm list

	Expected outcome value under optimal targeting	Optimal targeting vs. benchmarks		
		all-gift	no-gift	random-gift
	(1)	(2)	(3)	(4)
Panel A: Share of individuals that should receive the gift				
A1. Share treated	0.33			
Panel B: Results for primary outcome variable				
B1. Net donation amount	17.61***	2.14***	2.20***	2.17***
(1st year)	(0.97)	(0.82)	(0.81)	(0.58)
Panel C: Results for secondary outcome variables				
C1. Donation probability	0.503***	0.007	0.025**	0.016*
(1st year)	(0.013)	(0.013)	(0.010)	(0.008)
C2. Net donation amount	32.94***	2.33*	3.75***	3.04***
(1st and 2nd year)	(1.66)	(1.41)	(1.41)	(0.10)
C3. Donation probability	0.582***	0.001	0.017*	0.009
(1st and 2nd year)	(0.013)	(0.013)	(0.009)	(0.008)

Notes: This table documents the out-of-sample performance of our estimated optimal targeting rule, focusing on the warm list. The goal of optimal targeting is to maximize donations, net of costs. Panel A reports the share of individuals that, according to the rule, should receive the gift. Panel B reports the expected consequences of our rule for net donations as our main outcome. Panel C, instead, focuses on secondary outcomes. The columns can be interpreted as follows. Column 1 reports the expected value of the outcomes under optimal targeting. For example, we expect that, under optimal targeting, the donations, net of costs, would be 17.61 euro. Columns 2–4 show how optimal targeting changes the outcomes relative to three benchmark scenarios: everybody receives the gift (Column 2), no one receives the gift (Column 3), and the gift is randomly assigned to half of the sample (Column 4). Methodologically, the optimal targeting rules are estimated with Exact Policy-Learning Trees and a search depth of two (Zhou *et al.*, 2018). Donations are measured in euro. Standard errors are in parentheses. ***/**/* indicate statistical significance at the 1%/5%/10% level.

relative to the benchmarks. The table provides three sets of insights. First, Panel A reports that the estimated targeting rule recommends deviating from the no-gift and all-gift benchmarks. Specifically, it assigns the gift to 33% of the warm-list individuals (Column 1). This percentage is reasonably close to the percentage of individuals with a sorted effect that exceeds the gift's costs (see Figure 2).

Second, Panel B documents that, compared to our benchmarks, the charity would benefit substantially from applying the estimated optimal targeting rule. In the first year after the experiment, the average net donation under the estimated optimal targeting rule is 17.61 euro (Column 2). This value implies that, under the estimated optimal targeting rule, the average donation, net of costs, is 2.14 euro (13.8%) higher than if everybody received the gift (Column 3), 2.20 euro (14.3%) higher than if no one received the gift (Column 4), and 2.17 euro (14.1%) higher than if the gift was randomly allocated to one half of the warm-list sample (Column 5). Accordingly, the estimated optimal targeting rule is significantly more profitable than all three benchmark policies. In conclusion, our techniques allow the fundraiser to increase net donations significantly and, hence, service and goods provision.

Third, Panel C documents that, by applying the estimated optimal targeting rule, the fundraiser would also impact outcomes besides net donations (labeled secondary outcomes). Thus, although we train our algorithm to maximize net donations, implementing the estimated rule would trigger secondary effects as a byproduct. One secondary outcome is the donation probability within the first postexperiment year (see Row C1). Specifically, by implementing the estimated optimal rule, the fundraiser would increase this probability by almost three percentage points (5%) compared to the no-gift benchmark. Our proposed targeting strategy, hence, not only maximizes net donations, but also broadens the donor base compared to a scenario without gifts. In contrast to this result, the donation probability under the estimated rule is not significantly higher than under the all-gift benchmark. Hence, although this benchmark endows many more individuals with the gift, it does not fundamentally increase the donation probability. This result suggests that intensive margin responses drive the difference in net donations between the benchmark and the estimated rule. Besides impacts on the donation probability, the table also reveals secondary effects on longer-term outcomes. For example, Row C2 of Table 3 demonstrates that, when using the outcome "aggregate donations made within two years after the experiment," the positive effects of applying the estimated optimal targeting rule persists. This is an important result from the fundraiser's perspective: It highlights that the machine-learning-induced increase in net donations would not curb subsequent donations.

The estimated optimal targeting rule in the cold list. Table 4 shows the out-of-sample performance of the estimated optimal targeting rule in the cold list. Due to the substantial size of the cold-list sample (17,000 individuals), all of the effects are very precisely estimated. Again, the results are very different from those for the warm list. One marked difference is that the estimated target group is much narrower in the cold list (Panel A): In line with the evidence from the sorted-effects model, the estimated optimal rule assigns the gift to just 1.4% of the cold-list individuals. Given this finding, it is not surprising that the average net donation under the estimated optimal rule (0.15 euro) is virtually identical to that under the no-gift benchmark (Column 3 in Panel B). By contrast, the estimated optimal targeting rule outperforms the all-gift (Column 2 in Panel B) and random-gift benchmarks (Column 4 in Panel B). The reason is that campaigns that apply these two benchmarks would result in losses. Table 4 also presents evidence on secondary effects (Panel C). Again, we find no evidence for pull-forward or delay effects. Further, there are only minimal impacts on the donation probability. To sum up, the potential of machine-learning-based optimal targeting in the cold list is

Table 4: Out-of-sample performance of targeting rule in the cold list

	Expected outcome value under optimal targeting	Optimal targeting vs. benchmarks		
		all-gift	no-gift	random-gift
	(1)	(2)	(3)	(4)
Panel A: Share of individuals that should receive the gift				
A1. Share treated	0.014			
Panel B: Results for primary outcome variable				
B1. Net donation amount	0.15***	0.97***	-0.005	0.48***
(1st year)	(0.02)	(0.10)	(0.012)	(0.05)
Panel C: Results for secondary outcome variables				
C1. Donation probability	0.009***	-0.007***	0.001	-0.003**
(1st year)	(0.001)	(0.003)	(0.001)	(0.001)
C2. Net donation amount	0.44***	0.96***	0.04	0.50***
(1st and 2nd year)	(0.07)	(0.13)	(0.06)	(0.07)
C3. Donation probability	0.017***	-0.006*	0.001	-0.003
(1st and 2nd year)	(0.001)	(0.003)	(0.001)	(0.002)

Notes: This table documents the out-of-sample performance of our estimated optimal targeting rule, focusing on the cold list. The goal of optimal targeting is to maximize donations, net of costs. Panel A reports the share of individuals that, according to the rule, should receive the gift. Panel B reports the expected consequences of our rule for net donations as our main outcome. Panel C, instead, focuses on secondary outcomes. The columns can be interpreted as follows. Column 1 reports the expected value of the outcomes under optimal targeting. For example, we expect that, under optimal targeting, the donations, net of costs, would be 0.15 euro. Columns 2–4 show how optimal targeting changes the outcomes relative to three benchmark scenarios: everybody receives the gift (Column 2), no one receives the gift (Column 3), and the gift is randomly assigned to half of the sample (Column 4). Methodologically, the optimal targeting rules are estimated with Exact Policy-Learning Trees and a search depth of two (Zhou *et al.*, 2018). Donations are measured in euro. Standard errors are in parentheses. ***/**/* indicate statistical significance at the 1%/5%/10% level.

very limited, perhaps because the cold list consists of many notorious nondonors. This insight was not clear *a priori* and could only be established with a flexible, data-driven approach such as ours.

5.4 Characteristics of Predicted Net Donors and Net Receivers

A common theme in the fundraising literature is understanding the characteristics of individuals who give to charitable causes (Andreoni and Payne, 2013). Our machine-learning approach allows us to extend this literature by performing a broader descriptive analysis: Instead of merely describing the characteristics of givers, we can differentiate between the predicted net donors and predicted net recipients (i.e., individuals who increase donation by less than the gift's cost). The analysis, hence, reveals the characteristics of the individuals who, according to our estimated targeting rule, should receive the gift and contrasts them with the characteristics of those who should not be targeted.

