Uplift Forest for Multiple Treatments and Continuous Outcomes

Conference Paper · December 2020

CITATIONS
2

READS
1,290

2 authors:

Robin M. Gubela
Humboldt-Universität zu Berlin
13 PUBLICATIONS **84** CITATIONS
SEE PROFILE

Stefan Lessmann
Humboldt-Universität zu Berlin
150 PUBLICATIONS **4,513** CITATIONS
SEE PROFILE

Dec 14th, 12:00 AM

Uplift Forest for Multiple Treatments and Continuous Outcomes

Robin M. Gubela
Humboldt-Universität zu Berlin, robin.gubela@hu-berlin.de

Stefan Lessmann
Humboldt-University of Berlin, stefan.lessmann@hu-berlin.de

Uplift Forest for Multiple Treatments and Continuous Outcomes

Completed Research Paper

Robin M. Gubela
Humboldt-Universität zu Berlin
Spandauer Str. 1, 10178 Berlin
robin.gubela@hu-berlin.de

Stefan Lessmann
Humboldt-Universität zu Berlin
Spandauer Str. 1, 10178 Berlin
stefan.lessmann@hu-berlin.de

Abstract

Artificial intelligence (AI) models support digital commerce through enabling the personalization of communication and services. The paper focuses on predictive models for targeting marketing actions. Extracting information on customers' preferences from big data, such models forecast the reaction of customers to marketing stimuli. Prior research uses models for treatment effect estimation and focuses on a single treatment. In practice, digital marketers have many options to approach customers with different messages and over diverse channels. The support of these decisions requires multiple treatment models. The paper proposes an algorithm for multiple treatments and continuous outcomes. The tree-based algorithm relies on a new splitting criterion, which identifies the optimal treatment by optimizing outcome heterogeneities between treatments. It supports different evaluation measures and experimental designs by incorporating propensity scores. Empirical results on two real-world marketing data sets confirm the effectiveness of our approach compared to state-of-the-art benchmarks.

Keywords: Digital Commerce, Random Forests, Uplift Modeling, Multiple Treatments

Introduction

Digital commerce shapes societies, organizations, and the daily life from billions of people around the globe. Expenditures in digital advertising exceed 100 billion US-dollars (Ha 2019). Beyond reaching a larger audience than traditional mass media, digital advertisements facilitate personalizing marketing communication to individual customers' preferences and attitudes. Artificial intelligence (AI) models enable marketers to extract these preferences from large amounts of customer-centric data. Specifically, marketing employs response models that predict the likelihood of a customer to perform a desired action (e.g., signing a contract, buying a product). Such models are widely used in organizations to predict various outcome variables (e.g., attrition, failure, fraud) for various subjects (e.g., clients, machines, card transactions). We use the term *unit* to refer to the subject the behavior of which is predicted. Response models do not provide a causal relationship between an action (hereafter referred to as *treatment*) and the predicted outcome. Issuing an e-mail with a product promotion to a person might have little impact on his/her buying decision if that person was already planning to buy the product.

In contrast to response models, an uplift model predicts a unit's responsiveness to a treatment. Establishing a causal link between the action and the outcome requires data from a treatment and a control group. Uplift models forecast individual-level treatment effects (ITE), also known as conditional average treatment effects (CATE) (e.g., Knaus et al. 2020) as they depend on a unit's characteristics (e.g., a shopper's browsing behavior). Per unit, an uplift model estimates both the sign and the strength of a treatment's impact on its desired behavior, which allows targeting units according to their relative ITE in decreasing order.

Much uplift research analyzes the effectiveness of a single treatment to increase the likelihood of gaining a dichotomous return (e.g., click-through rates) by forecasting binary outcomes (e.g., Devriendt et al. 2018). Apart from binary outcomes, only a few studies have explored continuous outcomes (e.g., Gubela et al. 2020; Rudaś and Jaroszewicz 2018). Predicting continuous outcomes allows measuring the magnitude of a unit's action, such as a shopping basket value. Continuous outcome settings are typically more in line with business-related metrics. For example, Revenue Qini metrics measure a model's performance in terms of customer expenditures to support targeting decision-making in marketing (e.g., Gubela et al. 2017).

In contrast to the single treatment setting, multiple treatment uplift modeling comprises applications where several treatments co-exist (e.g., Rzepakowski and Jaroszewicz 2012). A unit obtains the treatment that has the highest predicted ITE relative to treatment alternatives. Consider alternative treatments in the form of different marketing incentives (e.g., a promotional message, an informational message, a brand-building message), or different channels to communicate a marketing message (e.g., e-mail, telephone call, in-app targeting). The examples show the availability of multiple treatments in practice. The question is which treatment is most effective for an individual unit. While many approaches exist to address the single treatment uplift problem, the number of multiple treatment uplift approaches is limited. Also, related research focuses on methods to predict binary outcomes. A recent study further introduces propensity score matching (PSM) for multiple treatment uplift modeling with binary outcomes as an algorithm-agnostic data pre-processing step (Olaya et al. 2020).

The paper proposes a novel algorithm for multiple treatments and continuous outcomes, which we call *multiple treatment revenue uplift forest* (MTRUF). Our algorithm grounds on random forests (Breiman 2001) and introduces a new splitting criterion to grow individual decision trees. The splitting criterion maximizes outcome heterogeneities between treatments at a node to identify the optimal treatment per unit. We design MTRUF in such a way that it generalizes previous approaches and extends their applicability. For example, MTRUF supports experimental and observational studies as well as different evaluation metrics. The former feature stems from incorporating propensity scores to correct for selection bias. However, unlike PSM, we do not require the matching part, which has the advantage to avoid losing a significant number of units. Also, the method's use of propensity scores is not limited toward a single evaluation metric.

An empirical contribution emerges from the validation of MTRUF against competitive multiple treatment uplift approaches. Few studies are available using empirical e-commerce data with multiple treatments. The paper employs two real-world data sets based on several e-mail and e-coupon treatments. We demonstrate the method's practical utility to support targeting decision-making by conducting several steps of empirical analysis. We report model results in terms of business outcomes and detail the relative quantities of the proposed treatments, which prior work disregards. Our results confirm the effectiveness of the proposed method.

The paper is organized as follows. Section 2 provides the theoretical background. Section 3 reviews the related literature. Section 4 introduces the proposed method. Section 5 describes the setup of the experimental analysis and reports its results. Section 6 concludes the paper.

Theoretical Background

We first introduce our notation. Let $\boldsymbol{X}_i$ be a vector of covariates that characterize an observational unit $i = 1, \dots, N$. A unit receives a discrete treatment $T_i = t$ from a set of K mutually exclusive treatments, that is, $t \in T = \{0, \dots, K\} \in \mathbb{Z}_0^+$. Note that $T_i = 0$ represents units that either obtained a reference treatment (e.g., a placebo) or no treatment. The literature refers to such units as the control group. A unit with a positive-valued treatment index is part of the treatment group. Let $n_t \in \mathbb{Z}_0^+$ denote the total of units that receive treatment t, which sums up to N across treatments. Let $Y_i \in \mathbb{R}_0^+$ denote a unit's continuous outcome. In contrast to conversion uplift modeling, which is based on binary outcomes, the literature refers to continuous outcomes as revenue uplift modeling (Gubela et al. 2017). Further, let $e(t, \boldsymbol{X_i}) = P(T_i = t|\boldsymbol{X_i})$ for $\forall\, t \in \{0, \dots, K\}$ denote the propensity score that generalizes to multiple treatments and describes a unit's probability to receive treatment t conditional on covariates $\boldsymbol{X_i}$ (e.g., Imbens 2000).

In the observational study, the selection of unobservables is an important concern. The paper focuses on the potential outcome framework. The potential outcomes are a function of the treatments $\{Y_i(t = 0), \dots, Y_i(t = K)\}$. Since a unit is assigned a single treatment, only the outcome corresponding to that

treatment is observed. For example, if $t = 1$, we only observe a (continuous) outcome $Y_i(t = 1)$, while $Y_i(t)$ with $t \neq 1$ are counterfactuals. This issue is referred to as the fundamental problem of causal inference (Holland 1986).

Causal inference under the potential outcomes framework requires three assumptions (e.g., Imbens and Wooldridge 2009). According to the *conditional unconfoundedness assumption*, the potential outcomes need to be independent of corresponding treatments conditional on covariates, that is, $\{Y_i(t = 0), \dots, Y_i(t = K)\} \perp T_i|\boldsymbol{X_i}$ where $\perp$ denotes orthogonality. The *stable unit treatment value assumption* (SUTVA) implies that a unit's potential outcome must be independent of the allocation of a treatment to a different unit. The *overlap assumption* refers to the propensity score $e(t, \boldsymbol{X_i}) \in]0; 1[$ (excluding the bounds). We refer to Lopez and Gutman (2017) for further details.

