ICIS 2022 Proceedings

Data Analytics for Business and Societal Challenges

Dec 12th, 12:00 AM

The Best of Two Worlds – Using Recent Advances from Uplift Modeling and Heterogeneous Treatment Effects to Optimize Targeting Policies

Jannik Rößler
University of Cologne, roessler@wim.uni-koeln.de

Richard Guse
Karlsruhe Institute of Technology, richard.guse@kit.edu

Detlef Schoder
University of Cologne, schoder@wim.uni-koeln.de

The Best of Two Worlds: Using Recent Advances from Uplift Modeling and Heterogeneous Treatment Effects to Optimize Targeting Policies

Completed Research Paper

Jannik Rößler
Cologne Institute for Information Systems, University of Cologne
Pohligstraße 1, 50969 Cologne
roessler@wim.uni-koeln.de

Richard Guse
Department of Economics and Management, Karlsruhe Institute of Technology
Kaiserstraße 89, 76133 Karlsruhe
richard.guse@kit.edu

Detlef Schoder
Cologne Institute for Information Systems, University of Cologne
Pohligstraße 1, 50969 Cologne
schoder@wim.uni-koeln.de

Abstract

The design of targeting policies is fundamental to address a variety of practical problems across a broad spectrum of domains from e-commerce to politics and medicine. Recently, researchers and practitioners have begun to predict individual treatment effects to optimize targeting policies. Although different research streams, that is, uplift modeling and heterogeneous treatment effect propose numerous methods to predict individual treatment effects, current approaches suffer from various practical challenges, such as weak model performance and a lack of reliability. In this study, we propose a new, tree-based, algorithm that combines recent advances from both research streams and demonstrate how its use can improve predicting the individual treatment effect. We benchmark our method empirically against state-of-the-art strategies and show that the proposed algorithm achieves excellent results. We demonstrate that our approach performs particularly well when targeting few customers, which is of paramount interest when designing targeting policies in a marketing context.

Keywords: Causal inference, individual treatment effects, uplift modeling, heterogeneous treatment effects, targeting policies

Introduction

The design of targeting policies is fundamental to address a variety of practical problems across a broad spectrum of domains from e-commerce to politics and medicine (Gupta et al., 2020). For example, online platforms send different promotions to different customers; politicians use ads, direct mail, and phone calls to push residents to vote for their candidacies; and doctors seek to treat sick patients with the correct treatment.

Marketing is one domain that has benefited significantly from the design of targeting policies (Simester et al., 2019, 2020). Companies such as Spotify (Fernández-Loría et al., 2021), Booking.com (Goldenberg et al., 2020), Uber (Z. Zhao & Harinen, 2019), and Microsoft (Research, 2019) are already designing and deploying targeting policies to identify large-scale marketing actions it can take to reach customers. For example, Booking.com designed a targeting policy that turned an underperforming promotions campaign with over 100 million visit instances into a campaign with a viable return on investment and a significant increase in completed bookings (Goldenberg et al., 2020). Spotify increased the engagement of customers by 28% using a proper targeting policy on a dataset with more than half a billion individuals (Fernández-Loría et al., 2021).

Recently, researchers from two different research communities – *uplift modeling* and *heterogeneous treatment effects* (HTE) – have begun to optimize targeting policies by predicting the causal effect of a treatment at the individual or subgroup level, henceforth referred to as the *individual treatment effect* (ITE) (Ascarza, 2018; Athey et al., 2019; Devriendt et al., 2018; Gubela et al., 2019; Hitsch & Misra, 2018; Simester et al., 2019). For example, Ascarza (2018) showed targeting customers with the highest sensitivity to a treatment in a given marketing campaign to be significantly more effective than targeting customers based on churn probability. Similarly, Hitsch and Misra (2018) demonstrated that predicting the causal effect of a treatment yields significantly greater profits than ignoring the causal effect.

Most literature in these two research streams focuses on the development of new approaches (Gubela et al., 2019). Specifically, researchers have focused on the *direct estimation* of ITEs, that is, on modifying conventional (mostly non-parametric) machine learning algorithms, especially with tree-based algorithms (Athey & Imbens, 2016; Radcliffe & Surry, 2011; Rzepakowski & Jaroszewicz, 2010; Sołtys et al., 2015; Wager & Athey, 2018). Modifying these tree-based algorithms, researchers from both research streams have found promising techniques to increase effectiveness even further (Gubela et al., 2019), such as splitting explicitly on the ITE (Athey & Imbens, 2016; Hansotia & Rukstales, 2002), distributing treatment and control samples evenly in child nodes (Rzepakowski & Jaroszewicz, 2012; Su et al., 2009), and using ensemble learning (e.g., random forest) (Athey et al., 2019; Sołtys et al., 2015). In addition, the uplift modeling literature has demonstrated that avoiding small leaves and considering the values of the parent node when calculating the split criterion reduces variance and overfitting (Rzepakowski & Jaroszewicz, 2012). The HTE literature has shown that using different samples for partitioning the covariate space and estimating the ITE score can substantially reduce bias (Athey & Imbens, 2016).

Despite these promising results and recent developments, however, ITE estimation is rarely employed in real-world applications because of various practical challenges, including weak model performance, a lack of reliability, and high complexity (Fernández & Provost, 2019; Gupta et al., 2020; Hitsch & Misra, 2018). Various researchers have demonstrated that most of the methods still suffer from high variance – not only across different datasets, but also across different cross-validation folds of the same dataset – and that the methods' performance often depends largely on the data and application contexts (Athey & Imbens, 2015; Devriendt et al., 2018; Rößler et al., 2021).

As researchers have begun to integrate uplift modeling and HTE more recently (Rößler & Schoder, 2022; Zhang et al., 2021), one opportunity to overcome the practical challenges that have emerged is to combine recent advances from both research streams and develop new methods that employ the best ideas from both worlds. To the best of our knowledge, however, no literature has developed any methods by combining findings from uplift modeling and HTE. Thus, we ask the following research question (RQ):

RQ: Can we make the prediction of individual treatment effects more efficient by combining recent advances from the uplift modeling and heterogeneous treatment effects literature?

To address this research question, and motivated by recent advances in uplift modeling and HTE literature, we propose a new, tree-based, direct estimation method. Our method has four main characteristics: a splitting criterion that optimizes on the ITE explicitly, takes into account the value of the parent node, and accounts for the sample sizes of treatment and control group in both child nodes; the honest approach, that is, using different samples for partitioning the covariate space and estimating the ITE score; the use of ensemble learning rather than a single decision tree; and the definition of termination rules. We compare our new method with approaches from both research streams using three empirical datasets. We focus on data with a binary treatment indicator (i.e., customers do or do not receive a treatment) and a binary response variable (e.g., customers do or do not stay with the company). Further, we assume that the data are coming from randomized controlled experiments (i.e., A/B tests) or that the data satisfy the unconfoundedness assumption and the stable unit treatment value assumption (Athey & Imbens, 2015).

The remainder of this paper is organized as follows. In the next section, we review the literature on uplift modeling and HTE. We then describe our proposed method, which again is motivated by recent advances in both research streams. This is followed by the presentation of the experimental setup describing the datasets, methods evaluated, and evaluation procedure. We then present our results. Finally, we discuss methodological, theoretical, and managerial implications, and conclude our paper with a summary.

Related Work

As the breadth of causal inference and machine learning applications has grown, research attention has increasingly turned to targeting individuals whose likelihood of response is most positively affected by a treatment (Hitsch & Misra, 2018). Rather than predicting customers' response probability, as is done in response modeling (Coussement et al., 2015), the targeting decisions in uplift modeling and HTE are based on the ITE (Gubela et al., 2017).