Table 5: Characteristics of individuals in the warm list

	Individuals targeted by the algorithm				
	Yes (net donors)		**No (net receivers)**		Std.
	Mean	Std. dev.	Mean	Std. dev.	diff.
	(1)	(2)	(3)	(4)	(5)
Panel A: Donation history before the experiment					
A1. Num. donations prev. 8 years	4.097	2.827	3.900	2.829	6.934
A2. Max. donation prev. 8 years (euro)	39.94	44.61	34.05	41.89	13.63
A3. Total donations prev. 8 years (euro)	130.98	150.69	123.39	187.39	4.460
A4. Donations 1 year ago (euro)	22.69	36.49	19.53	34.60	8.891
A5. Donations 2 years ago (euro)	17.91	30.30	16.89	28.77	3.466
A6. Donations 3 years ago (euro)	16.61	27.49	15.62	27.52	3.593
A7. Donations 4 years ago (euro)	16.71	28.33	15.37	27.21	4.813
A8. Donations 5 years ago (euro)	15.69	24.48	15.03	29.96	2.410
Panel B: Geospatial information					
B1. Elevation (meters)	308.66	6.266	321.38	9.524	157.80
B2. In 300 meters proximity:					
Number of restaurants	10.86	13.30	6.528	7.711	39.88
Number of supermarkets	1.062	1.371	1.086	1.362	1.748
Number of medical facilities	10.17	13.95	9.298	12.041	6.703
Number of cultural facilities	0.240	0.796	0.050	0.241	32.27
Number of churches	1.166	1.515	0.934	1.460	15.60
B3. Distance to main station (km)	3.247	2.521	3.245	1.867	0.053
B4. Distance to city hall (km)	2.927	2.237	3.196	1.856	13.11
B5. Distance to main church (km)	2.986	2.365	3.218	1.836	10.99
B6. Distance to airport (km)	5.427	1.236	5.483	1.960	3.408
B7. Travel time to main station (minutes)	18.42	11.79	17.50	7.560	9.371
Panel C: Socioeconomic characteristics					
C1. Female dummy	0.507		0.539		6.459
C2. Single dummy	0.503		0.496		1.464
C3. Widowed dummy	0.050		0.052		0.974
C4. Age (years)	68.08	18.23	68.72	18.34	3.488
C5. Residency duration (years)	7.423	1.690	7.439	1.659	0.951
Observations	787		15,67		

Notes: This table describes characteristics of the warm-list individuals who, according to our estimated optimal targeting rule, should receive a gift (predicted net donors) or should not receive a gift (predicted net recipients). Particularly, it reports the means and standard deviations of all observed characteristics for the predicted net donors (Columns 1–2) and the predicted net recipients (Columns 3–4). It also shows the standardized difference (Column 5). The residency duration in the urban area is censored after 8 years. Further, we measure travel time to the main station using public transportation at 9 : 00am on weekdays. For dummy variables, the first moment is sufficient to infer the entire distribution. Rosenbaum and Rubin (1983) classify absolute standardized differences (std. diff.) of more than 20 as *large*.

Characteristics of predicted net donors and net receivers in the warm list. Table 5 focuses on the warm list. It reports the means and standard deviations of all observed characteristics for predicted net donors (Columns 1–2) and predicted net receivers (Columns 3–4). It also shows the standardized difference, a standard balance diagnostic (Column 5). The table highlights some apparent differences between the two groups. For example, before the experiment, the predicted net donors donated on average more and also more frequently than the predicted net recipients (Panel A). They also live in more central areas that are characterized by (a) lower altitudes and (b) a more lively environment with more churches, restaurants, and cultural facilities (Panel B).[28] By contrast, the differences in socioeconomic characteristics are not very pronounced. If anything, predicted net donors are more frequently female (Panel C). When interpreting these results, keep in mind that the characteristics are not necessarily reflecting channels through which the impacts of the gift operate. The patterns might instead mirror effects working through correlated unobservable variables. For example, the donation history may approximate individuals' general willingness to give, and geospatial information could proxy income. In this vein, these and similar unobservable variables might channel the responses to the gift. We consider it a strength of our approach that the machine-learning algorithm can pick up the underlying forces that shape individuals' reactions to the gift without the need to collect data or explicitly model the relationships.

Characteristics of predicted net donors and net receivers in the cold list. Table 6 reports similar results for the cold list. The predicted net donors from the cold list also live in more central areas than the respective predicted net recipients, but the areas now tend to be less lively. Regarding the socioeconomic characteristics, there are more pronounced differences compared to the warm list. Relative to predicted net recipients, predicted net donors tend to have a higher likelihood of being females, singles, and newly settled residents.

Decision trees. A second, natural way to describe the estimated targeting rule is to plot decision trees. However, because we use a cross-fitting approach to evaluate the estimated rule's out-of-sample performance, we estimate 20 different trees. Considering one single tree is, hence, not very informative. Furthermore, the trees do not identify the causal effect of the donor characteristics on the net donations. Rather, they indicate correlations between donor characteristics and the gift's causal effect. For these reasons, we neither present nor causally interpret single trees in the main body of the paper; we do provide example trees in Appendix F (Figure F.1).[29]

[28]In the urban area we study, the topology correlates with distance to the city center. Concretely, individuals living in lower altitudes live on average closer to the city hall and the main church.

[29]The tree splits are based on the previous donation amount and the elevation at the home address in the warm list. In the cold list, the tree splits are based on the number of restaurants near the residence and the distance to the city hall and airport. The trees do not use socioeconomic characteristics.

Table 6: Characteristics of individuals targeted by the algorithm in the cold list

	Individuals targeted by the algorithm				
	Yes (net donors)		**No (net receivers)**		Std.
	Mean	Std. dev.	Mean	Std. dev.	diff.
	(1)	(2)	(3)	(4)	(5)
Panel A: Geospatial information					
A1. Elevation (meters)	313.47	8.183	316.15	10.34	28.71
A2. In 300 meters proximity:					
Number of restaurants	5.482	3.524	10.40	11.67	57.08
Number of supermarkets	0.888	1.122	1.296	1.502	30.77
Number of medical facilities	13.73	10.20	10.68	13.17	25.89
Number of cultural facilities	0.040	0.196	0.146	0.532	26.43
Number of churches	1.100	1.017	1.177	1.538	5.957
A3. Distance to main station (km)	1.995	0.890	2.874	2.028	56.14
A4. Distance to city hall (km)	1.364	0.645	2.816	1.885	103.02
A5. Distance to main church (km)	1.588	0.743	2.803	1.932	83.03
A6. Distance to airport (km)	4.143	1.038	5.567	1.642	103.68
A7. Travel time to main station (minutes)	12.50	6.312	16.18	8.680	48.45
Panel B: Socioeconomic characteristics					
B1. Female dummy	0.558		0.503		11.04
B2. Single dummy	0.713		0.642		15.24
B3. Widowed dummy	0.024		0.017		4.518
B4. Age (years)	47.58	21.00	48.41	19.30	4.132
B5. Residency duration (years)	5.677	2.964	5.973	2.818	10.22
Observations	251		17,174		

Notes: This table describes characteristics of the cold-list individuals who, according to our estimated optimal targeting rule, should receive a gift (predicted net donors) or should not receive a gift (predicted net recipients). Particularly, it reports the means and standard deviations of all observed characteristics for the predicted net donors (Columns 1–2) and the predicted net recipients (Columns 3–4). It also shows the standardized difference (Column 5). The residency duration in the urban area is censored after 8 years. Further, we measure travel time to the main station using public transportation at 9 : 00am on weekdays. For dummy variables, the first moment is sufficient to infer the entire distribution. Rosenbaum and Rubin (1983) classify absolute standardized differences (std. diff.) of more than 20 as *large*.

5.5 Relevant Type of Data for Optimal Targeting

As discussed before, we estimate the optimal targeting rule using socioeconomic characteristics, donation history, and geospatial information. The next step of our analysis explores which of these data are especially powerful to estimate targeting rules. It also investigates if our algorithms require all the data to estimate effective optimal targeting rules. Besides being interesting in itself, studying this topic is vital for charities. Data collection is costly, and, frequently, some forms of data (such as socioeconomic characteristics) are unavailable. In many settings, a charity might only have access to address data before sending out written solicitations. Hence, from a charity's perspective, the usefulness and feasibility of machine-learning-based optimal targeting critically depend on the data required to target net donors effectively.