Experimental settings rely on a randomized controlled trial (RCT) in which the treatments are randomly allocated to units. Hence, an RCT's treatment assignment process is independent from a unit's covariates. The average observed value of a treatment equals its average counterfactual values for units that received a different treatment, which implies that counterfactuals can be disregarded by assumption (Morgan and Winship 2015).

In a single treatment setting, the process of conducting an RCT and comparing outcomes from a treatment and control group to quantify a treatment's effectiveness is known as *A/B testing*. For the multiple treatment setting, *A/K testing* follows the same approach but compares outcomes as induced by multiple treatments (including $t = 0$). Conducting an RCT ensures both conditional unconfoundedness and overlap (e.g., Haupt et al. 2019), which (at least partly) explains its popularity in the uplift literature. Under the potential outcomes framework, and specifically regarding the conditional unconfoundedness assumption, there is no selection on unobservables (Imbens 2000).

In the following, we distinguish between the single and multiple treatment setting considering response and uplift modeling with binary and continuous outcomes. A single treatment response model takes units into account that received a treatment and estimates a binary and continuous outcome for *single treatment conversion response modeling* and *single treatment revenue response modeling*, respectively. Formally, a single treatment revenue response model predicts $E(Y_i(t > 0)|\boldsymbol{X_i})$. From the predictions of treatment units, a corresponding model subtracts the forecasts of control group units, that is, $E(Y_i(t > 0)|\boldsymbol{X_i}) - E(Y_i(t = 0)|\boldsymbol{X_i})$ in the revenue uplift case. Uplift research of the single treatment setting aims to identify whether a treatment yields more benefits than a reference treatment (e.g., Knaus et al. 2020). A large body of the uplift literature focuses on *single treatment conversion uplift modeling* estimating binary outcomes (e.g., Devriendt et al. 2018; Guelman et al. 2015; Kane et al. 2014). Examples include product purchases and click-through rates. Recent studies propose continuous outcome predictions (e.g., Gubela et al. 2020; Rudaś and Jaroszewicz 2018). *Single treatment revenue uplift modeling* facilitates capturing the magnitude of a unit's expected behavioral change in terms of a continuous outcome, for example, customer expenditures.

Multiple treatment research aims to explore performance differences between treatments. *Multiple treatment conversion response modeling* and *multiple treatment revenue response modeling* consider binary and continuous outcomes per treatment, respectively, but disregard the sample from a reference treatment. For illustration, if $t = \{0,1,K\}$, a multiple treatment revenue response model would predict both $E\big(Y_{i,r}(t = 1)\big|\boldsymbol{X_i}\big)$ and $E(Y_{i,r}(t = K)|\boldsymbol{X_i})$ to identify the treatment with the highest predicted value per unit. *Multiple treatment conversion uplift modeling* and *multiple treatment revenue uplift modeling* follow a similar design but incorporate control group data to contrast the treatment-specific forecasts from binary and continuous outcomes. Continuing the example, a multiple treatment revenue uplift model estimates $E\big(Y_{i,r}(t = 1)\big|\boldsymbol{X_i}\big) - E\big(Y_{i,r}(t = 0)\big|\boldsymbol{X_i}\big)$ and $E\big(Y_{i,r}(t = K)\big|\boldsymbol{X_i}\big) - E\big(Y_{i,r}(t = 0)\big|\boldsymbol{X_i}\big)$. As in the single treatment setting, estimating continuous outcomes is more aligned with campaign objectives than binary outcome predictions (e.g., campaign revenues or return on investment). We refer to a corresponding revenue uplift model's treatment-specific outputs $\hat{\tau}_i(t) \in \mathbb{R}$ as ITE. Higher ITE imply higher impacts of a treatment on a unit's estimated response. Zero ITE imply that a treatment has no effect. The treatment with the highest ITE is optimal, that is, $\hat{\tau}_i^*(t) = argmax(\hat{\tau}_i(t = 0), \dots, \hat{\tau}_i(t = K))$. Figure 1 summarizes the single and multiple treatment modeling strategies. *ST* and *MT* refer to the single and multiple treatment setting, respectively. The paper focuses on the multiple treatment revenue uplift modeling setting, which Figure 1 highlights using bold face.

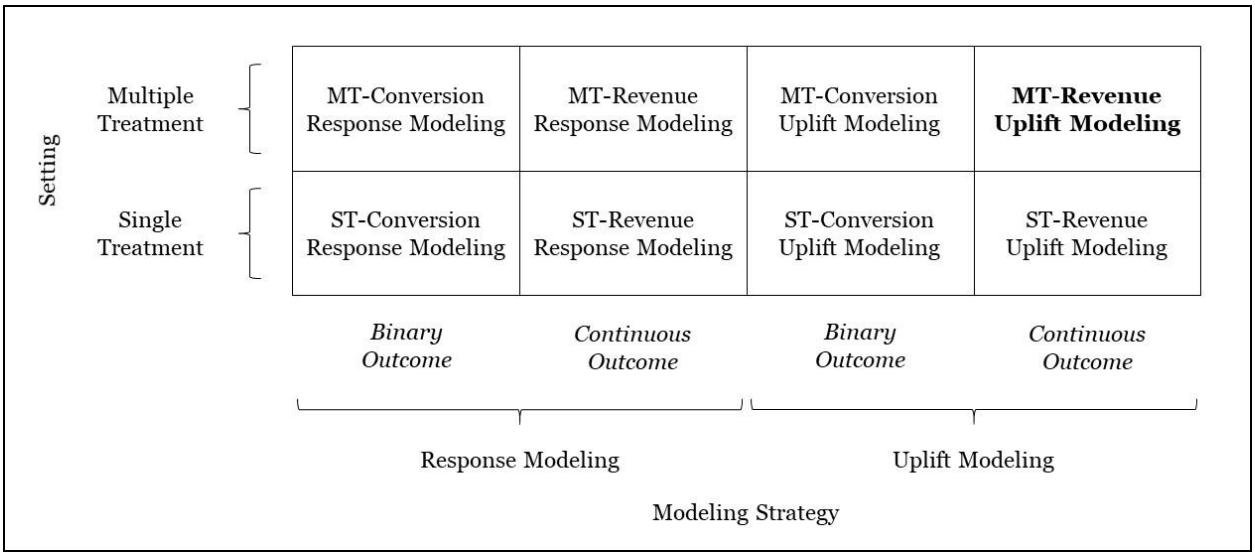

Figure 1. Single and Multiple Treatment Response and Uplift Modeling Strategies

Related Literature

In the following, we develop a taxonomy of model types for multiple treatment uplift modeling. We review both the uplift and causal inference literature to provide a holistic view of the field. Figure 2 systematizes the types of multiple treatment uplift models.

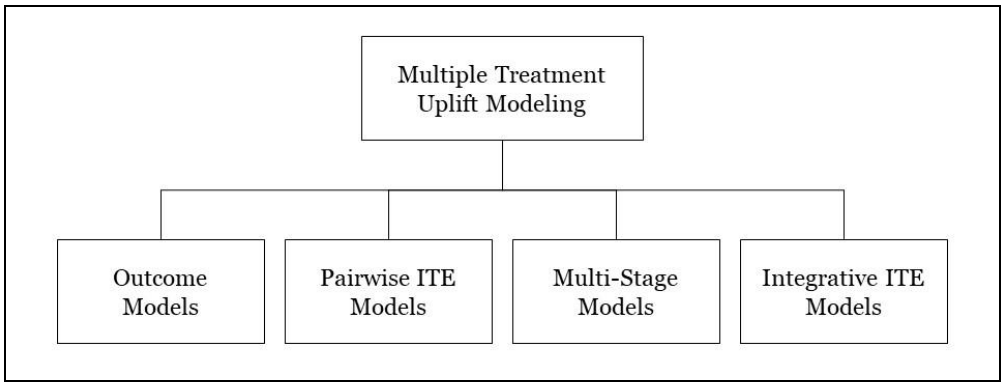

Figure 2. Taxonomy of Multiple Treatment Uplift Model Types

O*utcome models* predict targeted responses and generalize from the single treatment setting to multiple treatments in different ways. Akin to ITE models for single treatments, *pairwise ITE models* output ITE. In contrast to outcome models, they do not require calculating differences between outcome predictions ex post. Pairwise ITE models compare ITE from each treatment against the control group to select the treatment per unit with the largest incremental effect. *Multi-stage models* apply several consecutive predictive models to determine the most beneficial treatment. *Integrative ITE models* are single predictive models that assess each treatment's relative merit and propose the optimal treatment among the alternatives based on the evaluation. The model types imply different levels of effort in terms of model management and monitoring. For instance, while some outcome models require the full range of K models, pairwise ITE models need $K - 1$ models, and integrative ITE models no more than a single model to handle. Going forward, we offer a brief overview of prior work on each type of model.