The ITE is defined as the difference between an individual's outcome when subject or not subject to a treatment (Hitsch & Misra, 2018). The ITE, however, is not directly observable because of the *fundamental problem of causal inference* (Holland, 1986), which states that only one of the potential outcomes is observed for each customer. To overcome this issue, researchers use data from a randomized, controlled experiment (i.e., A/B test), that is, they split the population into two subpopulations (Ascarza, 2018), a treatment group subject to a treatment (e.g., a marketing treatment) and another disjoint control group not subject to any treatment. Uplift modeling and HTE utilize data from randomized controlled experiments and approximate the ITE by using the *conditional average treatment effect* (CATE). Following the Rubin causal model (Rubin, 1974), the CATE is defined as

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x] \tag{1}$$

where $Y_i(1)$ is the outcome of individual *i* being subject to a treatment,$Y_i(0)$ is the outcome of individual *i* not being subject to a treatment, and X_i is a K-component vector of features. Other causal inference procedures such as using (deep) instrumental variables (Hartford et al., 2017; Syrgkanis et al., 2019), propensity scores (Rosenbaum & Rubin, 1983) or observational data (Nie & Wager, 2021) for ITE estimation are fruitful areas for future research, but are beyond the scope of this paper.

Gutierrez and Gérardy (2017) distinguish between three different classes of ITE methods in uplift modeling and HTE: *two-model*, *class transformation*, and *direct*. *Two-model* methods predict the CATE as the difference between the conditional expectation from two independent models; *class transformation* methods estimate the CATE by using a new, modified outcome variable; and *direct* methods modify an existing supervised machine learning algorithm, with tree-based algorithms in particular, to train on the CATE directly. While researchers agree that evaluating a single algorithm is not enough to improve a targeting policy substantially (Devriendt et al., 2018; Fernández-Loría et al., 2021), recent research shows that direct methods typically outperform two-model and class transformation methods (Devriendt et al., 2018; Guelman et al., 2012; Hitsch & Misra, 2018; Rößler et al., 2021; Sołtys et al., 2015). Thus, our focus in this paper is on direct ITE estimation methods. Table 1 presents an overview of different tree-based methods, including their characteristics and differences.

In the HTE literature, researchers in the areas of statistics and econometrics have mainly proposed to modify nonparametric statistical methods to calculate the CATE directly. For example, Su et al. (2009) proposed the Interaction Tree (IT) to find treatment effects among subgroups of individuals. Their method follows the standard CART (Breiman et al., 1984) convention, but with the so-called G statistic as splitting criterion. Later, Su et al. (2012) provided another algorithm based on the CART (Breiman et al., 1984) convention, the Causal Inference Tree (CIT). The authors proposed to utilize the likelihood ratio test statistic as split criterion. More recently, the Causal Tree / Honest Tree (CT) (Athey & Imbens, 2016) and its extensions Causal Forest (CF) (Wager & Athey, 2018) and Generalized Random Forest (GRF) (Athey et al., 2019) have been proposed to estimate ITEs. All approaches modify the conventional CART (Breiman et al., 1984) procedure in two ways. First, the authors use different samples for partitioning the covariate space and estimating the CATE scores in the leaves. Before growing a tree, they split the training sample into an estimation sample S^{est}, which is used only for CATE score estimation, and a training sample S^{tr}, which is used only for selecting tree splits. The authors refer to this approach as *honesty*. Second, the authors modify the splitting criterion of a conventional CART (Breiman et al., 1984) algorithm such that it penalizes small leaf sizes by incorporating the within-leaf variance. That is, the authors use an estimator for the negative expected mean squared error (EMSE) that depends on S^{tr} and N^{est} (size of the estimation sample S^{est}).

Method	Base	Splitting Criterion	Account for treatment-to-control ratio imbalance	Penalize small leaf size	Account for size imbalance between the children	Consider the value of the parent node	Honesty
IT	CART	G-statistic	No	No	No	No	No
CIT	CART	Likelihood ratio test statistic	No	No	Yes	No	No
CT	CART	Negative expected mean squared error	No	Yes	Yes	No	Yes
CF / GRF	Random Forest						
DDP	CHAID	$\Delta\Delta p$-criterion	No	No	No	No	No
UT	CART	Distribution divergence measures	Yes	Yes	Yes	Yes	No
UB	Bagging						
URF	Random Forest						
IDDP (this paper)	Random Forest	IDDP	Yes	Yes	Yes	Yes	Yes

Table 1. Overview of tree-based methods

Similar to the HTE literature, the uplift modeling literature mainly modifies nonparametric models to directly optimize on the causal effect. One of the first modifications was proposed by Hansotia and Rukstales (2002). The authors suggested using a splitting statistic that maximizes the difference of the differences (DDP) between the treatment and control response probability in the left and right child node using the CHAID (Kass, 1980) convent*ion*. Rzepakowski and Jaroszewicz (2012), changed the CART (Breiman et al., 1984) algorithm by incorporating a split criterion that is based on information theory, using one of three different distribution divergence measures. These include Kullback-Leibler divergence, squared Euclidean distance, and chi-squared divergence. At each split, the goal is to maximize the difference in response distribution between the treatment and control group by subtracting the divergence of the parent node from the conditional divergence of each child node. This method is referred to as Uplift Tree (UT). This splitting statistic was also used with Bagging (UB) (Sołtys et al., 2015) and Random Forest (URF) (Guelman et al., 2012; Sołtys et al., 2015).

Although the direct methods appear to be most promising, various researchers have shown that no method significantly outperforms another on all datasets. Instead, most of the methods lack robustness and suffer from variance and reliability (Athey & Imbens, 2015; Devriendt et al., 2018; Rößler et al., 2021). For example, Devriendt et al. (2018) evaluated ten different methods, including four direct estimation methods, on four real-world datasets. The authors found that no method consistently outperformed the others and, instead, that methods lacked robustness in terms of performance across different datasets. Rößler et al. (2021) compared two-model, class transformation, and direct methods and found that none of them outperformed the others on all datasets. The authors proposed instead to combine approaches from all three method classes to obtain the best results. Researchers have also noted that ITE estimation is rarely employed in real-world applications, not only because of the practical challenges mentioned above but also because of the perception that the ITE prediction may be too difficult or that it may not improve traditional approaches such as response modeling (Gupta et al., 2020; Hitsch & Misra, 2018).

Zhang et al. (2021), most recently, enabled the development of new, potentially more robust methods when they began to unify uplift modeling and HTE literature. The authors provided a unified view of these two research streams, emphasizing their inherent connection and urging researchers and practitioners to consider literature from both communities when optimizing ITE estimation. Further, Rößler and Schoder (2022) undertook an extensive comparison of methods from HTE and uplift modeling using synthetic and real-world datasets, and argued that using the ITE can substantially improve a targeting policy – but only if "academics and practitioners evaluate various methods from both uplift modeling and HTE" (Rößler & Schoder, 2022, p. 1).

Motivated first by the need for more robust solutions and, second, by the recent integration of uplift modeling and HTE, we propose another example of a direct estimation algorithm that builds upon recent advances from both literatures.

Algorithm

In the following, we describe the four main characteristics of our new, tree-based, direct estimation algorithm. First, we elaborate on the splitting criterion of the tree-based algorithm, which optimizes on the ITE explicitly (Hansotia & Rukstales, 2002), takes into account the value of the parent node (Rzepakowski & Jaroszewicz, 2012), and accounts for the sample sizes of treatment and control group in both child nodes (Rzepakowski & Jaroszewicz, 2012; Su et al., 2012). We then describe how we use different samples for partitioning the covariate space and estimating the ITE score, referred to as honest approach (Athey & Imbens, 2016). Third, we explain why we use ensemble learning (i.e., random forest) instead of a single decision tree. Finally, we define the termination rules of our tree-based algorithm.