Table 7: Relevant data types in the warm list

Share treated	Expected donations under optimal targeting	Optimal targeting vs. all-gift	no-gift	random-gift	all variables
(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Socioeconomic characteristics					
0.55	15.71***	0.24	0.29	0.27	-1.91**
	(0.79)	(0.86)	(0.77)	(0.58)	(0.89)
Panel B: Donation history before the experiment					
0.12	17.20***	1.73**	1.79**	1.76***	-0.41
	(0.97)	(0.82)	(0.81)	(0.58)	(0.55)
Panel C: Geospatial information					
0.49	17.40***	1.93**	1.98**	1.95***	-0.22
	(0.91)	(0.84)	(0.79)	(0.58)	(0.61)
Panel D: Socioeconomic characteristics and donation history					
0.11	17.05***	1.58*	1.64**	1.61***	-0.56
	(0.96)	(0.84)	(0.79)	(0.58)	(0.56)
Panel E: Socioeconomic characteristics and geospatial information					
0.48	16.97***	1.50*	1.55*	1.53***	-0.64
	(0.89)	(0.84)	(0.80)	(0.58)	(0.62)
Panel F: Donation history and geospatial information					
0.33	17.61***	2.14***	2.20***	2.17***	0
	(0.97)	(0.82)	(0.81)	(0.58)	

Notes: This table documents the out-of-sample performance of targeting rules that rely only on selected subsets of our variables, focusing on the warm list. The rule in Panel A is based only on socioeconomic characteristics, in Panel B on the donation history, in Panel C on geospatial information, in Panel D on socioeconomic characteristics and the donation history, in Panel E on socioeconomic characteristics and geospatial information, and in Panel F on the donation history and geospatial information. Column 1 reports the share of individuals who, according to the respective rule, should receive the gift. Column 2 reports expected donations under optimal targeting. Columns 3–5 show how optimal targeting changes the outcomes relative to three benchmark scenarios: the all-gift (Column 3), no-gift (Column 4), and random-gift (Column 5) benchmarks. Column 6 compares the rule that only uses the subset of variables to the rule that relies on the full set of data. Methodologically, the rules are estimated with an Exact Policy-Learning Trees and a search depth of two (Zhou *et al.*, 2018). Donations are measured in euro. Standard errors are in parentheses. ***/**/* indicate statistical significance at the 1%/5%/10% level.

Relevant data in the warm list. Table 7 focuses on the warm list and explores the relevance of the different data types for the performance of the estimated optimal targeting rule. For this purpose, it evaluates the performance of several estimated targeting rules, each using only a subset of the available data. As before, the table benchmarks these more sparsely estimated rules against the all-gift (Column 3), no-gift (Column 4), and random-gift (Column 5) benchmarks. Additionally, Column 6 compares these newly estimated rules to the estimated optimal (baseline) rule obtained when using all data (reported in Table 3).

The results are as follows: First, our baseline rule clearly outperforms a rule that relies only on socioeconomic characteristics. Also, the rule based only on socioeconomic characteristics does not outperform the three benchmarks (Panel A). As socioeconomic data are often hard to obtain, these results might not be problematic for charities. Second, our baseline rule does not significantly dominate rules that either use only data on past donations (Panel B) or use only geospatial information (Panel C). Similar to our baseline rule, these two more sparsely estimated rules also significantly outperform the three benchmarks. These findings suggest that the data on past donations and geospatial information are substitutes in targeting. Hence, charities that do not have access to details on individuals' donation histories might instead rely on publicly available geospatial data only. Third, Panels D–F further emphasize that the socioeconomic characteristics are of very limited use for machine-learning-based optimal targeting. Adding them does not improve the rules that rely solely on the donation history (Panels B and D) or geospatial information (Panels C and F). Furthermore, a rule that combines the donation history with the geospatial information performs as well as a rule that uses all the data. The reason is that our baseline rule is not relying on the socioeconomic characteristics at all.

Relevant data in the cold list. Table 8 reports similar analyses for the cold list. To that end, it restricts the data either to socioeconomic characteristics or geospatial information. It turns out that the socioeconomic characteristics are also redundant in the cold list: Once again, our baseline targeting rule does not use socioeconomic characteristics. Thus, the results do not change when using only the geospatial information instead of all data (see Panel B).

We draw two main conclusions from this subsection. First, in the warm list, the fundraiser can significantly improve its campaigns' profits by relying only on widely available geospatial information. Put differently, the fundraiser does not necessarily need access to detailed data on past donations, and socioeconomic characteristics seem to be of little use for optimal targeting. This finding raises the attractiveness of our approach for charities that, for the sake of simplicity or due to data-collection costs, prefer to rely on a single data source. Second, in the cold list, the potential benefits of machine-learning-based optimal targeting are very limited. Given the data available, the dominant strategy is not to send the gift to individuals in the cold list.

Table 8: Relevant data types in the cold list

Share treated	Expected donations under optimal targeting	Optimal targeting vs. all-gift	no-gift	random-gift	all variables
(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Socioeconomic characteristics					
0.015	0.14***	0.96***	-0.014**	0.47***	-0.01
	(0.02)	(0.10)	(0.007)	(0.05)	(0.014)
Panel B: Geospatial information					
0.014	0.15***	0.97***	-0.005	0.48***	0
	(0.02)	(0.10)	(0.012)	(0.05)	

Notes: This table documents the out-of-sample performance of targeting rules that rely only on selected subsets of our variables, focusing on the cold list. The rule in Panel A is based on socioeconomic characteristics only, and in Panel B on geospatial information only. Column 1 reports the share of individuals who, according to the respective rule, should receive the gift. Column 2 reports the expected donations under optimal targeting. Columns 3–5 show how optimal targeting changes the outcomes relative to three benchmark scenarios: the all-gift (Column 3), no-gift (Column 4), and random-gift (Column 5) benchmarks. Column 6 compares the rule that uses only the subset of variables to the rule that relies on the full set of data. Methodologically, the rules are estimated with Exact Policy-Learning Trees and a search depth of two (Zhou *et al.*, 2018). Donations are measured in euro. Standard errors are in parentheses. ***/**/* indicate statistical significance at the 1%/5%/10% level.

5.6 Sensitivity Analysis

Our final step is to investigate the robustness of the estimated optimal targeting rules to different estimation approaches. First, we consider alternative depths of the Exact Policy-Learning Tree (one and three).[30] Second, we compare the results of the Exact Policy-Learning Trees to the results of standard CARTs (Breiman *et al.*, 1984). Third, for the CARTs, we also consider one version in which we use cross-validation to select the tree depth in a data-driven way.[31] Fourth, we employ weighted Logit as a standard estimator. Fifth, we use two alternative machine-learning estimators: Logit Lasso (Hastie *et al.*, 2016) and classification forests (Breiman, 2001).[32] For the Logit estimator, we consider two different model specifications. The baseline specification includes all observed characteristics linearly (24 variables in the warm list and 16 variables in the cold list). The flexible specification additionally includes squared terms of continuous variables and first-order interactions between most characteristics (321 variables in the

[30]For the cold list, Exact Policy-Learning Trees with a depth of three are infeasible due to computational constraints.

[31]We use the Gini index for tree splitting and a 10-fold cross-validation procedure to select the optimal tree depth. In the warm list, the number of terminal leaves varies between four and nine across the 20 different cross-fitted trees, with an average of 4.6. In the cold list, the number of terminal leaves varies between two and eight across the 20 different cross-fitted trees, with an average of 2.3.

[32]We build 1,000 trees for the classification forest. We draw a 50% random subsample with replacement for each tree and randomly select 50% of the baseline characteristics. We use the Gini index for tree splitting. We restrict the minimum size of the terminal leaf to 50 observations.

warm list and 149 variables in the cold list). The Logit-Lasso method selects the relevant characteristics from the flexible model specification.[33]

Table G.1 in Online Appendix G reports the results of the sensitivity analysis for the warm list, and Table G.2 focuses on the cold list. In the warm list, our baseline optimal targeting strategy clearly dominates the alternative specifications: The baseline optimal targeting rules of all the alternative estimators yield lower net donations than our main specification. Specifically, the Exact Policy-Learning Tree with depth one and the Logit with the flexible model specification have the lowest out-of-sample performance. The CART with the cross-validated tree depth is the only alternative estimator that also significantly outperforms the all-gift and no-gift benchmark allocation rules. For the cold list, the Logit-Lasso specification and the CART with cross-validated tree depth yield 0.01 euro higher net donations than our main specification. However, our main finding that the fundraiser should not send the gift to cold-list individuals persists.

6 Conclusion

This paper studies machine-learning-based optimal targeting of fundraising instruments by exploiting data from a natural field experiment. The underlying idea of optimal targeting is that fundraisers can maximize a campaign's profits by directing a fundraising instrument to individuals who increase their donations in response to the instrument by more than its cost. We label those individuals *net donors*. However, charities do not observe the set of net donors. We employ a machine-learning algorithm to estimate the relationship between individual characteristics and the potential donors' response to small unconditional gifts. Based on this algorithm, we can predict the subset of net donors and, hence, estimate machine-learning-based optimal targeting rules.