Outcome models estimate responses per unit and treatment. We identify three instantiations of an outcome model. First, the separate model approach (SMA) (Lo and Pachamanova 2015) generalizes the two model

approach (e.g., Cai et al. 2011) to accommodate multiple treatments. SMA estimates outcomes per treatment and control group and derives ITE ex post by subtracting the treatment-specific forecasts by the control group forecasts. Second, regression adjustment (e.g., Linden et al. 2016) uses a regression model and imputes counterfactual outcomes from treatments that a unit has not received. Average treatment effects are obtained by comparing the counterfactual outcomes among the treatments. Third, doubly robust models estimate outcomes and include forecasts of treatment assignments. Examples include multiple treatment meta-learners (Zhao and Harinen 2020) and causal neural networks (Schwab et al. 2019).

Pairwise ITE models predict ITE by contrasting each treatment consecutively against the control group. The treatment with the highest ITE per test set unit is chosen. Akin to outcome models, an advantage of pairwise ITE models is that available implementations of single treatment algorithms can be used without adaption. For example, Olaya et al. (2020) apply an uplift random forest (URF) for empirical analysis in terms of binary outcomes and label corresponding models as Naive Uplift Approach.

Multi-stage models employ several predictive models. Prior research considers two-stage procedures. Lo and Pachamanova (2015) introduce a first-stage SMA and a second-stage cluster analysis. The heuristic solves a linear optimization problem by maximizing ITE and response rates per cluster and treatment subject to budget and cluster size constraints. Chen et al. (2015) present a first-stage outcome model with interaction effects between the treatments and covariates. The second-stage regularized regression inputs the predicted probabilities, covariates, and a treatment and outputs expected outcomes.

Integrative ITE models are learners that output the most effective treatment per unit after assessing its benefit relative to competing treatments. Integrative ITE models capitalize on lower levels of effort than the above model types as they only require a single model. For binary responses, prior work proposes a transformation to facilitate multiclass classification (Olaya et al. 2020), a causal k-nearest neighbor algorithm (Guelman et al. 2014), and uplift decision trees with information-theoretical divergence measures (Rzepakowski and Jaroszewicz 2012). Regarding the continuous response setting, Imai and Ratkovic (2013) adapt support vector machines by putting separate sparsity (LASSO) constraints to the pre-treatment and heterogeneity parameters. The proposed truncated linear probability model estimates the difference in truncated values of predicted outcomes based on a latent variable's transformed values. Hu et al. (2020) extend Bayesian additive regression trees to accommodate multiple treatments by forecasting conditional responses. Zhao et al. (2017) introduce another decision trees-based integrative ITE model called contextual treatment selection (CTS).

In summary, several types of multiple treatment uplift models exist that differ in their characteristics and levels of effort regarding model management and monitoring. The approach proposed in this paper belongs to the family of integrated ITE models. We employ representatives of other model types as benchmark in the empirical evaluation. Specifically, we consider SMA due to its empirical popularity in prior studies in the field of multiple treatment conversion uplift modeling (e.g., Olaya et al. 2020). An extension to multiple treatment revenue uplift modeling is straightforward. Next, we employ the causal forest (Athey et al. 2019) as a pairwise ITE model to accommodate multiple treatments. As a popular approach to predict continuous responses, the causal forest yields notable empirical performance in the single treatment case (e.g., Gubela et al. 2020). Multi-stage models have not entered previous evaluations (Olaya et al. 2020; Zhao et al. 2017; Zhao and Harinen 2020) and are also not regarded here. Among the integrative ITE models, we select CTS based on its promising performance in marketing settings (e.g., Olaya et al. 2020; Zhao et al. 2017).

CTS is a particularly interesting integrative ITE model that shares methodological similarities with the approach proposed here. For example, it also supports continuous outcomes. To that end, CTS estimates the conditional expectation of a leaf node by using the observed outcome per treatment and the predicted outcome in the parent node of a treatment weighted by a regularization parameter. Zhao et al. (2017) introduce a new evaluation metric to determine the gain from splitting a node, known as *expected response*. Splits are selected that lead to the highest increase of the expected response. CTS multiplies the sample fraction with the highest conditional expectation for the parent node and left and right child nodes, respectively, and subtracts the sum of estimated outcomes from the child nodes by the estimated outcome from the parent node. The algorithm creates a leaf node if a split's gain is negative. Zero-gain nodes are split although splitting does not yield an immediate advantage. Matching is an integral part of the expected response metric. If the predicted treatment matches the observed treatment per unit, its observed response is divided by the treatment probability to get an unbiased estimate of the expected value. Unmatched units are dropped, which might comprise large numbers of observations (e.g., Saito et al. 2020). Zhao et al. (2017)

further propose modified uplift curves that illustrate a model's performance in terms of the expected response as a function of the targeting fraction after matching treatments per unit.

Multiple Treatment Revenue Uplift Forest

We design MTRUF as an integrative ITE model. Compared to other multiple treatment approaches, integrative ITE models are typically more efficient in that they avoid repetitive training of a model for each treatment. Unlike other integrative ITE models that forecast binary responses, MTRUF predicts continuous outcomes. Predicting responses of this scale facilitates quantifying a unit's likely business value conditional on a treatment. Since a marketer is typically interested in top-line growth, we argue that targeting units based on continuous response predictions (e.g., in the form of revenue gains) better aligns uplift model outcomes with business objectives. We detail the design of the proposed approach in the following and clarify methodological differences to CTS.

A general issue in tree learning concerns the high variance from node splitting. Random forests overcome this shortcoming. MTRUF grows a set of trees from bootstrap samples of the training data. The forest prediction is then calculated as a simple average over the individual trees. To increase diversity among trees, random forests determine optimal splits from a random subsample of the covariates, drawn individually for each split. MTRUF has similar characteristics. Like random forests, MTRUF draws bootstrap training samples with replacement. This contrasts bootstrapping procedures of alternative uplift forests such as CTS and URF. MTRUF further facilitates specifying the number of random covariates to consider for node splitting. As with many other regression forests, MTRUF takes the average values across tree estimations of continuous responses for reasons of robustness against outliers and to avoid privileging forecasts from specific trees. MTRUF selects the treatment with the highest corresponding outcome value. To increase computational speed, MTRUF features parallel computing.

MTRUF is an ensemble of individual *multiple treatment revenue uplift trees* (MTRUT), which we detail as follows. Let S denote a candidate split that separates the covariate space $\boldsymbol{X}$ into a left and right subspace, ϕ_ℓ and ϕ_r, respectively. Further, let n_ℓ and n_r denote the number of units belonging to the left and right subspace, respectively. Let $u \in U = \{0, \dots, K-1\} \in \mathbb{Z}_0^+$ and $v \in V = \{1, \dots, K\} \in \mathbb{Z}^+$. Further, let $\widehat{Y}_r(\boldsymbol{X}, t = u) \in \mathbb{R}_0^+$ and $\widehat{Y}_r(\boldsymbol{X}, t = v) \in \mathbb{R}_0^+$ denote the estimates of the conditional expected mean outcomes in a parent node as a function of covariates for a treatment from its corresponding treatment set. Finally, let $\widehat{Y}_r(\boldsymbol{X} \in \phi_\ell, t = u) \in \mathbb{R}_0^+$ and $\widehat{Y}_r(\boldsymbol{X} \in \phi_\ell, t = v) \in \mathbb{R}_0^+$ as well as $\widehat{Y}_r(\boldsymbol{X} \in \phi_r, t = u) \in \mathbb{R}_0^+$ and $\widehat{Y}_r(\boldsymbol{X} \in \phi_r, t = v) \in \mathbb{R}_0^+$ denote the estimates of the conditional expected mean outcomes in a left and right child node, respectively, for treatments $t = u$ and $t = v$ corresponding to the treatment sets U and V. Pairwise comparisons with $u = v$ do not add information to node splitting and are thus avoided.