Splitting Criterion

Hansotia and Rukstales (2002) propose using the ITE explicitly by splitting a decision node and thus maximize the difference of the differences between the treatment and control response probability in the left and right child node:

$$\Delta\Delta p = |E[Y(1) - Y(0)|X\in\Phi_L] - E[Y(1) - Y(0)|X\in\Phi_R]| \tag{2}$$

where Φ is a subset of the feature space associated with the current decision node, and Φ_L and Φ_R are the left and right child nodes, respectively. We refer to this method as *DDP*. Note that equations 2–5, and figures 1 and 2 represent an arbitrary split that can take place on any level in a decision tree. Figure 1 provides an example of the DDP splitting criterion, where the criterion takes its maximum value (i.e., 2).

The problem with the DDP splitting criterion is that it introduces three major issues into the optimization of the ITE. First, the splitting criterion maximizes the absolute gain rather than the incremental gain of the ITE, which potentially introduces overfitting. Consider, for example a scenario in which we have several features to choose from for a specific split to divide the sample into left and right child nodes. Naturally, we would choose to split on a specific feature if it maximizes the gain, and we would reject all other features with a smaller gain compared to the selected feature. We would stop splitting if any split on all of the features resulted in no further improvement. However, the gain achieved with the DDP splitting criterion is always positive, potentially leading to a very deep tree.

Second, directly connected to the first issue, the splitting criterion eventually leads to very small leaves. The method will continue splitting even if the split decreases the ITE. Although the intention can be that a split decreases the ITE in the first place for a much higher increase in the subsequent splits, this eventually leads to leaves of very small sizes (e.g., see the left child node in Figure 1), which in turn results in high variance and overfitting.

Third, the splitting criterion neither takes into account the difference in size of the resulting child nodes nor does it consider the treatment-to-control ratio, as other researchers have noted (Devriendt et al., 2018; Radcliffe & Surry, 2011; Rzepakowski & Jaroszewicz, 2012). For example, imagine that in the left child node in Figure 1, we have only one sample in the treatment group with a positive response. Thus, $E[Y(1) - Y(0)|X\in\Phi_L] = 1$. However, when a split leads to an uneven distribution of control and treatment samples in the child nodes, small populations will be overemphasized (Radcliffe & Surry, 2011) and variance will increase (Y. Zhao et al., 2017b).

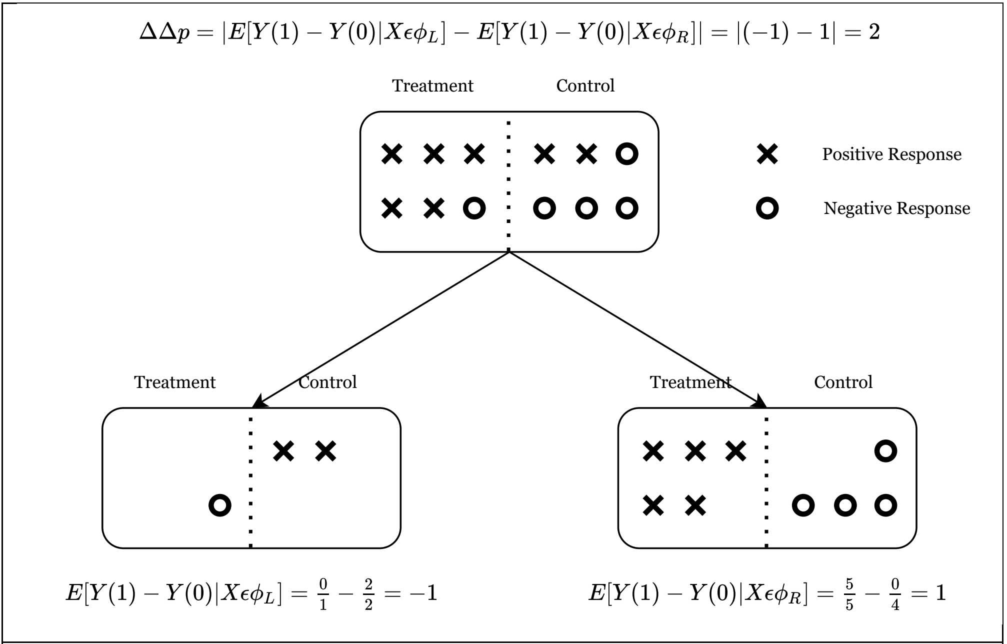

$\Delta\Delta p = |E[Y(1) - Y(0)|X\epsilon\phi_L] - E[Y(1) - Y(0)|X\epsilon\phi_R]| = |(-1) - 1| = 2$
Treatment
Control
Positive Response
Negative Response
Treatment
Control
Treatment
Control
$E[Y(1) - Y(0)|X\epsilon\phi_L] = \frac{0}{1} - \frac{2}{2} = -1$
$E[Y(1) - Y(0)|X\epsilon\phi_R] = \frac{5}{5} - \frac{0}{4} = 1$

Figure 1. Example of the DDP splitting criterion

Drawing from ideas present in uplift modeling (Rzepakowski & Jaroszewicz, 2012), we address the first two issues by considering the ITE value of the parent node in the splitting criterion to reduce variance and overfitting. We eventually want the gain (incremental ITE) to be negative if the split does not exceed the absolute ITE value of the parent node. Thus, we take the absolute ITE value of the parent node and subtract it from the $\Delta\Delta p$ value yielding a possible range between -1 and 2:

$$\Delta\Delta p^{*} = \Delta\Delta p - |E[Y(1) - Y(0)|X \in \Phi| \qquad (3)$$

By subtracting the absolute value of the parent node, we limit the frequency of positive $\Delta\Delta p^*$ values and thus ensure fewer splits. This approach does not perform a split if the difference of the ITE values of the child nodes ($\Delta\Delta p$) is less than the absolute ITE value of the parent node ($|E[Y(1) - Y(0)|X \in \Phi|$). Thus, the higher the ITE value of the parent node, the more difficult it is to find good splits. Thus, we both reduce the likelihood of creating a deeper tree and create leaves with few samples. Henceforth, we refer to this direct approach as a more invariant version of $\Delta\Delta p$, that is, invariant DDP (IDDP).

To address the third issue, we normalized our split by considering the sample sizes of the treatment and control groups in both child nodes, as suggested in the uplift modeling (Rzepakowski & Jaroszewicz, 2012) and HTE (Su et al., 2009) literature. More specifically, we extended the equation from Rzepakowski & Jaroszewicz (2012) such that it has the same range as their divergence functions (i.e., ranging between 0 and 2) by adding one to the $\Delta\Delta p^*$ term, dividing it by three, and multiplying it by two. Let $n_t(\Phi)$ be the number of treatment samples, $n_c(\Phi)$ the number of control samples, and $n(\Phi)$ the number of all samples in the current (parent) node. The first term of Eq. 4 penalizes the imbalance of treatment and control samples in the current parent node by calculating the entropy (H) of the proportion of treatment samples and control samples. The two following terms account for different leaf sizes:

$$\mathrm{I}(\Phi,\Phi_l,\Phi_r) = \mathrm{H}\left(\frac{n_t(\Phi)}{n(\Phi)},\frac{n_c(\Phi)}{n(\Phi)}\right)\, 2\,\frac{1+\Delta\Delta p^{*}}{3}$$
$$+\frac{n_t(\Phi)}{n(\Phi)}\mathrm{H}\left(\frac{n_t(\Phi_l)}{n(\Phi)},\frac{n_t(\Phi_r)}{n(\Phi)}\right)+\frac{n_c(\Phi)}{n(\Phi)}\mathrm{H}\left(\frac{n_c(\Phi_l)}{n(\Phi)},\frac{n_c(\Phi_r)}{n(\Phi)}\right)+\frac{1}{2} \qquad (4)$$