Our paper's first key message is that, in the warm list, machine-learning-based optimal targeting substantially boosts the charity's net donations. In our application, net donations increase by about 14% compared to the all-gift and no-gift benchmarks. Notably, the benefits of machine-learning-based optimal targeting even materialize when relying only on widely available geospatial data. Hence, charities can easily apply the proposed strategies to raise additional funds, net of costs. The second message is that, in the cold list, the approach does not raise donations sufficiently to cover the fundraising instrument's additional costs. We conclude that the fundraiser should not target cold-list individuals at all. We also document that the increase in net donations stems from heterogeneities in donors' responses to fundraising activities. Previous literature has suggested that such heterogeneities exist, for example, by providing mixed evidence on the effectiveness of unconditional gifts on giving (Falk, 2007; Yin *et al.*, 2020; Alpizar

[33] We specify the penalty λ of the Logit Lasso that minimizes the misclassification error using a 10-fold cross-validation approach.

et al., 2008).

In conclusion, our paper demonstrates that machine-learning-based optimal targeting can significantly increase the cost effectiveness of fundraising. One particularly noteworthy benefit of our applied machine-learning toolkit is that it allows charities to target fundraising efforts agnostically in a wide variety of contexts. Thus, to optimize targeting, charities do not need to develop a theoretical foundation or make strong assumptions on the functional relationship between individual characteristics and giving. We are, therefore, confident that the proposed approach offers an accessible way forward to improve the effectiveness of fundraising. We are also looking forward to future, evolving research applying similar techniques to alternative fundraising instruments and different settings. It will also be interesting to see how our findings generalize across these alternative tools and environments.

References

ABE, N., VERMA, N., APTE, C. and SCHROKO, R. (2004). Cross Channel Optimized Marketing by Reinforcement Learning. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge discovery and data mining*, pp. 767–772.

ADENA, M. and HUCK, S. (2019). Personalized Fundraising: A Field Experiment on Threshold Matching of Donation. *WZB Discussion Paper No. SP II 2019-306*.

ALDASHEV, G., MARINI, M. and VERDIER, T. (2014). Brothers in Alms? Coordination Between Nonprofits on Markets for Donations. *Journal of Public Economics*, **117**, 182–200.

— and VERDIER, T. (2010). Goodwill Bazaar: NGO Competition and Giving to Development. *Journal of Development Economics*, **91** (1), 48–63.

ALPIZAR, F., CARLSSON, F. and JOHANSSON-STENMAN, O. (2008). Anonymity, Reciprocity, and Conformity: Evidence from Voluntary Contributions to a National Park in Costa Rica. *Journal of Public Economics*, **92** (5–6), 1047–1060.

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 and Organization*, **156**, 86–102.

ANDREONI, J. (1988). Privately Provided Public Goods in a Large Economy: The Limits of Altruism. *Journal of Public Economics*, **35** (1), 57–73.

—, BROWN, E. and RISCHALL, I. (2003). Charitable Giving by Married Couples. Who Decides and Why Does It Matter? *Journal of Human Resources*, **38** (1), 111–133.

— and PAYNE, A. (2003). Do Government Grants to Private Charities Crowd Out Giving or Fund-Raising? *American Economic Review*, **93** (3), 792–812.

— and — (2011). Is Crowding Out Due Entirely to Fundraising? Evidence from a Panel of Charities. *Journal of Public Economics*, **95**, 334–343.

— and — (2013). Charitable Giving. In A. Auerbach, R. Chetty, M. Feldstein and E. Saez (eds.), *Handbook of Public Economics*, vol. 5, *1*, Elsevier, pp. 1–50.

— and VESTERLUND, L. (2001). Which Is the Fair Sex? Gender Differences in Altruism. *Quarterly Journal of Economics*, **116** (1), 293–312.

ASCARZA, E. (2018). Retention Futility: Targeting High-Risk Customers Might be Ineffective. *Journal of Marketing Research*, **55** (1), 80–98.

ATHEY, S. and IMBENS, G. W. (2019). Machine Learning Methods that Economists Should Know About. *Annual Review of Economics*, **11**, 685–725.

— and WAGER, S. (2021). Policy Learning with Observational Data. *Econometrica*, **forthcoming**.

BAKSHY, E., ECKLES, D., YAN, R. and ROSENN, I. (2012). Social Influence in Social Advertising: Evidence from Field Experiments. *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 146–161.

BEYGELZIMER, A. and LANGFORD, J. (2009). The Offset Tree for Learning with Partial Labels. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 129–138.

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

—, FRIEDMAN, J., STONE, C. and OLSHEN, R. (1984). *Classification and Regression Trees*. CRC Press.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C. and NEWEY, W. (2017). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, **107** (5), 261–265.

—, ESCANCIANO, J., ICHIMURA, H., NEWEY, W. and ROBINS, J. (2018a). Locally-Robust Semiparametric Estimation. *arXiv:1608.00033*.

—, FERNÁNDEZ-VAL, I. and LUO, Y. (2018b). The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *Econometrica*, **86** (6), 1911–1938.

CUI, G., WONG, M. L. and LUI, H.-K. (2006). Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming. *Management Science*, **52** (4), 597–612.

DE VRIES, N., REIS, R. and MOSCATO, P. (2015). Clustering Consumers Based on Trust, Confidence and Giving Behaviour: Data-Driven Model Building for Charitable Involvement in the Australian Not-for-Profit Sector. *PloS One*, **10** (4), 1–28.

DEUTSCHER SPENDENRAT (2016). Bilanz des Helfens. *Press Release*.

DONG, L., RATTI, C. and ZHENG, S. (2019). Predicting Neighborhoods' Socioeconomic Attributes Using Restaurant Data. *PNAS*, **116** (31), 15447–15452.

DROUVELIS, M. and MARX, B. (2021). Can Charitable Appeals Identify and Exploit Belief Heterogeneity? *CESifo Working Paper No. 8855*.

DUFWENBERG, M. and KIRCHSTEIGER, G. (2004). A Theory of Sequential Reciprocity. *Games and Economic Behavior*, **47** (2), 268–298.

ECKEL, C., HERBERICH, D. and MEER, J. (2018). It's Not the Thought that Counts: A Field Experiment on Gift Exchange and Giving at a Public University. In *The Economics of Philanthropy: Donations and Fundraising*, MIT Press.

ELLICKSON, P., KAR, W. and REEDER, J. (2020). Machine Learning Marketing Component Effects: Applying Causal Forests to Targeted Email Promotions. *mimeo*.

FALK, A. (2007). Gift Exchange in the Field. *Econometrica*, **75**, 1501–1511.

—, BECKER, A., DOHMEN, T., ENKE, B., HUFFMAN, D. and SUNDE, U. (2018). Global Evidence on Economic Preferences. *Quarterly Journal of Economics*, **133** (4), 1645–1692.

FAN, Q., HSU, Y.-C., LIELI, R. P. and ZHANG, Y. (2019). Estimation of Conditional Average Treatment Effects with High-Dimensional Data. *arXiv:1908.02399*.

FONG, N., ZHANG, Y., LUO, X. and WANG, X. (2019). Targeted Promotions on an E-book Platform: Crowding Out, Heterogeneity, and Opportunity Costs. *Journal of Marketing Research*, **56** (2), 310–323.

GLAESER, E., KIM, H. and LUCA, M. (2020). Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity. In Abraham, Jarmin, Moyer and Shapiro (eds.), *Big Data for 21st Century Economic Statistics*.

—, KOMINERS, S., LUCA, M. and NAIK, N. (2018). Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life. *Economic Inquiry*, **56** (1), 114–137.

GNEEZY, U., KEENAN, E. A. and GNEEZY, A. (2014). Avoiding Overhead Aversion in Charity. *Science*, **346** (6209), 632–635.

GUBELA, R., LESSMANN, S. and JAROSZEWICZ, S. (2020). Response Transformation and Profit Decomposition for Revenue Uplift Modeling. *European Journal of Operational Research*, **283** (2), 647–661.

GUELMAN, L., GUILLÉN, M. and PÉREZ-MARÍN, A. (2015). Uplift Random Forests. *Cybernetics and Systems*, **46** (3-4), 230–248.