We conduct several steps to determine the gain $G(S) \in \mathbb{R}$ from splitting a parent node into child nodes. We calculate the difference of continuous outcomes for all pairs of treatments in the parent node and in the left and right child nodes after a split. Furthermore, we weight the continuous outcomes of both child nodes according to their number of units relative to the total number of units N to take their different quantities into account. We derive the gain for a split by calculating the difference between the parent node's estimated outcome and the sum of relative estimated outcomes from the left and right child nodes as follows:

$$\begin{aligned} G(S) = &- \sum_{u=0}^{K-1} \sum_{v=1}^{K} \left(\widehat{Y}_r(\boldsymbol{X}, t = u) - \widehat{Y}_r(\boldsymbol{X}, t = v)\right)^2 \\ &+ \sum_{u=0}^{K-1} \sum_{v=1}^{K} \left(\widehat{Y}_r(\boldsymbol{X} \in \phi_\ell, t = u) - \widehat{Y}_r(\boldsymbol{X} \in \phi_\ell, t = v)\right)^2 * \frac{n_\ell}{N} \\ &+ \sum_{u=0}^{K-1} \sum_{v=1}^{K} \left(\widehat{Y}_r(\boldsymbol{X} \in \phi_r, t = u) - \widehat{Y}_r(\boldsymbol{X} \in \phi_r, t = v)\right)^2 * \frac{n_r}{N} \end{aligned} \tag{1}$$

Comparing outcomes ex post splitting a node with the ex ante outcome reveals whether splitting a node provides a financial value. This is the case if $G(S) > 0$. Per node in the tree, all splits are evaluated and the one with the highest (positive) gain is chosen. To determine splits of continuously-scaled covariates, the

histogram-based estimator considers several split points relative to a variable's distribution. MTRUT creates a leaf node if no split yields a positive gain. Also, a parent node becomes a leaf if the number of treatment-specific units in a child node is lower than the minimal split size. This is to limit a tree's complexity by ensuring that leaves do not contain too few observations. Parent nodes further become leaves if the tree has grown to its maximal depth. A unit is supposed to obtain the optimal treatment from the set of treatments, which is the treatment with the highest estimated outcome value.

Notably, MTRUT follows a different paradigm for split identification than CTS. MTRUT favors the treatment with the highest outcome heterogeneity from pairwise treatment comparisons at a node, whereas CTS selects the treatment with the maximal expected response value. MTRUT assigns units to treatment groups whose outcomes differ significantly. This ensures that a unit obtains the treatment with the highest value difference compared to alternatives. On the other hand, to calculate the estimates of the conditional expected mean outcomes in a child node, MTRUT does not consider the sum of a treatment's observed outcome and predicted response in a parent node weighted by a regularization parameter, as CTS does. As shown in Equation 1, MTRUT follows a different strategy to determine splits by considering parent node predictions for split selection without the need of regularization. In terms of the splitting criterion, secondary differences to CTS include MTRUT discarding cases of $G(S) = 0$ to sustain an improvement in predicted outcomes for further subtrees and provide additional business value.

We highlight further differences of the proposed method compared to the related approach of CTS and other multiple treatment learners. MTRUT generalizes to both experimental and observational settings by using treatment-dependent propensity scores $e(t, \boldsymbol{X_i})$ as the inverse of the corresponding responses. In RCT settings without confounding effects, as prevalent in much uplift research, propensity scores are a valuable asset yielding higher precision of treatment effect estimates (e.g., Williamson et al. 2014). MTRUT's design of using propensity scores contrasts CTS and several other multiple treatment learners, such as SMA, uplift k-nearest neighbors, and uplift decision trees. In contrast to PSM as applied by Olaya et al. (2020), our design to use propensity scores generalizes toward several metrics, namely the expected response and Qini metrics. Moreover, MTRUT differs from CTS in that is does not drop large numbers of observations due to matching. Therefore, MTRUT might perform better than CTS in settings with limited data amounts while sustaining unbiased estimates. Lastly, MTRUT generalizes to different evaluation metrics, that is, Revenue Qini curves and modified uplift curves. An important contribution of Zhao et al. (2017) is the new approach for performance evaluation beyond existing methods such as Qini. On the other hand, Qini metrics still enjoy popularity (e.g., Olaya et al. 2020). We suggest that the ability to consider the expected response and also Qini metrics is an advantage of MTRUT.

Experimental Analysis

The objective of our experiments is to empirically validate the benefit of our proposed approach relative to competitive multiple treatment uplift models. In the following, we clarify the empirical data, choice of uplift models, and evaluation criteria before reporting empirical results.

Data

We use two real-world data sets with e-mail and e-coupon treatments. The Hillstrom e-mail marketing data (Hillstrom 2008) comprises 64,000 observations of online retail customers in the United States of America. The data contains three types of treatments: a men-specific merchandising e-mail, a women-specific merchandising e-mail, and no e-mail. Treatment allocation is based on randomization. A customer has received a treatment with a 33.3% probability. Available covariates quantify a customer's behavioral patterns. More specifically, the data set lists a customer's visit to a particular commercial website, his/her product purchases and purchase volumes in US dollars within a two-week period after executing the campaign. Other variables include the period since the latest purchase and historical expenditures during the prior year from different customer segments for men-related or women-related merchandising items. Also, location-specific information, the distribution channel to complete a product purchase transaction during the prior year, and an indicator of the customer status is available.

The second data set is based on an online commerce retailer's e-coupon campaign targeting customers from several member states of the European Union. Corresponding data is non-disclosed and obtained from an industry collaborator. After fixing variable types, dropping constant-valued variables, handling missing

values, and extracting relevant meta-information about the treatments based on our partner's domain knowledge, the data set contains 69,496 observations and 24 variables. Besides product purchases, spend amounts are captured and measured in euro currency. Customer expenditures have been normalized by our industry partner for reasons of confidentiality. Four variables indicate a coupon discount treatment. Absolute discounts are 10€ and 15€. Percentage discounts are a function of the customer's basket value and have 10% and 15% value, respectively. Another indicator variable refers to control group units without a coupon. Customers receive a treatment with a randomized probability of 73.3% after a maximal number of nine pageviews. The remaining covariates quantify a customer's surfing behavior during the current shop session and from a previous session. Examples refer to the number of views and the time spent on specific shop pages, whether products are added to a shopping cart, and technical aspects.

Models

The empirical study comprises several multiple treatment revenue uplift models. In addition to MTRUF, our experiments include CTS as a natural benchmark, as explained above. We further use a SMA outcome model and choose random forests as base learner. This choice is motivated by the SMA-based random forests' empirical results in the recent benchmark study considering both Qini and expected response metrics (Olaya et al. 2020). The SMA model is a recognized approach in several empirical studies in the field of multiple treatment conversion uplift modeling (e.g., Lo and Pachamanova 2015). Also, we employ the causal forest with honest splitting as a pairwise ITE model, which is a recognized causal method with a sound theoretical fundament. The causal forest has shown strong performance in settings that involve a single treatment, which motivates us to consider it in our multiple treatment setting. To our knowledge, the causal forest has not yet been evaluated in terms of the expected response metric. To facilitate a comparison on equal ground, we specify all forests with 500 trees, and five random candidate covariates per split.

Evaluation Criteria

Assessing uplift models is challenging due to the fundamental problem of causal inference, which states that a unit's actual and predicted outcomes are unobservable per treatment (Holland 1986). We use two metrics to assess performance per targeted decile: Revenue Qini curves for multiple treatments, and modified uplift curves, which we detail as follows.

Qini curves or gains charts for uplift (Radcliffe 2007) quantify the cumulative incremental responses of an uplift model. They are an established uplift metric to assess binary response predictions for single treatment problems (e.g., Guelman et al. 2015). Revenue Qini curves extend Qini curves toward continuous outcomes (Gubela et al. 2017). Revenue Qini curves illustrate an uplift model's performance according to its predicted vector of ITE, which is derived by employing a revenue uplift model. The vector is sorted in decreasing order so that units with high ITE receive a higher position than units with low ITE. This is because units with high ITE are likely responding favorably to a treatment by spending larger amounts. Assuming that units with similar ITE will behave similarly, the sorted vector is divided into ten similar-sized buckets, which are referred to as *targeted deciles* for decision-support. Uplift models achieve identical results by targeting none or the whole population, which corresponds to deciles 0 and 10, respectively. Revenue Qini curves are incremental as they consider both treatment and control groups to determine ITE. Further, they are cumulative as the decile-wise values are calculated by including values from preceding deciles. While recent research develops Qini curves for multiple treatment conversion uplift models (e.g., Zhao and Harinen 2020), we extend Qini curves toward assessing continuous outcomes from multiple treatments and express them in terms of cumulative incremental revenue. We use decile-wise averages across the treatments, which we contrast against their control group counterparts. Similar to Radcliffe (2007) and Rzepakowski and Jaroszewicz (2012) in the single treatment case, we add control group weights based on decile-specific numbers of treated and control units to scale toward a quantity as follows:

$$R(\tau) = \sum_{t=1}^{K} \hat{Y}_{r,k,\delta} - \hat{Y}_{r,\delta}(t=0) \frac{\sum_{t=1}^{K} N_{r,k,\delta}}{N_{r,\delta}(t=0)} \; \forall t \in T \tag{2}$$

Let δ denote a targeted decile. Furthermore, N_δ denotes the number of units corresponding to a decile, $\sum_{t=1}^{K} \hat{Y}_{r,k,\delta}$ and $\sum_{t=1}^{K} N_{r,k,\delta}$ as well as $\hat{Y}_{r,\delta}(t=0)$ and $N_{r,\delta}(t=0)$ denote the estimated conditional expected

mean outcomes per decile and the corresponding quantities of units as part of the treatment and the control group, respectively.

