Gubela, Robin M.; Lessmann, Stefan; Stöcker, Björn

Multiple Treatment Modeling for Target Marketing Campaigns: A Large-Scale Benchmark Study

Information Systems Frontiers

Provided in Cooperation with:
Springer Nature

Suggested Citation: Gubela, Robin M.; Lessmann, Stefan; Stöcker, Björn (2022) : Multiple Treatment Modeling for Target Marketing Campaigns: A Large-Scale Benchmark Study, Information Systems Frontiers, ISSN 1572-9419, Springer US, New York, NY, Vol. 26, Iss. 3, pp. 875-898, https://doi.org/10.1007/s10796-022-10283-4

Information Systems Frontiers (2024) 26:875–898
https://doi.org/10.1007/s10796-022-10283-4

Multiple Treatment Modeling for Target Marketing Campaigns: A Large-Scale Benchmark Study

Robin M. Gubela[1] · **Stefan Lessmann**[1] · **Björn Stöcker**[2]

Accepted: 24 April 2022 / Published online: 2 June 2022

Abstract
Machine learning and artificial intelligence (ML/AI) promise higher degrees of personalization and enhanced efficiency in marketing communication. The paper focuses on causal ML/AI models for campaign targeting. Such models estimate the change in customer behavior due to a marketing action known as the individual treatment effect (ITE) or uplift. ITE estimates capture the value of a marketing action when applied to a specific customer and facilitate effective and efficient targeting. We consolidate uplift models for multiple treatments and continuous outcomes and perform a benchmarking study to demonstrate their potential to target promotional monetary campaigns. In this use case, the new models facilitate selecting the optimal discount amount to offer to a customer. Large-scale analysis based on eight marketing data sets from leading B2C retailers confirms the significant gains in the campaign return on marketing when using the new models compared to relevant model benchmarks and conventional marketing practices.

Keywords Artificial Intelligence · Business Analytics · Personalization · Uplift Modeling · Machine Learning

1 Introduction

Big data, business analytics, and artificial intelligence (AI) induce business and societal transformations (e.g., Pappas et al., 2018). These technologies have received much interest in information systems (IS) and management research (e.g., Mikalef et al., 2020). Uncovering valuable knowledge from mining large data (e.g., Martens et al., 2016) creates business value in a plethora of use cases (e.g., Chen et al., 2012). IS research uses machine learning and artificial intelligence (ML/AI) for sentiment analysis (Mendon et al., 2021), anti-phishing (Abbasi et al., 2015), donor retention (Kauten et al., 2021), and corporate planning (Smiti & Soui, 2020). ML/AI for personalized marketing enhance a firm's competitiveness and drives organizational decisions, such as tailoring product promotions to individual customers based on deep customer insights (e.g., Khan et al., 2009; Wedel & Kannan, 2016). We focus on the personalization of individual-level push e-promotions such as e-mail marketing or the dynamic inclusion of marketing stimuli into a visited website. In online sales, individual-level personalization is economically more meaningful than tailoring offers to less granular focus groups (Zhang & Wedel, 2009). However, decisions associated with individual customer-level actions (e.g., whether to contact, nudge, incentivize the customer) should be taken based on their marginal costs and benefits (e.g., Gubela et al., 2020).

The identification and explanation of cause-and-effect relationships has received a lot of attention in IS research. Examples include the studies of Ioannou et al. (2022) and Lee and Lee (2012). Taking a formal, statistical perspective, the potential outcome framework (POF) may be seen as the gold standard to estimate the causal effect of a treatment (e.g., sending a newsletter) on an outcome of interest (e.g., an e-shop visit). POF applications in the IS literature include (Tafti & Shmueli, 2020), (Luong et al., 2021), and (Wang et al., 2021). Uplift models capitalize on scalable machine learning techniques and support targeting decisions by establishing this causal effect (e.g., Gubela et al.,