HARBAUGH, W., MAYR, U. and BURGHART, D. (2007). Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations. *Science*, **316** (5831), 1622–1625.

HASSELL, H. and MONSON, J. (2014). Campaign Targets and Messages in Direct Mail Fundraising. *Political Behavior*, **36** (2), 359–376.

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2016). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

HEBLICH, S., TREW, A. and ZYLBERBERG, Y. (2020). East Side Story: Historical Pollution and Persistent Neighborhood Sorting. *Journal of Political Economy*, **forthcoming**.

HITSCH, G. and MISRA, J. (2018). Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. *SSRN preprint 3111957*.

KANG, J., KUZNETSOVA, P., LUCA, M. and CHOI, Y. (2013). Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1443–1448.

KIM, Y., STREET, N., RUSSELL, G. and MENCZER, F. (2005). Customer Targeting: A Neural Network Approach Guided by Genetic Algorithms. *Management Science*, **51** (2), 264–276.

KITAGAWA, T. and TETENOV, A. (2018). Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, **86** (2), 591–616.

KIZILCEC, R., BAKSHY, E., ECKLES, D. and BURKE, M. (2018). Social Influence and Reciprocity in Online Gift Giving. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–11.

KLEINBERG, J., LUDWIG, J., MULLAINATHAN, S. and OBERMEYER, Z. (2015). Prediction Policy Problems. *American Economic Review*, **105** (5), 491–495.

KNAUS, M. (2020). Double Machine Learning Based Program Evaluation Under Unconfoundedness. *arXiv:2003.03191*.

—, LECHNER, M. and STRITTMATTER, A. (2020). Heterogeneous Employment Effects of Job Search Programs: A Machine Learning Approach. *Journal of Human Resources*, **forthcoming**.

—, — and — (2021). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. *Econometrics Journal*, **24** (1), 134–161.

KNITTEL, C. and STOLPER, S. (2019). Using Machine Learning to Target Treatment: The Case of Household Energy Use. *NBER Working Papers 26531*.

LANDRY, C., LANGE, A., LIST, J., PRICE, M. and RUPP, N. (2010). Is a Donor in Hand Better Than Two in the Bush? Evidence from a Natural Field Experiment. *American Economic Review*, **100** (3), 958–983.

LIST, J. (2011). The Market for Charitable Giving. *Journal of Economic Perspectives*, **25** (2), 157–80.

LIST, J. A. and PRICE, M. K. (2012). Charitable Giving Around the World: Thoughts on How to Expand the Pie. *CESifo Economic Studies*, **58** (1), 1–30.

MANSKI, C. (2004). Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, **72** (4), 1221–1246.

MEER, J. (2017). Does Fundraising Create New Giving? *Journal of Public Economics*, **145**, 82–93.

MORGAN, J. (2000). Financing Public Goods by Means of Lotteries. *The Review of Economic Studies*, **67** (4), 761–784.

NAME-CORREA, A. and YILDIRIM, H. (2013). A Theory of Charitable Fund-Raising with Costly Solicitations. *American Economic Review*, **103** (2), 1091–1107.

NEWMAN, G. E. and SHEN, J. Y. (2012). The Counterintuitive Effects of Thank-You Gifts on Charitable Giving. *Journal of Economic Psychology*, **33** (5), 973–983.

OKTEN, C. and WEISBROD, B. (2000). Determinants of Donations in Private Nonprofit Markets. *Journal of Public Economics*, **75** (2), 255–272.

RAJAN, S., PINK, G. and DOWN, W. (2009). Sociodemographic and Personality Characteristics of Canadian Donors Contributing to International Charity. *Nonprofit and Voluntary Sector Quarterly*, **38** (3), 413–440.

ROBINS, J., ROTNITZKY, A. and ZHAO, L. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, **89** (427), 846–866.

ROCKOFF, J., JACOB, B., KANE, T. and STAIGER, D. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance Policy*, **6** (1), 43–74.

ROSE-ACKERMAN, S. (1982). Charitable Giving and "Excessive" Fundraising. *The Quarterly Journal of Economics*, **97** (2), 193–212.

ROSENBAUM, P. and RUBIN, D. (1983). The Central Role of Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70** (1), 41–55.

RUBIN, D. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66** (5), 688–701.

SCHLEGELMILCH, B. and DIAMANTOPOULOS, A. (1997). Characteristics Affecting Charitable Donations: Empirical Evidence from Britain. *Journal of Marketing Practice: Applied Marketing Science*, **3** (1), 14–28.

SCHWARTZ, E., BRADLOW, E. and FADER, P. (2017). Customer Acquisition via Display Advertising Using Multi-armed Bandit Experiments. *Marketing Science*, **36** (4), 500–522.

SEMENOVA, V. and CHERNOZHUKOV, V. (2020). Debiased Machine Learning of Conditional Average Treatment Effects and Other Causal Functions. *Econometrics Journal*, **forthcoming**.

SIMESTER, D., TIMOSHENKO, A. and ZOUMPOULIS, S. (2020). Targeting Prospective Customers: Robustness of Machine-Learning Methods to Typical Data Challenges. *Management Science*, **66** (6), 2495–2522.

SMITH, A., SEILER, S. and AGGARWAL, I. (2021). Optimal Price Targeting. *mimeo*.

SVERDRUP, E., KANODIA, A., ZHOU, Z., ATHEY, S. and WAGER, S. (2020). Policytree: Policy Learning via Doubly Robust Empirical Welfare Maximization over Trees. *Journal of Open Source Software*, **5** (50), 1–6.

TINKELMAN, D. and MANKANEY, K. (2007). When is Administrative Efficiency Associated with Charitable Donations? *Nonprofit and Voluntary Sector Quarterly*, **36**, 41–64.

VESTERLUND, L. (2003). The Informational Value of Sequential Fundraising. *Journal of Public Economics*, **87** (3-4), 627–657.

WIEPKING, P. and JAMES, R. (2013). Why Are the Oldest Old Less Generous? Explanations for the Unexpected Age-Related Drop in Charitable Giving. *Ageing and Society*, **33** (3), 486–510.

WINTERICH, K. P., MITTAL, V. and AQUINO, K. (2013). When Does Recognition Increase Charitable Behavior? Toward a Moral Identity-Based Model. *Journal of Marketing*, **77** (3), 121–134.

YIN, B., LI, Y. and SINGH, S. (2020). Coins Are Cold and Cards Are Caring: The Effect of Pregiving Incentives on Charity Perceptions, Relationship Norms, and Donation Behavior. *Journal of Marketing*, **84** (6), 57–73.

ZADROZNY, B. (2003). *Policy Mining: Learning Decision Policies from Fixed Sets of Data*. Ph.D. Thesis, University of California, San Diego.

ZHAO, Y., ZENG, D., RUSH, A. and KOSOROK, M. (2012). Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, **117** (499), 1106–1118.

ZHOU, Z., ATHEY, S. and WAGER, S. (2018). Offline Multi-action Policy Learning: Generalization and Optimization. *arXiv:1810.04778v2*.

ZIMMERT, M. and LECHNER, M. (2019). Nonparametric Estimation of Causal Heterogeneity Under High-Dimensional Confounding. *arXiv:1908.0877*.

Online Appendix to
"Optimal Targeting in Fundraising"

Tobias Cagala, Ulrich Glogowsky,
Johannes Rincke, and Anthony Strittmatter

Sections:

Cagala: Deutsche Bundesbank (tobias.cagala@bundesbank.de); Glogowsky: Johannes Kepler University Linz and CESifo (ulrich.glogowsky@jku.at); Rincke: University of Erlangen-Nuremberg (johannes.rincke@fau.de); Strittmatter: CREST-ENSAE, Institut Polytechnique Paris and CESifo (anthony.strittmatter@ensae.fr).