A recent proposal to evaluate multiple treatment uplift models refers to modified uplift curves, which are based on the expected response metric (Zhao et al. 2017). The metric matches the treatment per unit as proposed by a multiple treatment revenue uplift model with the observed treatment and drops the remaining treatment alternatives. Matched units are pipelined to further analysis. As an example, MTRUF fulfills the following condition to identify the optimal treatment T_i^* in terms of expected responses for continuous outcomes, weighted by the inverse of the propensity score:

$$\mathrm{T_i^*} = \arg\max_{t=0,\ldots,\mathrm{K}} \frac{E(Y_i|\boldsymbol{X_i}, t)}{e(t, \boldsymbol{X_i})} \mathbb{I}\{T_i = t\} \; \forall t \in T \tag{3}$$

The Iversion bracket $\mathbb{I}(\cdot)$ has a value of one in case of a treatment match. Modified uplift curves sort the differences of expected responses between treatment and control groups in decreasing order and assign the most suitable units to the first targeted deciles. They further consider the amount of outcome from control group units per decile. To this end, the curves do not start at a revenue value of zero (because outcomes from control group customers are already captured before targeting a subpopulation) and end in different values for the same reason.

In the uplift literature, Qini curves are well-known and widely used to assess model performance for single and multiple treatment cases. Modified uplift curves are a recent solution to address multiple treatment problems. Another difference to Qini curves is that modified uplift curves correct for bias by using treatment matching. Therefore, modified uplift curves are a better choice for observational studies, in which covariates affect both response and treatment variables. We remark that the proposed approach features propensity scores. To this end, it is not bound to an evaluation based on modified uplift curves to alleviate selection bias. MTRUF provides unbiased estimates for both the Qini and expected response metrics, which other approaches might not (e.g., SMA). Despite the modified uplift curves' advantage of bias correction, the treatment matching implies a loss of a signification number of (unmatched) observations, which applies to both experimental and observational settings. In contrast, Qini curves consider all observations to support targeting decision-making. In experimental settings, where covariates do not impact the treatments due to randomized treatment allocation, the relative advantage of Qini curves becomes clear. Both metrics evaluate unbiased estimates. However, Qini curves regard all observations, while modified uplift curves drop significant shares. Lastly, Qini curves assume that the treatment units with the highest estimated scores behave similarly to their control group analogs. However, exact counterparts cannot be observed, and corresponding units might have divergent characteristics (Rzepakowski and Jaroszewicz 2012). This also applies to modified uplift curves that similarly rank units. Based on this discussion, we consider both metrics to evaluate the multiple treatment revenue uplift models.

We draw ten bootstrap folds with replacement per data set. Per fold, we randomly split the data into a 70% training and a 30% testing partition, respectively. The procedure helps us increase our results' robustness in terms of standard errors of the mean values across the folds, which we calculate by dividing the corresponding standard deviation by the square root of the underlying number of units. The analysis is implemented in R and executed using a compute server with 288 gigabytes of random-access memory and 24 cores at 3.4 gigahertz featuring parallelization to speed up the process.

Results

The first part of the subsection clarifies the individual-level treatment effect distributions per data set, treatment and model. The second part reports the results from the Qini analysis. The third part exposes the results in terms of the expected response metric and clarifies the proposed treatments' relative quantities.

Before examining likely outcomes from customer targeting, we study how much assigning a treatment would alter a customer's spending behavior conditional on covariates. Thus, we analyze the predicted ITE distributions of the different treatments for the MTRUF, CTS, causal forest, and SMA (RF) models, as the difference between the predicted values for a treatment and a control group. Figure 3 presents the model-specific distributions across the bootstrap folds per data set and treatment. The dashed vertical lines denote the average values per treatment.

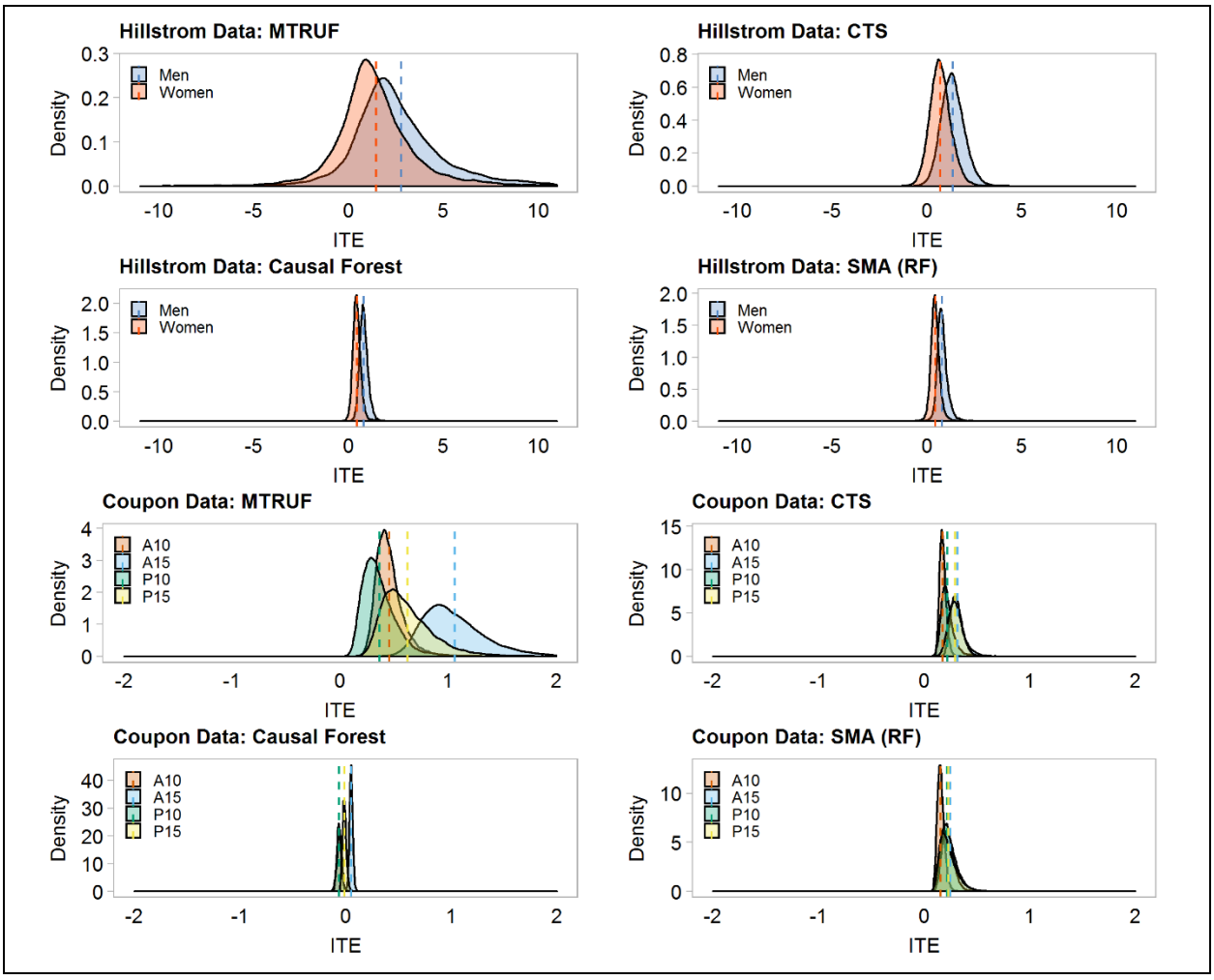

Figure 3. Predicted ITE Distributions Per Data Set, Model and Treatment

Figure 3 suggests that the model curves peak above an estimated ITE value of zero per data set, which implies that the treatments positively affect many customers. The average predicted ITE values per model and treatment confirm this observation. The outcomes from the treatment groups are, on average, higher than the control group outcomes. We justify this result in light of the previous campaign's revenue amount on the full data set per treatment/control group. Regarding the Hillstrom data, for example, the absolute (relative) revenue regarding the men, women, and control group is $30,311.69 ($0.47), $23,038.11 ($0.36), and $13,908.33 ($0.22), respectively. Second, we recognize treatment-dependent differences in terms of the levels of estimated ITE per data set. Considering the Hillstrom data, the men-specific treatment yields higher predicted ITE values than the women-specific treatment, which the ITE distributions and average values demonstrate. Therefore, the men-specific treatment is more effective in altering customer spending behavior than the women-specific treatment, primarily in the case of MTRUF. Regarding MTRUF and the coupon data, corresponding differences between treatments are particularly high. Allocating 15€ coupons yields the highest treatment effects, followed by 15% coupons. Third, we discern model-specific differences of the ITE distributions. Considering the Hillstrom data, MTRUF returns a mean (median) value of 2.11 (1.72) while CTS has 1.03 (0.99), the causal forest has 0.62 (0.60), and SMA (RF) has 0.63 (0.60) across treatments. MTRUF and CTS have higher value ranges of estimated ITE than the causal forest and SMA (RF). MTRUF's estimated ITE distributions per treatment significantly differ in terms of the coupon data compared to the other three multiple treatment revenue uplift models. We explain these remarkable differences considering the proposed algorithm's splitting criterion, which selects the treatment with the

highest outcome heterogeneity from pairwise treatment comparisons at a node. The other models do not focus on maximizing outcome heterogeneities between treatments.