✉ Robin M. Gubela
robin.gubela@hu-berlin.de

Stefan Lessmann
stefan.lessmann@hu-berlin.de

Björn Stöcker
bjoern.stoecker@baur.de

[1] Humboldt-Universität zu Berlin, Spandauer Str. 1, 10178 Berlin, Germany

[2] BAUR Versand, Bahnhofstraße 10, 96224 Burgkunstadt, Germany

2019). They facilitate a nuanced understanding of a customer's likely future behavior and how this behavior changes when assigning treatment to that customer. This provides actionable advice on how to interact with different customers. More formally, considering customer characteristics, an uplift model estimates the *individual treatment effect* (ITE) of a marketing treatment on a customer's behavior (e.g., Devriendt et al., 2018). The ITE is also known as the *conditional average treatment effect* (e.g., Knaus et al., 2021). Estimates of the ITE can be understood as the return on treating a customer. To see this, consider a marketing treatment meant to stimulate purchases. The ITE estimates the change in an individual customer's probability to buy if receiving treatment.

A common targeting strategy is to rank order candidate recipients of a campaign, that is candidate recipients of treatment, based on their model-estimated ITE and to target customers in that order until the available budget is exhausted (e.g., Devriendt et al., 2020). This way, the budget is spent on customers for which a treatment creates the largest value (e.g., increases the probability to buy the most), which ensures efficient utilization of the budget and, more generally, marketing resources. Hence, the ITE is the cornerstone of an efficient targeting policy.

Prior work develops uplift models for the *single treatment case* to address the question "*Which customers shall receive treatment?"* (e.g., Devriendt et al., 2018). A campaign that targets coupons with fixed face value to selected customers exemplifies this setting. Approaches to estimate the ITE can then be distinguished into *conversion uplift modeling* (e.g., Kane et al., 2014) and *revenue uplift modeling* (e.g., Gubela et al., 2020). The former involves predicting binary response variables such as coupon redemption whereas the latter emphasizes heterogeneity in customer values and predicts continuous response variables such as spending amount.

Treatments usually differ in both their configurations and effects. For example, Sawant et al. (2018) investigate different forms of advertising a product while Montaguti et al. (2016) study financial vs. non-financial campaigns. In coupon targeting, an effective strategy of price promotions (Wu et al., 2021), we can also distinguish treatment configurations in the form of different face values. How customers react to different treatment configurations also shows much diversity. A discount is likely to increase the buying probability of a price-sensitive customer substantially whereas a non-monetary treatment might have little effect. A non-monetary treatment in the form of, e.g., product information, a positive review, etc. could nonetheless have a sizeable impact on the buying probability of another customer who used to buy a higher-priced competing product from another vendor. We also observe much variability in treatment effects in our empirical benchmarking experiment. As intuition would suggest, coupons with higher face values tend to be associated with higher ITEs.

To address variability in treatment configurations and effects, the *multiple treatment case* considers several local treatments to answer the question "*Which customers shall receive which treatment?"* (e.g., Rzepakowski & Jaroszewicz, 2012). Allocating the most effective treatment to a customer from among different alternatives facilitates a higher degree of individualization in customer targeting decisions compared to the single treatment case. The business goal to raise the campaign return on marketing is the same as in the single treatment case.

Research on multiple treatment models is scarce and focuses on binary response variables (see Olaya et al., 2020). Recently, two multiple treatment models for continuous outcomes have been proposed in target marketing: the contextual treatment selection (CTS) algorithm (Zhao et al., 2017) and the uplift forest for multiple treatments and continuous outcomes (MTRUF) (Gubela & Lessmann, 2020b). However, the literature lacks a large-scale examination and systematic comparison of continuous outcome models for multiple treatments in marketing. We argue that investigating the utility of corresponding approaches over relevant benchmarks increases the confidence in the new modeling setting and helps marketers raise the effectiveness of their targeting campaigns. Against this background, the paper aims to answer the following research question:

Do multiple treatment models for continuous outcomes realize more campaign return on marketing than multiple treatment models for binary outcomes and single treatment models?