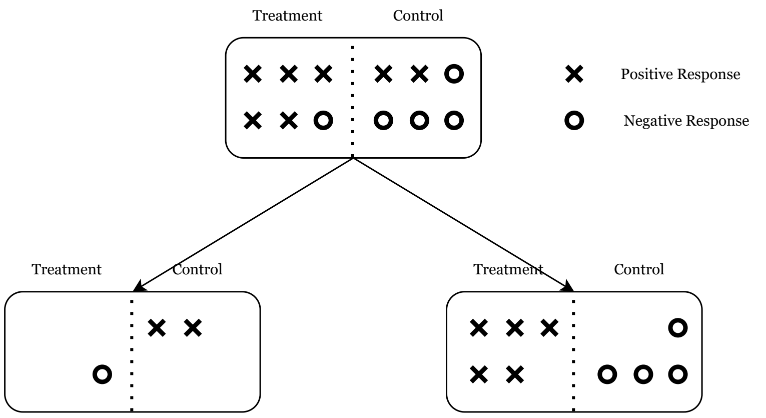
where the entropy is defined as $\mathrm{H(p,q)} = (-p * \log_2(p)) + (-q * \log_2(q))$. Similar to Rzepakowski and Jaroszewicz (2012) we added 0.5 to the last term such that the second part of Eq. 4 is always at least 0.5. The normalization term J(A) punishes unbalanced splits and favors balanced splits. If, however, we do not account for very small values of J(A), we would inflate even a small $\Delta\Delta p^*$ value. For example, consider a $\Delta\Delta p^*$ value of 0.1 and a J(A) value of 0.001. This would result in in an overly inflated IDDP value of 100 (see Eq. 5). We can prevent this by adding 0.5 to the normalization term; instead, the IDDP value would be approximately 2. With Eq. 3 and Eq. 4 taken together, we define our final splitting criterion as:

$$\mathrm{IDDP} = \frac{\Delta\Delta\mathrm{p}^*}{\mathrm{I}(\Phi, \Phi_\mathrm{l}, \Phi_\mathrm{r})} \tag{5}$$

Similar to Zhao et al. (2017a), we include a regularity term to avoid outliers that would affect the ITE estimation. This means that the greater the difference in response rates (i.e., treatment and control response rates) between parent and child node, the more the child nodes' response rates are shifted to the response rates of the parent.

Figure 2 is an example of the IDDP splitting criterion. In contrast to the DDP splitting criterion, the IDDP criterion takes into account all the issues mentioned above. Thus, in the example, the IDDP criterion returns a much lower value because of the imbalance between treatment and control sample size in the left and right child nodes.

$$\Delta\Delta p^* = \Delta\Delta p - |E[Y(1) - Y(0)|X\epsilon\phi]| = 2 - (\tfrac{5}{6} - \tfrac{2}{6}) = 1.5$$



$$E[Y(1) - Y(0)|X\epsilon\phi_L] = \tfrac{0}{1} - \tfrac{2}{2} = -1 \qquad E[Y(1) - Y(0)|X\epsilon\phi_R] = \tfrac{5}{5} - \tfrac{0}{4} = 1$$

Normalization

Imbalance in parent node with respect to treatment and control samples: $H(\frac{n_t(\phi)}{n(\phi)}, \frac{n_c(\phi)}{n(\phi)}) = H(\frac{6}{12}, \frac{6}{12}) = 1$

Imbalance between left and right child node with respect to the treatment samples: $H(\frac{n_t(\phi_l)}{n(\phi)}, \frac{n_t(\phi_r)}{n(\phi)}) = H(\frac{1}{12}, \frac{5}{12}) = 0.8250$

Imbalance between left and right child node with respect to the control samples: $H(\frac{n_c(\phi_l)}{n(\phi)}, \frac{n_c(\phi_r)}{n(\phi)}) = H(\frac{2}{12}, \frac{4}{12}) = 0.9591$

$$\begin{aligned} I(\phi, \phi_l, \phi_r) &= H(\tfrac{n_t(\phi)}{n(\phi)}, \tfrac{n_c(\phi)}{n(\phi)}) * 2^{\frac{1+\Delta\Delta p^*}{3}} + \tfrac{n_t(\phi)}{n(\phi)} H(\tfrac{n_t(\phi_l)}{n(\phi)}, \tfrac{n_t(\phi_r)}{n(\phi)}) + \tfrac{n_c(\phi)}{n(\phi)} H(\tfrac{n_c(\phi_l)}{n(\phi)}, \tfrac{n_c(\phi_r)}{n(\phi)}) + \tfrac{1}{2} \\ &= 1 * 2^{\frac{1+1.5}{3}} + \tfrac{6}{12} * 0.8250 + \tfrac{6}{12} * 0.9591 + \tfrac{1}{2} \\ &= 3.0588 \end{aligned}$$

$$IDDP = \frac{\Delta\Delta p^*}{I(\phi,\phi_l,\phi_r)} = \frac{1.5}{3.0588} = 0.4904$$

Figure 2. Example of the IDDP splitting criterion

Honesty

To reduce bias, we follow the *honesty* approach proposed by Athey and Imbens (2016). The authors write "that a model is honest if it does not use the same information for selecting the model structure … as for estimation given a model structure" (Athey & Imbens, 2016, p. 7353). To accomplish this, they suggest using different samples for partitioning the covariate space and estimating the CATE scores. Before growing a tree, they split the training sample into an estimation sample S^{est}, which is used only for CATE score estimation in the leaves, and a training sample S^{tr}, which is used only for selecting tree splits. Although the partition in training and estimation samples reduces the sample size for training the tree, the authors argue that the reduction in bias offsets this disadvantage.

Random Forest

The high probability of overfitting with a single decision tree is an important factor to consider for a tree-based algorithm (Grimmer et al., 2017; Wager & Athey, 2018). Multiple researchers report excellent performance of tree ensembles to overcome the problem of overfitting and improve generalizability. Thus, we use random forest for our IDDP approach as suggested in uplift modeling (Guelman et al., 2012; Sołtys et al., 2015; Y. Zhao et al., 2017a) and HTE (Grimmer et al., 2017; Wager & Athey, 2018).

Termination Rules

Finally, for IDDP we explicitly define the termination rules based on the splitting criterion and additional rules based on tree depth and leaf size. The algorithm to build the tree does reach a terminal node and does not continue splitting if one of the following criteria is met:

1. There is no split that results in a positive/ non-negative IDDP value.
2. The depth of the decision tree is greater than a user-defined parameter *max_depth*.
3. The terminal node size is smaller than a user-defined parameter *min_samples_leaf*.
4. The number of treatment samples in the terminal node is smaller than a user-defined parameter *min_samples_treatment*.

Experiment

In this section, we present the setup for our experiment on three real-world datasets between the proposed algorithm and other methods from both research streams.

Datasets

We performed the benchmarking on three real-world datasets. The setup was designed to consider different characteristics, such as varying sizes and response rates. Table 2 is an overview of the key properties of each dataset.

Hillstrom

The Hillstrom dataset is an email marketing campaign from MineThatData (Hillstrom, 2008). In line with previous research (Devriendt et al., 2018; Kane et al., 2014), we selected *visit* as the dependent response variable and considered only the promotion for women's merchandise. Overall, the dataset comprises eight covariates, including continuous and categorical features covering consumer behavior (e.g., dollar value spent in the past year, the channels the customer purchased from in the past year) and geospatial data (e.g., zip code) data. Note that we removed duplicates (i.e., identical customers).

Starbucks

The Starbucks dataset comes from a promotional campaign conducted by Starbucks via a reward mobile app (it was made available via the Udacity Data Scientists Nanodegree). The dataset comprises seven covariates, which were all anonymized.