In contrast to the distribution analysis of estimated individual-level treatment effects, Revenue Qini curves facilitate identifying customers who most likely respond positively to treatment with considerable spend volumes per targeted decile. Figure 4 illustrates model performance according to this metric per data set. A curve represents a model's cumulative incremental revenue, which we calculate as an average across the ten bootstrap folds per targeted decile. To increase the robustness of our results, the shaded error bars visualize the corresponding mean's standard error per decile. We further add a linear (grey) line per data set that represents the results from random targeting.

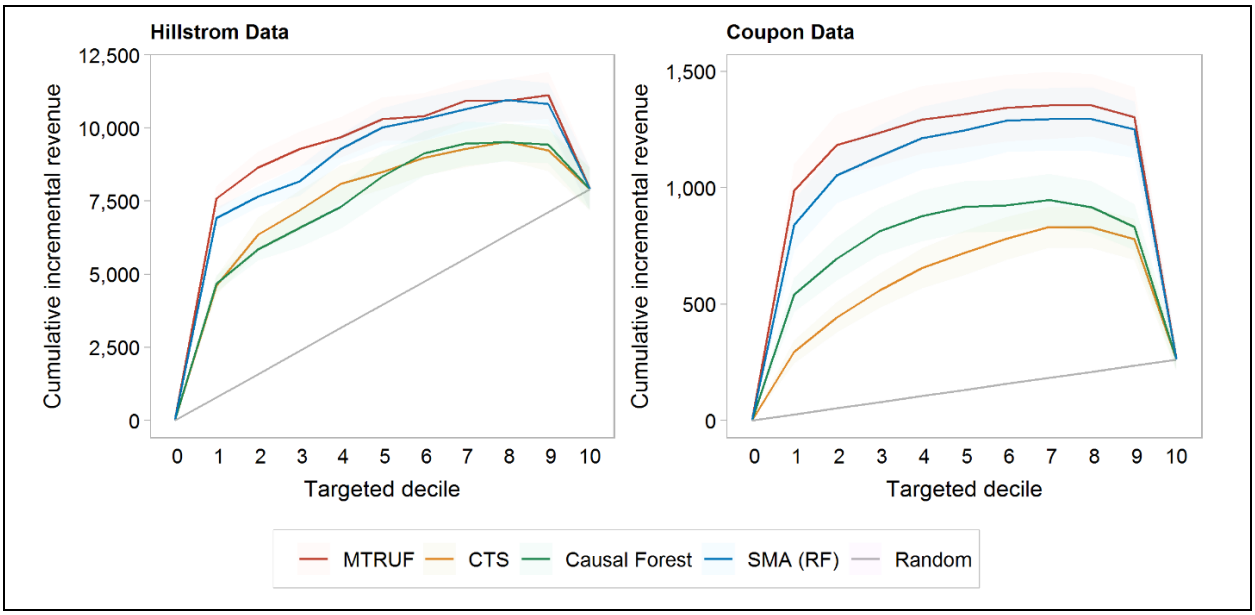

Figure 4. Cumulative Incremental Revenue Per Data Set, Model and Decile

We derive two findings from Figure 4. First, all models increase cumulative incremental revenue with the number of targeted deciles and significantly outperform the random targeting baseline. The results of both data sets confirm that the increases are exceptionally high in terms of the first decile, which indicates that the models generate a significant amount of financial value from targeting top-decile customers. Resource restrictions support the relevance of this decile for campaign management. It is also noteworthy that the models face deteriorating revenue amounts in terms of the last decile. The difference of cumulative incremental revenue between any of the deciles 1-9 and the last decile is particularly high in the coupon data. This suggests that the models succeed in targeting fewer customers with higher business value than targeting the whole population. Second, we recognize different model performances. For both data sets, MTRUF and SMA (RF) realize higher revenue amounts than CTS and the causal forest. Differences are most prevalent regarding the first and last deciles, which suggests that MTRUF and SMA (RF) target the top-fraction of customers who are likely spending considerable amounts. This finding further demonstrates that the models assign customers to the last decile with low or negative treatment effects. Compared to SMA (RF), MTRUF achieves better results across deciles (except for the Hillstrom data's eighth decile) with some remarkable differences in the first three deciles. Hence, we identify MTRUF as the most competitive multiple treatment uplift model considering the Qini metric. The causal forest outperforms CTS in terms of the coupon data. Regarding the Hillstrom data, CTS exceeds the causal forest from the second to the fifth decile and performs slightly worse than the causal forest in terms of the sixth, seventh, and ninth deciles. CTS does not generate as much cumulative incremental revenue as SMA (RF), which (partly) contradicts the few previous findings in the field of multiple treatment conversion uplift modeling (Olaya et al. 2020; Zhao and Harinen 2020). The standard errors support the robustness of these findings.

Figure 5 presents modified uplift curves to assess a model's performance per data set and decile based on the expected response (revenue) after treatment matching per unit. We plot shaded error bars to visualize

the standard errors across bootstrap folds. Control group outcomes are part of the expected response calculations, which explains the higher revenue values on the vertical axis compared to Figure 4.