A Descriptives and Balance of Observables

Table A.1: Means and standard deviations of observable characteristics

	Warm list		Cold list	
	Mean	Std. dev.	Mean	Std. dev.
	(1)	(2)	(3)	(4)
Panel A. Socioeconomic characteristics				
Female dummy	0.53		0.50	
Single dummy	0.50		0.64	
Widowed dummy	0.05		0.02	
Age (years)	68.51	18.30	48.40	19.32
Residency duration in urban area (years)	7.43	1.67	5.97	2.82
Panel B: Donation history before the experiment				
Number of donations previous 8 years	3.97	2.83	0	
Max. donations previous 8 years (euro)	36.02	42.90	0	
Total donations previous 8 years (euro)	125.9	176.0	0	
Donations 1 year ago (euro)	20.59	35.27	0	
Donations 2 years ago (euro)	17.23	29.29	0	
Donations 3 years ago (euro)	15.95	27.51	0	
Donations 4 years ago (euro)	15.82	27.59	0	
Donations 5 years ago (euro)	15.25	28.24	0	
Panel C: Geospatial information about home address				
Elevation (meters)	317.1	10.46	316.1	10.32
In 300 meters proximity:				
Number of restaurants	7.98	10.14	10.33	11.61
Number of supermarkets	1.08	1.36	1.29	1.50
Number of medical facilities	9.59	12.72	10.72	13.13
Number of cultural facilities	0.11	0.51	0.14	0.53
Number of churches	1.01	1.48	1.18	1.53
Distance to main station (km)	3.25	2.11	2.86	2.02
Distance to city hall (km)	3.11	2.00	2.79	1.88
Distance to main church (km)	3.14	2.03	2.79	1.93
Distance to airport (km)	5.46	1.75	5.55	1.64
Travel time to main station (minutes)	17.81	9.20	16.13	8.66
Observations	2,354		17,425	

Notes: This table describes the characteristics of cold and warm-list individuals. The donation history in the cold list is zero, because we only measure the donations to the specific fundraiser we cooperate with. The residency duration in the urban area is censored after 8 years. We measure travel time to the main station using public transportation at 9 : 00am on weekdays. For dummy variables, the first moment is sufficient to infer the entire distribution. Rosenbaum and Rubin (1983) classify absolute standardized difference (std. diff.) of more than 20 as *large*.

Table A.2: Balance of observable characteristics in the warm list

	Treatment group		Control group		Std.
	Mean	Std. dev.	Mean	Std. dev.	diff.
	(1)	(2)	(3)	(4)	(5)
Panel A: Socioeconomic characteristics					
Female dummy	0.53		0.53		1.22
Single dummy	0.50		0.49		2.04
Widowed dummy	0.05		0.06		3.98
Age (years)	68.57	18.31	68.45	18.31	0.69
Residency duration (years)	7.41	1.70	7.46	1.63	3.13
Panel B: Donation history before the experiment					
Num. donations prev. 8 years	3.94	2.81	3.99	2.85	1.98
Max. don. prev. 8 years (euro)	36.49	46.63	35.54	38.81	2.22
Total don. prev. 8 years (euro)	126.64	181.52	125.21	170.31	0.82
Donations 1 year ago (euro)	21.19	39.73	19.98	30.13	3.41
Donations 2 years ago (euro)	17.28	29.55	17.18	29.03	0.32
Donations 3 years ago (euro)	16.11	28.13	15.80	26.88	1.13
Donations 4 years ago (euro)	16.30	28.79	15.34	26.34	3.48
Donations 5 years ago (euro)	14.83	28.53	15.67	27.95	2.96
Panel C: Geospatial information about home address					
Elevation (meters)	317.4	10.58	316.9	10.35	4.89
In 300 meters proximity:					
Number of restaurants	7.86	10.15	8.10	10.14	2.42
Number of supermarkets	1.11	1.37	1.05	1.36	4.83
Number of medical facilities	9.52	12.52	9.66	12.91	1.03
Number of cultural facilities	0.11	0.52	0.11	0.50	0.22
Number of churches	1.03	1.50	1.00	1.46	1.83
Distance to main station (km)	3.24	2.04	3.25	2.18	0.59
Distance to city hall (km)	3.09	1.93	3.12	2.06	1.15
Distance to main church (km)	3.13	1.96	3.15	2.10	0.87
Distance to airport (km)	5.44	1.77	5.49	1.74	2.47
Travel time to main station (minutes)	17.72	8.80	17.90	9.58	1.94
Observations	1,180		1,174		

Notes: This table describes the characteristics of the warm-list individuals in the treatment and control groups. The residency duration in the urban area is censored after 8 years. We measure travel time to the main station using public transportation at 9 : 00am on weekdays. For dummy variables, the first moment is sufficient to infer the entire distribution. Rosenbaum and Rubin (1983) classify absolute standardized difference (std. diff.) of more than 20 as *large*.

Table A.3: Balance of observable characteristics in the cold list

	Treatment group		Control group		Std.
	Mean	Std. dev.	Mean	Std. dev.	diff.
	(1)	(2)	(3)	(4)	(5)
Panel A: Socioeconomic characteristics					
Female dummy	0.50		0.50		0.14
Single dummy	0.64		0.64		0.23
Widowed dummy	0.02		0.02		0.35
Age (years)	48.34	19.25	48.41	19.33	0.34
Residency duration (years)	5.96	2.82	5.97	2.82	0.36
Panel B: Geospatial information about home address					
Elevation (meters)	316.2	10.42	316.1	10.31	0.40
In 300 meters proximity:					
Number of restaurants	10.02	11.29	10.38	11.66	3.17
Number of supermarkets	1.31	1.50	1.29	1.50	1.14
Number of medical facilities	10.40	12.91	10.77	13.17	2.85
Number of cultural facilities	0.14	0.51	0.15	0.53	1.10
Number of churches	1.13	1.49	1.18	1.54	3.27
Distance to main station (km)	2.89	2.02	2.86	2.02	1.69
Distance to city hall (km)	2.81	1.88	2.79	1.88	0.90
Distance to main church (km)	2.81	1.93	2.78	1.93	1.19
Distance to airport (km)	5.55	1.65	5.55	1.64	0.07
Travel time to main station (minutes)	16.25	8.80	16.11	8.64	1.64
Observations	2,283		15,142		

Notes: This table describes the characteristics of the cold-list individuals in the treatment and control groups. The residency duration in the urban area is censored after 8 years. We measure travel time to the main station using public transportation at 9 : 00am on weekdays. For dummy variables, the first moment is sufficient to infer the entire distribution. Rosenbaum and Rubin (1983) classify absolute standardized difference (std. diff.) of more than 20 as *large*.

B Identification of ATEs and CATEs

To achieve identification, we have to assume that the stratified randomization was appropriately conducted, ensuring $p(z,x) = Pr(D_i = 1|Z_i = z, X_i = x) = Pr(D_i = 1|Z_i = z) = p(z)$ and $(Y_i(1), Y_i(-1)) \perp\!\!\!\perp D_i|Z_i = z$. Furthermore, we have to assume that the probability of receiving the gift is between zero and one, $0 < p(z) < 1$.

Under these assumptions, the following equations prove that the CATEs are identified from observable data:

$$
\begin{aligned}
\delta(x) &\overset{\text{LIE}}{=} E_{Z|X=x}[E[Y_i(1) - Y_i(-1)|Z_i = z, X_i = x]], \\
&\overset{\text{Experiment}}{=} E_{Z|X=x}[E[Y_i(1)|D_i = 1, Z_i = z, X_i = x] - E[Y_i(-1)|D_i = -1, Z_i = z, X_i = x]], \\
&\overset{\text{SUTVA}}{=} E_{Z|X=x}[E[Y_i|D_i = 1, Z_i = z, X_i = x] - E[Y_i|D_i = -1, Z_i = z, X_i = x]],
\end{aligned}
$$

where the first equality is an application of the law of iterative expectations (LIE), the second equality holds under the experimental design, and the last equality under the SUTVA. The identification proof for the ATEs follows from the LIE, $\delta = E[\delta(X_i)]$.