Churn

The Churn dataset compromises private data from a company in Germany with fixed-term contracts, which usually last a year, and auto-renewal. The firm is primarily characterized by two customers groups: customers who renew their contracts regularly, referred to in uplift modeling as *sure things* (Rößler et al., 2021); and customers who do not remain loyal to the company after the minimum contract period but who undergo auto-renewal if not reminded of a forthcoming cancelation, referred to in uplift modeling as *do-not-disturbs* (Rößler et al., 2021). The company established a targeting policy in order to determine the latter from among the remaining customers who can be convinced to stay. The goal was to avoid offering treatments to sure things and do-not-disturbs. While treating former customers would result in additional costs given that these customers stay with the company regardless of treatment, treating do-not-disturbs would result not only in additional costs but could also lead to negative effects, as customers might have stayed with the company had they not been reminded of a forthcoming cancelation.

Over one year, the firm ran several churn prevention campaigns to investigate whether a discount offer for the next contract period would prevent customers from churning and reduce customer turnout. In each trial, customers were selected whose contracts had not yet been canceled but who had the same contractual end date. These customers were divided randomly into control and treatment groups. All customers in the treatment group were subject to one of the following treatments: a discount for the next contract period or a gift. The cost for each treatment, including expenses for sending the offer via mail and the discount / gift itself, was 40€. Customers in the control group received no incentive. After the cancellation date, the company tracked who renewed and who canceled the contract. Overall, the company collected 34 features, including continuous and categorical features covering socio-demographic information, campaign details, and consumer behavior data. The company did not obtain customer relationship information, such as a customer life-cycle value.

Variables	Hillstrom	Starbucks	Churn
Description	Email marketing campaign	Email promotion campaign	Churn prevention campaign
Channel	E-mail	E-mail	Print
Public or private	Public	Public	Private
Number of observations	38,231	126,184	17,114
Number of treatment observations	19,159	63,112	10,193
Number of control observations	19,072	63,072	6,921
Treatment-to-control ratio	1:1	1:1	3:2
Number of covariates (excluding treatment and response variable)	8	7	32
Treatment response rate	16.02%	1.68%	63.02%
Control response rate	11.48%	0.73%	63.50%
Average treatment effect	4.54%	0.95%	-0.48%

Table 2. Overview of datasets used in the experiment

Methods

We compared our proposed direct approach (IDDP) with four state-of-the-art, direct estimation methods from both research streams: ΔΔp criterion (Hansotia & Rukstales, 2002) with random forest (DDP); uplift random forest with Euclidean distance (U-ED) (Sołtys et al., 2015); generalized random forest (GRF) (Athey et al., 2019); and Bayesian causal forest (BCF) (Hahn et al., 2020). Note that the DDP approach was originally intended for use with a single decision tree. However, as a baseline for comparison, we also use random forest for the DDP method.

We evaluated the following hyperparameters and their values for all methods: *n_estimators* [40, 140, 240], *max_depth* [10, 20, 30], *min_samples_leaf* [20, 40, 60], and *min_samples_treatment* [20, 40, 60]. The methods were implemented using AutoUM (Rößler and Schoder, 2022)[1] which is based on causalml (Chen et al., 2020), EconML (Research, 2019); and the Accelerated Bayesian Causal Forest package[2].

Evaluation

We split Hillstrom, Starbucks, and Churn dataset into 80% training and 20% testing while stratifying the treatment and response variable. We then used the training samples and 10-fold cross-validation to select the best hyperparameter setting for each method – again, using stratification with the treatment and response variable. Predictions and qini related metrics – that is, the deciles of the qini curve, the unscaled qini coefficient, and the weighted unscaled qini coefficient – were computed for each method, dataset, and validation fold. We then chose the best hyperparameter settings based on the average unscaled qini coefficient on the validation folds.

Typically, the performance of predictive algorithms is compared in terms of actual versus predicted outcomes (Gubela et al., 2019). However, in ITE estimation, the ground truth is not observed. Consequently, most researchers use a decile-based metric to evaluate the performance of a predictive model (Gubela et al., 2019). The qini curve and the unscaled qini coefficient, metrics typically used by researchers and practitioners (Devriendt et al., 2018; Gubela et al., 2019; Imai & Li, 2021; Rößler et al., 2021), assess the performance by comparing groups of customers rather than individual customers. The qini curve plots the cumulative difference in response rate between treatment and control samples as a function of the number of targeted samples ranked by the method from high to low (Devriendt et al., 2018). The optimal qini curve ranks treatment responder ahead of treatment non-responder in the treatment group and control non-responder ahead of control responder in the control group.

Ideally, the qini curve of an algorithm should achieve increasing, high uplift values in the first quantiles (e.g., 50% uplift in 10% decile, 60% uplift in 20% decile etc.) until the uplift saddles and eventually begins to decrease until it is equal to the ATE (i.e., the uplift value when targeting all customers). The unscaled qini coefficient serves as a single number metric. It is defined as the ratio of the area under the actual qini curve to the area under the diagonal – corresponding to random targeting (Radcliffe & Surry, 2011). In general, the higher the value, the better the model. More specifically, the unscaled qini coefficient defines how the model performs in comparison with random targeting. For example, a value of two means that the model is performing twice as well as random targeting, and a value of one implies that the model is performing as well as random targeting. As the unscaled qini coefficient does not take into account the importance of the first deciles, which is usually of paramount interest in marketing campaigns, we use the following weighted unscaled qini coefficient proposed by Gubela et al. (2017):

$$\text{Weighted Unscaled Qini} = \frac{(0.9 * Q_1 + 0.8 * Q_2 + \cdots + 0.1 * Q_9)}{\sum_i Q_i} \tag{6}$$

where $i = (1, \dots, 9)$ refers to the decile index and Q_i refers to the uplift value of decile i.

Finally, we compared and evaluated the models based on the average performance on the test set. That is, we used the 10 models from cross-validation with the best hyperparameter settings, estimated their scores on the independent test sample, and calculated the average qini curves, average unscaled qini coefficients, and weighted unscaled qini coefficients.

Results

The empirical results comprise the performance evaluations of five ITE estimation methods on three real-world datasets. The performance metrics capture how the marketing campaigns improve by using uplift modeling strategies in terms of qini curve, unscaled qini coefficient, and weighted unscaled qini coefficient. Table 3 illustrates the best hyperparameter settings for each method.

[1] https://github.com/jroessler/autoum

[2] https://github.com/socket778/XBCF

Dataset	Method	Hyperparameter			
		max_depth	n_estimators	min_samples_leaf	min_samples_treatment
Hillstrom	U-ED	10	140	40	20
	DDP	10	240	40	20
	GRF	30	40	40	40
	BCF	10	140	20	20
	IDDP	10	40	40	20
Starbucks	U-ED	10	240	40	40
	DDP	10	240	40	40
	GRF	10	240	40	20
	BCF	10	240	20	40
	IDDP	20	240	60	20
Churn	U-ED	10	140	40	20
	DDP	20	40	40	20
	GRF	10	40	60	20
	BCF	10	140	40	20
	IDDP	10	240	20	20

Table 3. Best hyperparameter for each method and dataset

Figures 3-5 depict model performance in terms of qini curve per ITE method and dataset. These figures show that most of the methods succeeded in outperforming random targeting, which is represented by the diagonal lines, on all datasets. However, BCF deviated from this pattern, as its performance on the Churn dataset was much weaker than random targeting. In this sense, Figure 5 demonstrates that BCF was not a suitable ITE estimation for that specific dataset. Although the performance was better than that of BCF, GRF appears to be another method that proves to be inadequate for the Churn dataset. The corresponding qini curve never exceeded random targeting with substantial margin – except for the last deciles, which are barely relevant for marketing practice (see Figure 5).