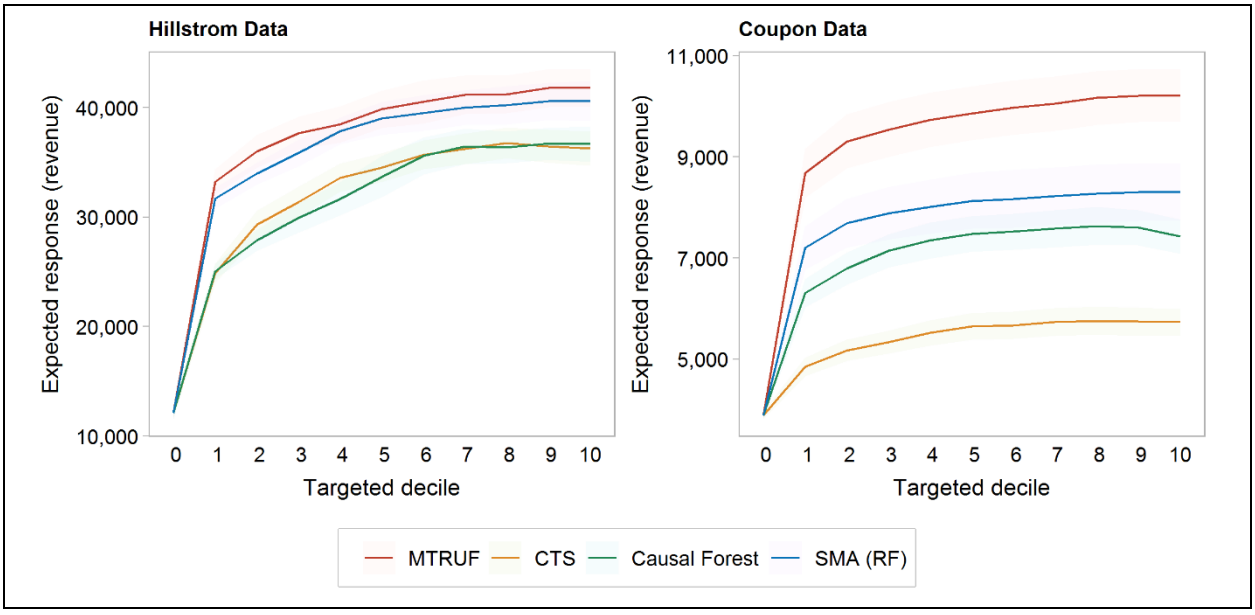

Figure 5. Expected Response (Revenue) Per Data Set, Model and Decile

The results from Figure 5 confirm the results of the Qini analysis. The expected response increases with the number of deciles, especially considering the Hillstrom data. Per data set, the highest increase of expected response refers to the first targeted decile. This observation emphasizes the advantage of employing a multiple treatment revenue uplift model to support campaign decision-making with resource constraints. Also, the models gain different levels of expected revenues. Overall, MTRUF achieves superior performance per data set. Performance differences of MTRUF relative to the other models are particularly pronounced regarding the coupon data. Without exception, MTRUF yields significantly higher business value than its competitors per bootstrap fold and decile. Regarding the Hillstrom data, MTRUF's differences in expected response against the second-best model, the SMA (RF), are less substantial compared to the coupon data. MTRUF and SMA (RF) notably outperform the other models when targeting ten percent of test set customers. This underlines their relative merit to identify the small fraction of customers most likely willing to generate high revenues due to being targeted if the predicted and observed treatments match. We justify the high performance of MTRUF and SMA (RF) considering the same reasons as discussed in terms of the Qini results. Akin to previous findings, the causal forest yields higher and similar performance compared to CTS in terms of the coupon and Hillstrom data, respectively. In the latter case, CTS outperforms the causal forest by targeting customers from deciles 2-5. The business advantage of SMA (RF) compared to CTS in terms of the expected response metric (at least marginally) contradicts prior research (Olaya et al. 2020; Zhao et al. 2017). The standard errors sustain the reliability of our findings.

As Figure 5 does not display the distinctions between the treatments in terms of the expected response metric, we analyze the relative quantities of proposed treatment allocation per data set, model, treatment, and decile in the following. Table 1 presents the corresponding results.

Table 1 discloses the following insights. Regarding the Hillstrom data, the models prefer to assign the men-specific treatment rather than the women-specific treatment or no treatment. In terms of the coupon data, the models generally emphasize allocating the 15€ treatment per decile. In contrast to SMA (RF) and CTS, this treatment is MTRUF's and the causal forest's predominant choice over alternative treatments. This finding particularly supports the corresponding results of MTRUF, as reported in Figure 3. Apart from the 15€ treatment, SMA (RF) and CTS suggest allocating high numbers of 15% coupons to customers. Also, SMA (RF) further prioritizes 10% coupons. MTRUF targets significantly fewer customers using a percentage coupon than the other models. We relate these findings to the preferred treatment's higher revenue levels compared to the alternative treatments and the control group.

Data Set	Model	Treatment	Quantities of Proposed Treatment Allocation Per Decile (%)									
			1	2	3	4	5	6	7	8	9	10
Hill-strom Data	MTRUF	Men	6.2	12.3	18.1	24.0	29.8	35.5	41.5	47.5	51.5	51.5
		Women	3.8	7.7	11.9	16.0	20.2	24.5	28.5	32.5	34.7	34.7
		Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.8	13.8
	CTS	Men	6.7	13.1	19.4	25.6	31.5	37.4	43.2	49.2	54.1	55.0
		Women	3.3	6.9	10.6	14.4	18.5	22.6	26.8	30.8	33.7	34.1
		Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	10.9
	Causal Forest	Men	7.2	14.7	22.1	29.1	35.9	42.6	49.3	56.1	62.7	66.9
		Women	2.8	5.3	7.9	10.9	14.1	17.4	20.7	23.9	27.3	29.3
		Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.8
	SMA (RF)	Men	7.0	14.1	21.2	27.8	34.3	40.8	47.6	54.5	61.6	65.9
		Women	3.0	5.9	8.8	12.2	15.7	19.2	22.4	25.5	28.3	29.6
		Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	4.5
Coupon Data	MTRUF	A_10	1.0	1.9	2.8	3.8	4.9	6.0	7.5	9.1	9.6	9.6
		A_15	6.2	13.1	20.5	27.8	35.2	42.5	49.6	55.9	58.0	58.0
		P_10	1.1	1.9	2.5	3.0	3.5	4.0	4.4	4.8	5.0	5.0
		P_15	1.7	3.1	4.2	5.4	6.5	7.5	8.6	9.8	10.3	10.3
		Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	7.2	17.2
	CTS	A_10	0.7	2.0	3.2	4.5	6.0	7.3	8.6	9.8	10.8	11.0
		A_15	4.1	7.3	10.8	14.5	18.0	21.6	25.3	29.0	32.3	33.0
		P_10	2.1	3.8	5.4	6.9	8.5	10.2	11.8	13.5	14.9	15.1
		P_15	3.1	6.9	10.5	14.1	17.5	20.9	24.3	27.8	30.9	31.6
		Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	9.4
	Causal Forest	A_10	0.2	0.3	0.4	0.6	0.7	0.9	1.0	1.2	1.4	1.6
		A_15	5.5	12.8	20.7	28.7	36.7	44.8	52.6	60.2	67.3	71.9
		P_10	1.7	2.5	3.2	3.8	4.3	4.8	5.3	5.8	6.4	6.9
		P_15	2.7	4.4	5.7	7.0	8.3	9.6	11.1	12.8	14.9	16.5
		Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.1
	SMA (RF)	A_10	0.8	2.5	4.3	6.3	8.7	11.6	14.6	17.4	18.5	18.5
		A_15	3.1	6.0	8.9	11.7	14.7	17.8	21.1	24.8	28.0	28.1
		P_10	3.0	6.5	9.8	13.0	15.7	17.9	19.9	21.7	23.2	23.3
		P_15	3.2	5.1	7.0	9.0	11.0	12.8	14.5	16.2	17.9	17.9
		Control	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.5	12.2

Table 1. Treatment Allocation Quantities Per Data Set, Model, Treatment, and Decile

We re-iterate that the expected response difference of MTRUF against the other models is most substantial in terms of the coupon data's first targeted decile. Here, MTRUF allocates the 15€ coupon to 12.7%, 51.2%, and 100% more customers than the causal forest, CTS, and SMA (RF). At the same time, MTRUF targets much fewer first-decile customers using a percentage discount compared to the remaining models. Considering the Hillstrom data's first decile, MTRUF and SMA (RF) target customers with higher expected outcomes while not allocating significantly different quantities of specific treatments compared to CTS and

the causal forest. A final note refers to the last two deciles per data set. While MTRUF and CTS provide fewer men-specific and women-specific e-mails accordingly, the causal forest and SMA (RF) models allocate (especially men-specific) e-mails to large subpopulations in terms of the Hillstrom data. Only 3.8% and 4.5% of customers do not obtain a treatment in the tenth decile according to the causal forest and SMA (RF), respectively. Corresponding values of MTRUF and CTS are 13.8% and 10.9%. Regarding the coupon data's last two deciles, MTRUF allocates remarkably fewer treatments to customers than the other models, followed by SMA (RF). The results show that refraining from targeting many corresponding customers yields a business value.

From the analyses, we learn that MTRUF is the most effective multiple treatment revenue uplift model. It outperforms several competitive approaches, namely, CTS (an integrative ITE model), the causal forest (a pairwise ITE model), and the separate model approach with random forests (an outcome model). In terms of the coupon data, for example, it allocates large numbers of 15€ coupons to appropriate subpopulations while de-prioritizing alternative treatments with percentage discounts. Regarding the last two deciles, it allocates significantly fewer coupons to customers than the other models. We attribute the encouraging results to MTRUF's methodological design. In contrast to the causal forest and SMA (RF), MTRUF is an integrative ITE model that explicitly controls for all treatments to optimize constructing specific types of trees. A critical difference to CTS is that the proposed algorithm's splitting criterion selects the treatment with the highest outcome heterogeneity at a node compared to treatment alternatives. CTS achieves mediocre results, which might result from the characteristics of its splitting criterion. For example, it allows splitting nodes with zero gain, which induces no further value improvement during subtree construction. Also, we note that CTS is not optimized toward Revenue Qini curves. Like CTS, the results show that the causal forest is not as competitive as MTRUF and SMA (RF). We remark that the original causal forest has not been developed toward multiple treatment problems. As a pairwise ITE model, it does not iteratively adapt its predictions for several treatments while building a model. It takes the difference of the average predicted revenue values in the leaf nodes from customers that received treatment and contrasts them against the related control group customer forecasts, which might be too simplistic.

SMA (RF) achieves remarkable performance in terms of both evaluation metrics. This finding is counterintuitive since the method forecasts treatment-specific outcomes and subtracts outcome differences ex post. The base learners do not optimize outcomes by choosing a treatment from several treatment options per node. SMA's high performance might be explained in that it organizes strong base learners per treatment by relying on established implementations of random forests as part of the "randomForest" R software package, which benefits from extended professional support. This contrasts the implementations of the other models. The original campaign's experimental design might further facilitate correct predictions of the mean differences because the covariate spaces are equal for the different treatment and control groups, as opposed to observational data where they differ per treatment.