C Identification with AIPW Scores

To prove that $\delta = E[\Gamma_i]$ and $\delta(x) = E[\Gamma_i | X_i = x]$, it is sufficient to prove that

$$E[Y_i(1)|X_i = x] = E[\Gamma_i(1)|X_i = x] = E\left[\mu_1(Z_i) + \frac{1+D_i}{2} \cdot \frac{Y_i - \mu_1(Z_i)}{p(Z_i)} \middle| X_i = x\right],$$
$$E[Y_i(-1)|X_i = x] = E[\Gamma_i(-1)|X_i = x] = E\left[\mu_{-1}(Z_i) - \frac{D_i - 1}{2} \cdot \frac{Y_i - \mu_{-1}(Z_i)}{1 - p(Z_i)} \middle| X_i = x\right].$$

We focus on $E[\Gamma_i(1)|X_i = x]$. We have:

$$E[\Gamma_i(1)|X_i = x] = E\left[\mu_1(Z_i) + \frac{1+D_i}{2} \cdot \frac{Y_i - \mu_1(Z_i)}{p(Z_i)} \middle| X_i = x\right], \tag{C.1}$$

$$= E\left[\frac{1+D_i}{2} \cdot \frac{Y_i}{p(Z_i)} \middle| X_i = x\right] + E\left[\left(p(Z_i) - \frac{1+D_i}{2}\right) \cdot \frac{\mu_1(Z_i)}{p(Z_i)} \middle| X_i = x\right], \tag{C.2}$$

$$= E_{Z|X=x}\left[E\left[\frac{1+D_i}{2} \cdot \frac{Y_i}{p(Z_i)} \middle| X_i = x, Z_i = z\right]\right] + E_{Z|X=x}\left[E\left[\left(p(Z_i) - \frac{1+D_i}{2}\right) \cdot \frac{\mu_1(Z_i)}{p(Z_i)} \middle| X_i = x, Z_i = z\right]\right], \tag{C.3}$$

$$= E_{Z|X=x}\left[E\left[\frac{1+D_i}{2} \cdot \frac{Y_i}{p(z,x)} \middle| X_i = x, Z_i = z\right]\right] + \underbrace{E_{Z|X=x}\left[E\left[\left(p(z,x) - \frac{1+D_i}{2}\right) \cdot \frac{\mu_1(Z_i)}{p(z,x)} \middle| X_i = x, Z_i = z\right]\right]}_{=0}, \tag{C.4}$$

$$= E_{Z|X=x} E\left[\left[\frac{1\{D_i = 1\} Y_i}{p(z,x)} \middle| X_i = x, Z_i = z\right]\right], \tag{C.5}$$

$$= E_{Z|X=x} E[[Y_i | D_i = 1, X_i = x, Z_i = z]], \tag{C.6}$$

$$= E_{Z|X=x} E[[Y_i(1) | D_i = 1, X_i = x, Z_i = z]], \tag{C.7}$$

$$= E_{Z|X=x} E[[Y_i(1) | X_i = x, Z_i = z]], \tag{C.8}$$

$$= E[Y_i(1)|X_i = x]. \tag{C.9}$$

In (C.1), we use the definition of $\Gamma_i(1)$. In (C.2), we just make a rearrangement. In (C.3), we apply the law of iterative expectations. In (C.4), we exploit that $p(z) = p(z,x)$ because of our experimental design (only the characteristics in Z_i have an impact on the probability of receiving the gift). Note that the second right-side term cancels, because $p(z,x) = E[(1+D_i)/2 | X_i = x, Z_i = z]$. In (C.5), we replace $(1+D_i)/2$ with the indicator function $1\{D_i = 1\}$. In (C.6), we apply the discrete law of iterative expectations backwards. In (C.7) and (C.8), we use $(Y_i(1), Y_i(-1)) \perp\!\!\!\perp D_i | Z_i = z$, the conditional-independence assumption, which holds by the experimental design. In

(C.9), we apply the law of iterative expectations backwards. This finishes the proof that $E[Y_i(1)|X_i = x] = E[\Gamma_i(1)|X_i = x]$. The proof that $E[Y_i(-1)|X_i = x] = E[\Gamma_i(-1)|X_i = x]$ proceeds analogous.

D Nuisance Parameters

Table D.1: Coefficients of nuisance parameters

	Logit Donation dummy		OLS Donations without gift		Donations with gift	
	(1)		(2)		(3)	
	Coef.	Std. err.	Coef.	Std. err.	Coef.	Std. err.
Panel A: Warm list						
Female dummy	-0.009	0.023	-3.67**	1.69	-5.00**	1.91
Single dummy	0.009	0.023	0.31	1.74	1.22	1.98
Widowed dummy	-0.041	0.050	-1.29	3.54	2.40	4.41
Age quintiles:						
2nd quintile	0.005	0.029	-3.96*	2.13	1.22	2.42
3rd quintile	0.004	0.030	-5.68**	2.20	-3.20	2.51
4th quintile	0.006	0.031	-2.61	2.31	1.12	2.64
Baseline willingness to donate quintiles:						
2nd quintile	-0.002	0.031	-29.64***	2.33	-25.76***	2.65
3rd quintile	0.001	0.030	-24.81***	2.23	-19.02***	2.55
4th quintile	0.003	0.030	-12.73***	2.23	-7.94***	2.54
Intercept	0.498***	0.031	37.13***	2.27	32.44***	2.59
Panel B: Cold list						
Female dummy	-0.003	0.046	-0.002	0.04	0.18	0.20
Single dummy	-0.003	0.052	-0.10**	0.05	-0.28	0.23
Widowed dummy	0.027	0.176	-0.01	0.17	0.93	0.77
Age quintiles:						
2nd quintile	0.009	0.065	0.09	0.06	0.22	0.28
3rd quintile	0.002	0.067	0.10	0.06	0.03	0.29
4th quintile	0.001	0.068	0.20***	0.06	0.45	0.30
Intercept	-1.891***	0.068	0.14**	0.07	0.32	0.30

Notes: This table shows the coefficients of the nuisance parameters. Donations (euro) are measured during the first year after the gift was sent. ***/**/* indicate statistical significance at the 1%/5%/10% level.

E Additional Results of the Sorted-Effects Model

Table E.1: Mean characteristics of the groups with the 10% largest and smallest treatment effects in the warm list

	10% Largest		10% Smallest		Difference		
	Mean	Std. err.	Mean	Std. err.	Mean	Std. err.	JP-val.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Socioeconomic characteristics							
Female dummy	0.49	0.05	0.41	0.05	0.08	0.08	1.00
Single dummy	0.55	0.05	0.49	0.05	0.06	0.08	1.00
Widowed dummy	0.11	0.03	0.04	0.03	0.06	0.04	0.94
Age (years)	70.84	1.72	67.45	1.90	3.39	3.10	1.00
Residency duration (years)	7.28	0.21	6.93	0.24	0.36	0.32	1.00
Panel B: Donation history before the experiment							
Num. donations prev. 8 years	3.64	0.27	4.42	0.29	-0.78	0.43	0.79
Max. don. prev. 8 years (euro)	55.05	6.05	70.13	8.36	-15.08	12.37	0.99
Total don. prev. 8 years (euro)	161.7	24.79	269.5	30.57	-107.9	43.85	0.37
Donations 1 year ago (euro)	25.53	4.94	50.67	6.96	-25.13	9.29	0.24
Donations 2 years ago (euro)	19.36	4.11	41.82	5.46	-22.46	7.81	0.17
Donations 3 years ago (euro)	26.21	3.74	36.08	4.88	-9.88	7.08	0.97
Donations 4 years ago (euro)	23.70	3.94	29.41	4.74	-5.71	7.22	1.00
Donations 5 years ago (euro)	21.33	4.22	31.81	4.52	-10.49	7.04	0.95
Panel C: Geospatial information about home address							
Elevation (meters)	315.4	1.10	321.3	1.51	-5.98	2.11	0.18
In 300 meters proximity:							
Number of restaurants	8.96	1.38	7.50	1.19	1.46	2.26	1.00
Number of supermarkets	0.79	0.15	1.12	0.14	-0.32	0.23	0.98
Number of medical facilities	9.21	1.54	9.42	1.46	-0.21	2.50	1.00
Number of cultural facilities	0.35	0.08	0.24	0.09	0.11	0.14	1.00
Number of churches	1.15	0.19	1.31	0.18	-0.16	0.29	1.00
Distance to main station (km)	3.70	0.24	3.49	0.26	0.21	0.40	1.00
Distance to city hall (km)	3.57	0.24	3.30	0.24	0.28	0.37	1.00
Distance to main church (km)	3.60	0.23	3.37	0.24	0.23	0.39	1.00
Distance to airport (km)	5.87	0.17	5.38	0.16	0.49	0.23	0.58
Time to main station (minutes)	19.75	1.11	17.99	1.05	1.75	1.84	1.00

Notes: This table shows the mean characteristics of the groups with the 10% largest and smallest treatment effects in the warm list. We report joint p-values (JP-value), which account for simultaneous inference on several characteristics. We employ the so-called *single-step* methods to control the family-wise error rate (see, e.g., Chernozhukov *et al.*, 2018b, for details). Standard errors are calculated with a multiplier bootstrap using 500 replications. The residency duration in the urban area is censored after 8 years. We measure travel time to the main station using public transportation at 9 : 00am on weekdays.