Most of the methods did show consistently good results with respect to the first deciles of the qini curve, which is of paramount importance given that it is "common practice in marketing to target only a small subset of customers with a campaign" (Gubela et al., 2017, p. 12). Thus, looking at Figures 3–5, we can see that companies can achieve high response rates while targeting fewer customers. A positive result shown in Figures 3–5 is that the IDDP approach in particular was performing well in the first deciles. In the second decile on the Hillstrom dataset, IDDP outperformed the second-best approach, GRF, by 62% and random targeting by 205% (see Figure 3). In the second decile on the Starbucks dataset, IDDP outperformed the second-best approach, BCF, by 21% and random targeting by 142% (see Figure 4). Finally, in the Churn dataset's first decile, IDDP outperformed the second-best approach, U-ED, by 81% and random targeting by 1600% (see Figure 5).

Figure 5 reveals that targeting the entire customer base on the Churn dataset results in a negative uplift value of -0.48%. Recall that targeting in the Churn dataset costs 40€ per individual, and so targeting all 10,193 individuals totals 407,720€. To turn this underperforming churn prevention campaign into a campaign with a viable effect, the company could use IDDP and target only 10% of the customers in the treatment group, reducing the number of contacts by 90%, from 10,193 to 1,020, while achieving an incremental response rate of 0.85%. Thus, the company can achieve response rates of 64,35% and 63.50% in the treatment and control group, respectively. Overall, the company can save 366,920€ and improve the incremental response rate by 277% (from -0.48% to +0.85%) compared to targeting all customers in the treatment group. Finally, the strategy is in line with the company's expectations. By targeting only 10% of the customers, the firm can omit sending large amounts of discounts to sure things and do-not-disturbs, while reducing the expenses per responder from 37.69€ to 3.72€.

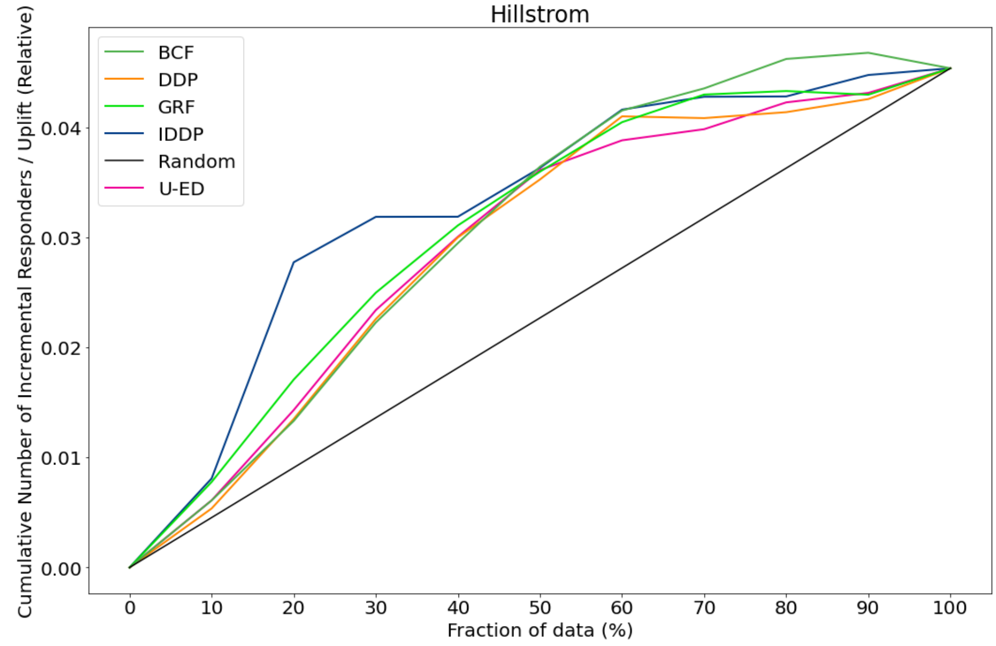


Figure 3. Qini curves on Hillstrom dataset for each method.

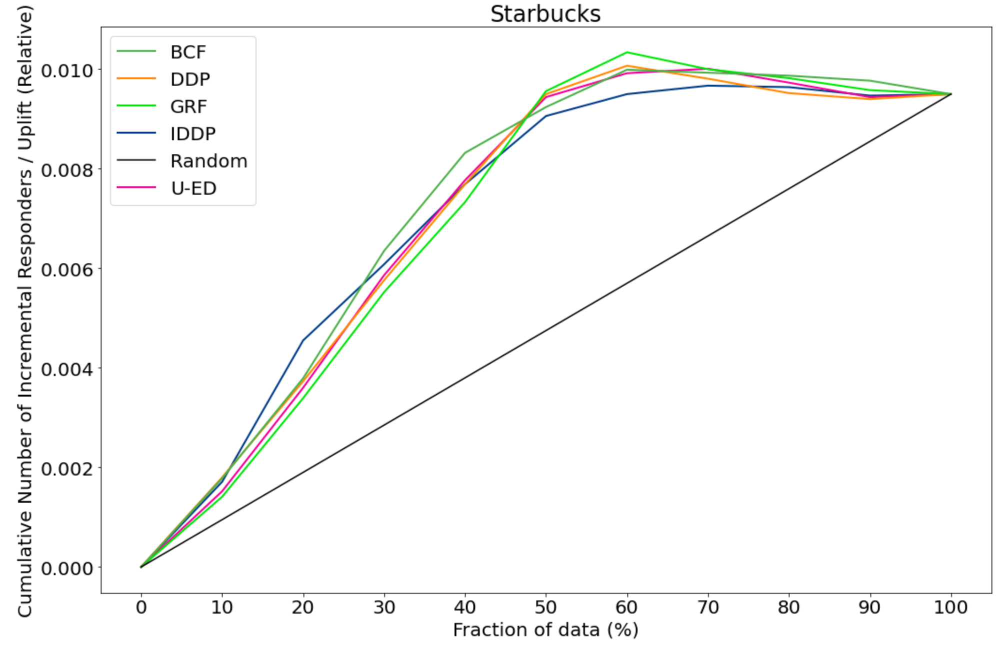


Figure 4. Qini curves on Starbucks dataset for each method.

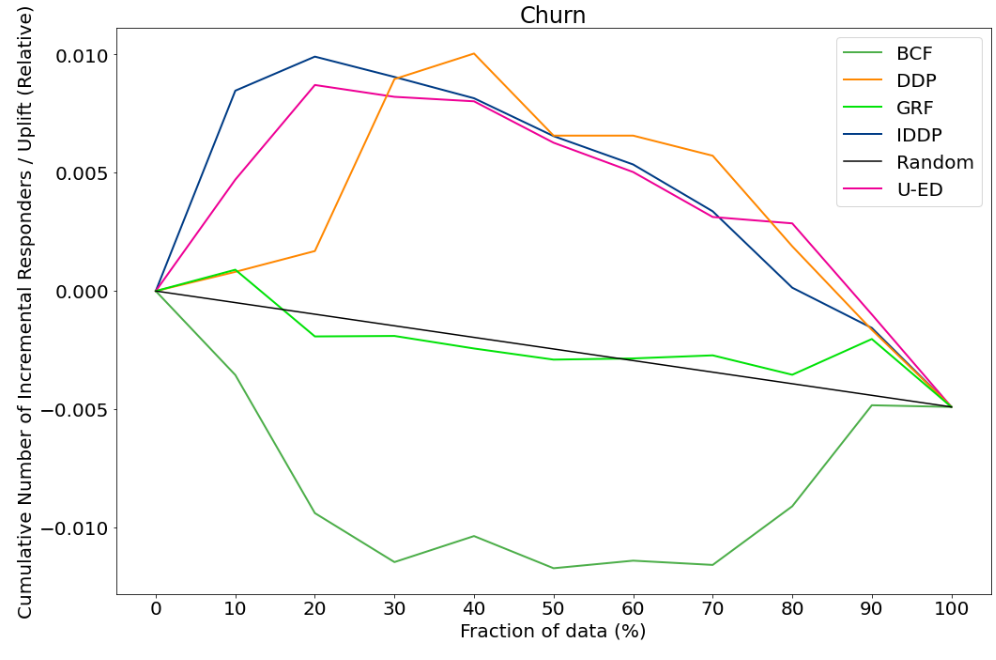


Figure 5. Qini curves on Churn dataset for each method.