Conclusion, Limitations, and Outlook

The paper has developed a novel integrative ITE model for multiple treatment uplift modeling, which estimates continuous outcomes. The proposed algorithm expands the limited research in the field of multiple treatment revenue uplift modeling. Using propensity scores, it adapts to experimental and observational studies. It corrects for possible selection bias without requiring matching the predicted and the observed treatment per unit as opposed to other methods such as PSM. Additional advantages compared to competitive multiple treatment uplift approaches include its capability to generalize toward different evaluation metrics, namely, Revenue Qini curves and modified uplift curves. MTRUF is an ensemble of many multiple treatment revenue uplift trees with specific characteristics. The algorithm follows the objective to determine the most effective treatment from several options by maximizing estimated outcome heterogeneities between pairs of treatments at a node to build individual trees successively. The procedure contrasts the splitting criterion of CTS, which chooses the treatment with the maximal expected response value.

Besides its methodological value, the paper provides an empirical value by testing MTRUF's performance relative to other (types of) multiple treatment uplift models. To this end, the empirical part has included CTS as an integrative ITE model, the causal forest as a pairwise ITE model, and SMA with random forests as a base learner as an instantiation of an outcome model. We have assessed model performance in terms of treatment-specific distributions of forecasted ITE, Revenue Qini curves, and modified uplift curves. To

conduct the analyses, we have used two empirical data sets with multiple e-mail and e-coupon treatments. To increase the robustness of our results, we have drawn ten bootstrap folds with replacement, sizing up the data sets to more than 1.3 million observations in total. As a result of the empirical analysis, we identify MTRUF as the approach that yields the relatively highest individual-level treatment effects, cumulative incremental revenue, and expected response in the form of revenue, which underlines its practical utility. Also, the paper has clarified the relative quantities of the proposed treatments.

The paper exhibits limitations that pave the way for future research. As prevalent in the data, we assume that a unit obtains only one treatment at a time. However, it could be the case that several concurrent campaigns are executed, so that a unit receives more than a single treatment at the same time (e.g., an e-mail and a catalog from the same service provider). Given the lack of related data, we leave an empirical analysis of parallel targeting strategies for future research. While the novel splitting criterion generalizes to a higher number of discrete treatments as considered in this paper, it does not address continuous treatment problems. Identifying related applications and extending MTRUT toward this objective might be an exciting path for future research. Also, the scope of the analysis is limited toward cross-sectional setups of e-mail and e-couponing use cases. External validation considering other settings and data sets is vital to confirm the encouraging results observed here. In particular, assessing the benefits and costs of the proposed approach in online learning applications is a promising avenue for future research. Also, a simulation study would be a very valuable asset to validate the proposed method's advantages further. Lastly, an analysis of the statistical properties of the ITE estimates from MTRUF is an important next step and left for future research.

References

Athey, S., Tibshirani, J., and Wager, S. 2019. "Generalized Random Forests," *The Annals of Statistics* (47:2), pp. 1148-1178.

Breiman, L. 2001. "Random Forests," *Machine Learning* (45:1), pp. 5-32.

Cai, T., Tian, L., Wong, P. H., and J., W. L. 2011. "Analysis of Randomized Comparative Clinical Trial Data for Personalized Treatment Selections," *Biostatistics* (12:2), pp. 270-282.

Chen, X., Owen, Z., Pixton, C., and Simchi-Levi, D. 2015. "A Statistical Learning Approach to Personalization in Revenue Management," *Available at SSRN 2579462*.

Devriendt, F., Moldovan, D., and 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), pp. 13-41.

Gubela, R. M., Lessmann, S., Haupt, J., Baumann, A., Radmer, T., and Gebert, F. 2017. "Revenue Uplift Modeling," in *Proceedings of the 38th International Conference on Information Systems (ICIS'17)*. Seoul, South Korea: AIS.

Gubela, R. M., Lessmann, S., and Jaroszewicz, S. 2020. "Response Transformation and Profit Decomposition for Revenue Uplift Modeling," *European Journal of Operational Research* (283:2), pp. 647-661.

Guelman, L., Guillén, M., and Pérez-Marín, A. M. 2015. "Uplift Random Forests," *Cybernetics and Systems* (46:3-4), pp. 230-248.

Guelman, L., Guillén, M., and Pérez Marín, A. M. 2014. "Optimal Personalized Treatment Rules for Marketing Interventions: A Review of Methods, a New Proposal, and an Insurance Case Study," *UB Riskcenter Working Paper Series, 2014/06*.

Ha, A. 2019. "US Digital Advertising Exceeded $100b in 2018 (IAB Report)." Retrieved 29th April, 2020, from https://techcrunch.com/2019/05/07/iab-internet-advertising-report/

Haupt, J., Jacob, D., Gubela, R. M., and Lessmann, S. 2019. "Affordable Uplift: Supervised Randomization in Controlled Experiments," in *Proceedings of the 40th International Conference on Information Systems (ICIS'19)*. Munich, Germany: AIS.

Hillstrom, K. 2008. "The Minethatdata E-Mail Analytics and Data Mining Challenge." Retrieved 15th June, 2020, from https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

Holland, P. W. 1986. "Statistics and Causal Inference," *Journal of the American Statistical Association* (81:396), pp. 945-960.

Hu, L., Gu, C., Lopez, M., Ji, J., and Wisnivesky, J. 2020. "Estimation of Causal Effects of Multiple Treatments in Observational Studies with a Binary Outcome," *Preprint* (arXiv:2001.06483v1).

Imai, K., and Ratkovic, M. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Applied Statistics* (7:1), pp. 443-470.

Imbens, G. W. 2000. "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika* (87:3), pp. 706-710.

Imbens, G. W., and Wooldridge, J. M. 2009. "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature* (47:1), pp. 5-86.

Kane, K., Lo, V. S., and 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), pp. 218-238.

Knaus, M. C., Lechner, M., and Strittmatter, A. 2020. "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence," *The Econometrics Journal* (doi:10.1093/ectj/utaa014).

Linden, A., Uysal, S. D., Ryan, A., and Adams, J. L. 2016. "Estimating Causal Effects for Multivalued Treatments: A Comparison of Approaches," *Statistics in Medicine* (35:4), pp. 534-552.

Lo, V. S., and Pachamanova, A. D. 2015. "From Predictive Uplift Modeling to Prescriptive Uplift Analytics: A Practical Approach to Treatment Optimization While Accounting for Estimation Risk," *Journal of Marketing Analytics* (3:2), pp. 79-95.

Lopez, M. J., and Gutman, R. 2017. "Estimation of Causal Effects with Multiple Treatments: A Review and New Ideas," *Statistical Science* (32:3), pp. 432-454.

Morgan, S. L., and Winship, C. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, New York: Cambridge University Press.

Olaya, D., Coussement, K., and Verbeke, W. 2020. "A Survey and Benchmarking Study of Multitreatment Uplift Modeling," *Data Mining and Knowledge Discovery* (34), pp. 273-308.

Radcliffe, N. 2007. "Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Models," *Direct Marketing Analytics Journal* (1), pp. 14-21.

Rudaś, K., and Jaroszewicz, S. 2018. "Linear Regression for Uplift Modeling," *Data Mining and Knowledge Discovery* (32:5), pp. 1275-1305.

Rzepakowski, P., and Jaroszewicz, S. 2012. "Decision Trees for Uplift Modeling with Single and Multiple Treatments," *Knowledge and Information Systems* (32:2), pp. 303-327.

Saito, Y., Sakata, H., and Nakata, K. 2020. "Cost-Effective and Stable Policy Optimization Algorithm for Uplift Modeling with Multiple Treatments," *Proceedings of the 2020 SIAM International Conference on Data Mining*: SIAM, pp. 406-414.

Schwab, P., Linhardt, L., and Karlen, W. 2019. "Perfect Match: A Simple Method for Learning Representations for Counterfactual Inference with Neural Networks," *Preprint* (arXiv:1810.00656v5).

Williamson, E. J., Forbes, A., and White, I. R. 2014. "Variance Reduction in Randomised Trials by Inverse Probability Weighting Using the Propensity Score," *Statistics in medicine* (33:5), pp. 721-737.

Zhao, Y., Fang, X., and Simchi-Levi, D. 2017. "Uplift Modeling with Multiple Treatments and General Response Types," in *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, pp. 588-596.

Zhao, Z., and Harinen, T. 2020. "Uplift Modeling for Multiple Treatments with Cost Optimization," *Preprint* (arXiv:1908.05372v3).