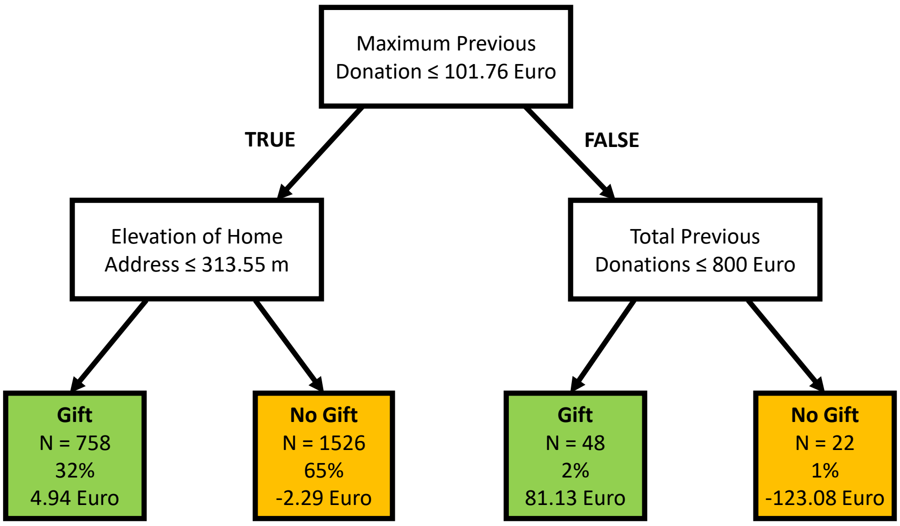
Table E.2: Mean characteristics of the groups with the 10% largest and smallest treatment effects in the cold list

	10% Largest		10% Smallest		Difference		
	Mean	Std. err.	Mean	Std. err.	Mean	Std. err.	JP-val.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Socioeconomic characteristics							
Female dummy	0.71	0.09	0.49	0.05	0.22	0.13	0.77
Single dummy	0.52	0.11	0.61	0.05	-0.09	0.12	0.99
Widowed dummy	0.11	0.02	0.06	0.01	0.05	0.03	0.66
Age (years)	55.75	3.75	51.84	1.71	3.92	4.15	0.99
Residency duration (years)	5.63	0.62	6.55	0.22	-0.92	0.70	0.94
Panel B: Geospatial information about home address							
Elevation (meters)	320.3	1.26	318.8	1.01	1.48	1.53	0.99
In 300 meters proximity:							
Number of restaurants	10.09	1.51	10.09	1.22	0.00	1.85	1.00
Number of supermarkets	1.34	0.18	1.22	0.16	0.12	0.25	1.00
Number of medical facilities	18.73	2.97	11.54	1.83	7.19	2.92	0.28
Number of cultural facilities	0.09	0.05	0.39	0.06	-0.30	0.08	0.02
Number of churches	1.12	0.30	1.42	0.14	-0.30	0.36	0.99
Distance to main station (km)	3.33	0.28	3.87	0.25	-0.54	0.31	0.76
Distance to city hall (km)	2.63	0.32	3.64	0.26	-1.02	0.33	0.09
Distance to main church (km)	2.91	0.30	3.71	0.27	-0.79	0.32	0.28
Distance to airport (km)	4.24	0.27	5.42	0.18	-1.18	0.39	0.10
Time to main station (minutes)	16.51	0.96	20.19	1.25	-3.68	1.22	0.10

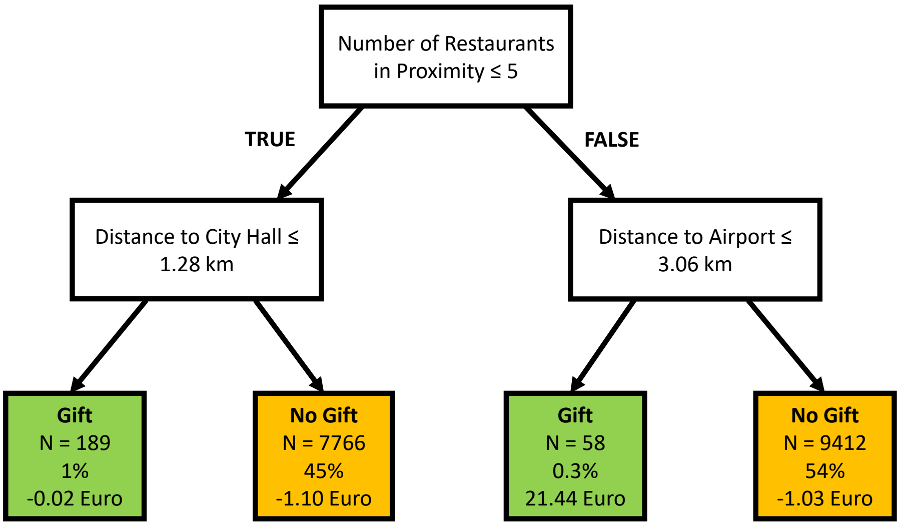
Notes: This table shows the mean characteristics of the groups with the 10% largest and smallest treatment effects in the cold list. We report joint p-values (JP-value), which account for simultaneous inference on several characteristics. We employ the so-called *single-step* methods to control the family-wise error rate (see, e.g., Chernozhukov *et al.*, 2018b, for details). Standard errors are calculated with a multiplier bootstrap using 500 replications. The residency duration in the urban area is censored after 8 years. We measure travel time to the main station using public transportation at 9 : 00am on weekdays.

F Example Decision Tree

Figure F.1: Illustration of decision tree



a) Warm list



b) Cold list

Notes: This figure illustrates one example decision tree. For each terminal leaf, it reports the total and relative number of observations and the effect of the gift on net donations during the first year after the experiment. To derive the decision trees, we estimate Exact Policy-Learning Trees with a search depth of two (Zhou *et al.*, 2018).

G Sensitivity Analysis

Table G.1: Results for alternative estimators in the warm list

	Share treated	Net donations	Optimal targeting vs. all-gift	Optimal targeting vs. no-gift
	(1)	(2)	(3)	(4)
Logit				
Baseline model	0.47	16.19***	0.72	0.78
		(0.90)	(0.81)	(0.82)
Flexible model	0.43	15.45***	-0.02	0.04
		(0.80)	(0.91)	(0.71)
Logit Lasso	0.83	15.86***	0.39	0.45
		(0.91)	(0.62)	(0.98)
Exact Policy-Learning Tree				
depth = 1	0.39	15.05***	-0.42	-0.36
		(0.77)	(0.98)	(0.61)
depth = 3	0.34	15.49***	0.02	0.08
		(0.88)	(0.87)	(0.76)
CART				
depth = 2	0.11	16.10***	0.63	0.68
		(0.94)	(0.93)	(0.69)
Cross-validated depth	0.33	17.40***	1.93**	1.98**
		(0.96)	(0.83)	(0.80)
Classification Forest	0.42	15.88***	0.41	0.47
		(0.83)	(0.89)	(0.73)

Notes: This table shows the results of our robustness checks for warm-list individuals. The outcome variable is the donation amount (euro) during the first year after the gift was sent. Standard errors are in parentheses. ***/**/* indicate statistical significance at the 1%/5%/10% level.

Table G.2: Results for alternative estimators in the cold list

	Share treated	Net donations	Optimal targeting vs. all-gift	no-gift
	(1)	(2)	(3)	(4)
Logit				
Baseline model	0.047	0.15***	0.96***	-0.01
		(0.04)	(0.10)	(0.03)
Flexible model	0.058	0.11***	0.93***	-0.05*
		(0.03)	(0.10)	(0.03)
Logit Lasso	0.0003	0.16***	0.97***	-0.002
		(0.02)	(0.10)	(0.001)
Exact Policy-Learning Tree				
depth = 1	0.07	0.06***	0.87***	-0.10***
		(0.02)	(0.10)	(0.01)
CART				
depth = 2	0.042	0.10***	0.92***	-0.06***
		(0.02)	(0.10)	(0.02)
Cross-validated depth	0.0006	0.16***	0.97***	-0.001**
		(0.02)	(0.10)	(0.0003)
Classification Forest	0.001	0.15***	0.97***	-0.005
		(0.02)	(0.10)	(0.003)

Notes: This table shows the results of our robustness checks for cold-list individuals. The outcome variable is the donation amount (euro) during the first year after the gift was sent. In the cold list, implementing an Exact Policy-Learning Tree with a depth of three is computationally infeasible. Standard errors are in parentheses. ***/**/* indicate statistical significance at the 1%/5%/10% level.