To simplify performance comparison of various techniques across the three datasets, Table 4 reports the average unscaled qini coefficients and the weighted unscaled qini coefficients for each method and dataset. Table 4 reveals that according to the average unscaled qini coefficients, IDDP was the best-performing method on Hillstrom and Churn dataset, and the second-best performing method on the Starbucks dataset. On the Starbucks dataset, IDDP was outperformed only by BCF. Although the remaining algorithms mostly performed worse than the IDDP, they achieved good results on all three datasets. Only on the Churn dataset did GRF and BCF perform poorly compared with ED, DDP, and IDDP, highlighting again that neither method is suitable for the Churn dataset.

Method	Metric					
	Unscaled Qini Coefficient			Weighted Unscaled Qini Coefficient		
	Hillstrom	Starbucks	Churn	Hillstrom	Starbucks	Churn
Uplift random forest with Euclidean distance (U-ED)	1.3081	1.5150	*3.7832*	0.4002	0.4101	*0.6164*
ΔΔp criterion with random forest (DDP)	1.3013	1.5148	3.5648	0.3973	0.4134	0.5472
Generalized random forest (GRF)	*1.3635*	1.5079	1.1108	*0.4076*	0.4045	0.4039
Bayesian causal forest (BCF)	1.3593	**1.5518**	-1.5068	0.3894	*0.4145*	0.4934
Invariant DDP (IDDP)	**1.4571**	*1.5167*	**3.9256**	**0.4274**	**0.4179**	**0.6691**

Table 4. Results of the experiment with the average unscaled qini coefficient.

According to the weighted unscaled qini coefficients, IDDP was the best-performing method on all three datasets. It is striking that IDDP is the best-performing method on the Starbucks dataset when using the weighted unscaled qini coefficient, but only the second-best performing method on the same dataset when using the unscaled qini coefficient. However, as described above, IDDP is outperforming the other methods, particularly in the second decile on the Starbucks dataset, resulting in a slightly better weighted unscaled qini coefficient. Overall, however, the remaining methods were also performing very well on most datasets. GRF and BCF achieved worse values than U-ED, DDP, and IDDP only on the Churn dataset.

Note that we also conducted analyses on synthetic datasets similar to Rößler and Schoder (2022). However, our results showed no differences in performance for all methods. We could not find any conditions (e.g., sample size, number of covariates, treatment-to-control-ratio) under which one method was outperforming the others.

Conclusion

In this study, we developed a new tree-based, direct estimation method based on recent advances in uplift modeling and HTE to predict the ITE and optimize targeting policies. Our proposed method has the four main characteristics: a splitting criterion that optimizes on the ITE explicitly, takes into account the value of the parent node, and accounts for the sample sizes of treatment and control group in both child nodes; the honest approach, that is, using different samples for partitioning the covariate space and estimating the ITE score; the use of ensemble learning (i.e., random forest) rather than a single decision tree; and the definition of termination rules. When tested on three datasets, our approach outperformed other state-of-the art methods. In the first deciles in particular, which are of paramount interest when designing targeting policies in a marketing context (Gubela et al., 2017), our approach increased response rates in our campaigns compared to other methods. For example, when targeting only 20% of the customers, our method increased response rates by 62%, 21%, and 81% (first, second, and third dataset, respectively) compared to other approaches.

More generally, our proposed method illustrated that by combining findings and ideas from two related research streams – uplift modeling and HTE – we could tackle major contemporary, practical challenges such as weak model performance and a lack of reliability (Fernández & Provost, 2019; Gupta et al., 2020; Hitsch & Misra, 2018). For example, our method performed very well not only across different cross-validation folds, but also across different datasets. Further, the proposed method is highly relevant for practitioners and analysts in charge of marketing campaigns because it performed extremely well when targeting few customers – a use case that is especially common in marketing campaigns (Gubela et al., 2017). Overall, we demonstrated that researchers from both uplift modeling and HTE have to work together to estimate ITEs in the best possible way.

We are aware that our research has some limitations that serve as excellent avenues for future research. First, although we used different real-world datasets, we were constrained to three datasets to evaluate our approach. Hence, our results might be difficult to generalize. For example, the differences in results between the Churn dataset and the other datasets might be due to the unique characteristics of the former, such as a negative ATE, more covariates, and a smaller sample size. Also, it is striking that on the Starbucks dataset the performance of the IDDP is non-optimal. This may be caused by the specifics of the dataset such as low response rates in the treatment and control groups. This imbalance in response rates can pose challenges for HTE and uplift modeling algorithms. Additional research is necessary to validate our results and to further improve our understanding of the proposed method in different targeting contexts and on different datasets. Further, other scholars should evaluate the proposed approach by conducting a simulation-based analysis to derive boundary conditions – that is, finding circumstances under which the performance differs by, for example, varying response rates in the control and treatment groups or varying number of samples.

Second, while we proposed a method that is theoretically constrained with respect to the type of input data (i.e., binary treatment variable and binary response variable), we invite fellow researchers to extend our proposed algorithm to contexts with continuous response variables and multiple treatment variables.

Third, although we evaluated different methods and various hyperparameter settings, other scholars should extend our analysis by taking into account other methods (e.g., two-model approach (Foster et al., 2011) or class transformation approach (Jaskowski & Jaroszewicz, 2012)) as well as further hyperparameter settings.

Fourth, we used only random forest to evaluate our method. Other scholars should extend the IDDP approach to other ensemble methods, such as bagging and boosting.

Our study has methodological, theoretical, and managerial implications. From a methodological perspective, we have designed a new, effective ITE estimation method, motivated by previous studies on ITE estimation and that uses a tree-based algorithm. This enabled us to improve the performance of current state-of-the-art methods in ITE estimation. We also demonstrated that our method performs especially well when targeting few customers.

From a theoretical perspective, we have shown the power of combining findings and recent advances from two related research streams, uplift modeling and HTE. While both streams have recently been integrated (Zhang et al., 2021), we demonstrated that combining methods and findings from both research streams is not only feasible but also beneficial to researchers. We highlighted that we targeting policies can be optimized by using the inherent symbiotic relationship between uplift modeling and HTE and leveraging prior knowledge. Thus, our study complements most recent research into unifying uplift modeling and HTE in that we demonstrated that both research streams can benefit from each other. Given the complexity of ITE estimation (Gupta et al., 2020), we urge researchers to consider both research streams when designing, evaluating, and optimizing ITEs. The results of this study provide new research opportunities not only for optimizing targeting policies in the domain of information systems and marketing but also for other domains that potentially benefit from optimizing targeting policies, such as political science (e.g., political campaigns) and health care (e.g., estimating drug effects).

From a managerial perspective, our study showed that using the ITE can substantially improve a targeting policy with a proper estimation method. For example, while the Bayesian causal forest method was the best-performing method on one dataset (i.e., Starbucks), it was the worst-performing method on another dataset (i.e., Churn). Based on our results, business managers and practitioners should evaluate a variety of methods from both research streams when optimizing targeting policies using the ITE.

References

Ascarza, E. (2018). Retention Futility: Targeting High-Risk Customers Might be Ineffective. Journal of Marketing Research, 55(1), 80–98. https://doi.org/10.1509/jmr.16.0163

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., & Imbens, G. W. (2015). Machine Learning Methods for Estimating Heterogeneous Causal Effects. Stat, 26.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. The Annals of Statistics, 47(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification And Regression Trees. Routledge. https://doi.org/10.1201/9781315139470

Chen, H., Harinen, T., Lee, J.-Y., Yung, M., & Zhao, Z. (2020). CausalML: Python Package for Causal Machine Learning. ArXiv:2002.11631. http://arxiv.org/abs/2002.11631

Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. Expert Systems with Applications, 42(22), 8403–8412. https://doi.org/10.1016/j.eswa.2015.06.054

Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. Big Data, 6(1), 13–41. https://doi.org/10.1089/big.2017.0104

Fernández, C., & Provost, F. (2019). Causal Classification: Treatment Effect vs. Outcome Prediction. Journal of Machine Learning Research. https://doi.org/10.2139/ssrn.3408524

Fernández-Loría, C., Provost, F., Anderton, J., Carterette, B., & Chandar, P. (2021). A Comparison of Methods for Treatment Assignment with an Application to Playlist Generation. ArXiv:2004.11532 [Cs, Econ, Stat]. http://arxiv.org/abs/2004.11532

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. Statistics in Medicine, 30(24), 2867–2880. https://doi.org/10.1002/sim.4322

Goldenberg, D., Albert, J., Bernardi, L., & Estevez, P. (2020). Free Lunch! Retrospective Uplift Modeling for Dynamic Promotions Recommendation within ROI Constraints. Fourteenth ACM Conference on Recommender Systems, 486–491. https://doi.org/10.1145/3383313.3412215

Grimmer, J., Messing, S., & Westwood, S. J. (2017). Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods. Political Analysis, 25(4), 413–434. https://doi.org/10.1017/pan.2017.15

Gubela, R., Bequé, A., Lessmann, S., & Gebert, F. (2019). Conversion Uplift in E-Commerce: A Systematic Benchmark of Modeling Strategies. International Journal of Information Technology & Decision Making, 18(03), 747–791. https://doi.org/10.1142/S0219622019500172

Gubela, R., Lessmann, S., Haupt, J., Baumann, A., Radmer, T., & Gebert, F. (2017, December 10). Revenue Uplift Modeling. ICIS 2017 Proceedings. https://aisel.aisnet.org/icis2017/DataScience/Presentations/24

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2012). Random Forests for Uplift Modeling: An Insurance Customer Retention Case. In K. J. Engemann, A. M. Gil-Lafuente, & J. M. Merigó (Eds.), Modeling and Simulation in Engineering, Economics and Management (Vol. 115, pp. 123–133). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30433-0_13

Gupta, V., Han, B. R., Kim, S.-H., & Paek, H. (2020). Maximizing Intervention Effectiveness. Management Science, 66(12), 5576–5598. https://doi.org/10.1287/mnsc.2019.3537

Gutierrez, P., & Gérardy, J.-Y. (2017). Causal Inference and Uplift Modelling: A Review of the Literature. Proceedings of The 3rd International Conference on Predictive Applications and APIs, 67, 1–13. https://proceedings.mlr.press/v67/gutierrez17a.html

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). Bayesian Analysis, 15(3), 965–1056. https://doi.org/10.1214/19-BA1195

Hansotia, B., & Rukstales, B. (2002). Incremental value modeling. Journal of Interactive Marketing; Philadelphia, 16(3), 35–46. https://doi.org/10.1002/dir.10035

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. Proceedings of the 34th International Conference on Machine Learning, 1414–1423. https://proceedings.mlr.press/v70/hartford17a.html

Hillstrom, K. (2008). MineThatData: The MineThatData E-Mail Analytics And Data Mining Challenge. https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

Hitsch, G. J., & Misra, S. (2018). Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3111957

Holland, P. W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association, 81(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Imai, K., & Li, M. L. (2021). Experimental Evaluation of Individualized Treatment Rules. Journal of the American Statistical Association, 0(0), 1–15. https://doi.org/10.1080/01621459.2021.1923511

Jaskowski, M., & Jaroszewicz, S. (2012). Uplift modeling for clinical trial data. ICML 2012 Workshop on Clinical Data Analysis.

Kane, K., Lo, V. S. Y., & Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. Journal of Marketing Analytics, 2(4), 218–238. https://doi.org/10.1057/jma.2014.18

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 29(2), 119–127. https://doi.org/10.2307/2986296

Nie, X., & Wager, S. (2021). Quasi-Oracle Estimation of Heterogeneous Treatment Effects. Biometrika, 108(2), 299–319. https://doi.org/10.1093/biomet/asaa076

Radcliffe, N. J., & Surry, P. D. (2011). Real-World Uplift Modelling with Significance-Based Uplift Trees. White Paper TR-2011-1, Stochastic Solutions, 1–33.

Research, M. (2019). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rößler, J., & Schoder, D. (2022). Bridging the Gap: A Systematic Benchmarking of Uplift Modeling and Heterogeneous Treatment Effects Methods. Journal of Interactive Marketing, 10949968221111084. https://doi.org/10.1177/10949968221111083

Rößler, J., Tilly, R., & Schoder, D. (2021). To Treat, or Not to Treat: Reducing Volatility in Uplift Modeling Through Weighted Ensembles. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2021.193

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688–701. https://doi.org/10.1037/h0037350

Rzepakowski, P., & Jaroszewicz, S. (2010). Decision Trees for Uplift Modeling. 2010 IEEE International Conference on Data Mining, 441–450. https://doi.org/10.1109/ICDM.2010.62

Rzepakowski, P., & Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems, 32(2), 303–327. https://doi.org/10.1007/s10115-011-0434-0

Simester, D., Timoshenko, A., & Zoumpoulis, S. I. (2019). Targeting Prospective Customers: Robustness of Machine-Learning Methods to Typical Data Challenges. Management Science, 66(6), 2495–2522. https://doi.org/10.1287/mnsc.2019.3308

Simester, D., Timoshenko, A., & Zoumpoulis, S. I. (2020). Efficiently Evaluating Targeting Policies: Improving on Champion vs. Challenger Experiments. Management Science, 66(8), 3412–3424. https://doi.org/10.1287/mnsc.2019.3379

Sołtys, M., Jaroszewicz, S., & Rzepakowski, P. (2015). Ensemble methods for uplift modeling. Data Mining and Knowledge Discovery, 29(6), 1531–1559. https://doi.org/10.1007/s10618-014-0383-9

Su, X., Kang, J., Fan, J., Levine, R., & Yan, X. (2012). Facilitating Score and Causal Inference Trees for Large Observational Studies. Journal of Machine Learning Research, 13, 2955–2994.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup Analysis via Recursive Partitioning (SSRN Scholarly Paper ID 1341380). Social Science Research Network. https://doi.org/10.2139/ssrn.1341380

Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., & Lewis, G. (2019). Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments. Advances in Neural Information Processing Systems, 32.

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

Zhang, W., Li, J., & Liu, L. (2021). A Unified Survey of Treatment Effect Heterogeneity Modelling and Uplift Modelling. ACM Computing Surveys, 54(8), 162:1-162:36. https://doi.org/10.1145/3466818

Zhao, Y., Fang, X., & Simchi-Levi, D. (2017a, June 30). Uplift Modeling with Multiple Treatments and General Response Types. 2017 SIAM International Conference on Data Mining. 2017 SIAM International Conference on Data Mining, Philadelphia, PA. https://doi.org/10.1137/1.9781611974973

Zhao, Y., Fang, X., & Simchi-Levi, D. (2017b). A Practically Competitive and Provably Consistent Algorithm for Uplift Modeling. 2017 IEEE International Conference on Data Mining (ICDM), 1171–1176. https://doi.org/10.1109/ICDM.2017.157

Zhao, Z., & Harinen, T. (2019). Uplift Modeling for Multiple Treatments with Cost Optimization. 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 422–431. https://doi.org/10.1109/DSAA.2019.00057