Our study contributes to the growing research of uplift modeling in the following ways. We conceptualize revenue uplift modeling for multiple treatments following Gubela and Lessmann (2020a) and consolidate models in this field. Estimating and evaluating continuous outcomes for multiple treatments facilitates identifying customers who likely generate high revenue, which binary outcome models disregard. In contrast to Gubela and Lessmann (2020a), who provide a first test of one uplift model in the new setting, we perform an extensive benchmark study to validate the state-of-the-art and compare the performance of related methods against multiple treatment conversion uplift models and single treatment uplift models. Other prior work on multiple treatment uplift models often emphasizes non-marketing applications (Rzepakowski & Jaroszewicz, 2012), considers few data sets (e.g., Zhao & Harinen, 2019; Zhao et al., 2017), or disregards continuous outcome models (e.g., Olaya et al., 2020). Our benchmark extends prior studies by considering eight real-world marketing data sets based on product-specific campaigns with electronic discounts from B2C retailers. The data sets have different characteristics in terms of size,

treatment/control group imbalances, and group-specific response rates. Beyond CTS and MTRUF, we also incorporate the popular model-agnostic separate model approach (SMA) (Lo & Pachamanova, 2015) and the recently developed causal forest algorithm (Athey et al., 2019), which prior work in IS has already used for e-commerce targeting (Luo et al., 2019), for the multiple treatment case. We find that especially MTRUF and CTS deliver significantly higher campaign return on marketing than common targeting heuristics, multiple treatment models for binary outcomes, and single treatment models across relevant evaluation metrics, data sets, and customer deciles.

The paper is organized as follows. Section 2 provides the theoretical background. Section 3 reviews the related literature. Section 4 presents the experimental design. Section 5 reports the results. Lastly, Section 6 concludes the paper, reflects on its limitations, and provides an outlook for future research.

2 Theoretical Background

Uplift models support customer targeting by forecasting a unit's behavioral change due to an intervention. In marketing campaigns, uplift models typically estimate the purchase likelihood of customers exposed to treatment (i.e., a *treatment group*) and customers not exposed to treatment (i.e., the *placebo/control group*) conditional on customer characteristics. Let $\boldsymbol{X}_i = \left(\boldsymbol{X}_{i,1}, \ldots, \boldsymbol{X}_{i,p}\right) \in \mathbb{R}^p$ be a p dimensional vector of covariates characterizing customer $i = 1, \ldots N$. Generally, bold and plain typeface refers to vectors and scalars, respectively. We depict the treatment indicator as $T_i \in \{0, \ldots, K\}$ with $K \in \mathbb{Z}^+$; $T_i > 0$ for a unit that obtained a specific treatment, otherwise $T_i = 0$. Moreover, let $Y_{i,c}(T_i) \in \{0, 1\}$ denote a binary response (conversion uplift modeling) and let $Y_{i,r}(T_i) \in \mathbb{R}_0^+$ denote a continuous response (revenue uplift modeling) as a function of treatment. For example, assuming a unit has received treatment $T_i = 1$ from $K = 2$ treatments, we refer to $Y_{i,c}(T_i = 1)$ as the observed response; and to $Y_{i,c}(T_i = 0)$ and $Y_{i,c}(T_i = 2)$ as counterfactual outcomes (e.g., Morgan & Winship, 2015). Response models ignore causality and forecast future realizations of a target variable as a function of $\boldsymbol{X}_i$ for binary and continuous outcomes, respectively, as follows:

$$\widehat{Y}_{i,c}(T_i) = P\left(Y_{i,c}(T_i)|\boldsymbol{X}_i\right) \in [0;1] \tag{1}$$

$$\widehat{Y}_{i,r}(T_i) = E\left(Y_{i,r}(T_i)|\boldsymbol{X}_i\right) \in \mathbb{R}^+ \tag{2}$$

In contrast, uplift models estimate ITE as the conditional mean difference between treatment and control group predictions given $\boldsymbol{X}_i$. The ITE of the K-th treatment based on binary and continuous outcomes is determined as follows:

$$\widehat{\tau}_{i,c}(T_i = K) = P\left(Y_{i,c}\left(T_i = K\right)|\boldsymbol{X}_i\right) - P\left(Y_{i,c}\left(T_i = 0\right)|\boldsymbol{X}_i\right) \in [-1;1] \tag{3}$$

$$\widehat{\tau}_{i,r}(T_i = K) = E\left(Y_{i,r}\left(T_i = K\right)|\boldsymbol{X}_i\right) - E\left(Y_{i,r}\left(T_i = 0\right)|\boldsymbol{X}_i\right) \in \mathbb{R} \tag{4}$$

Suppose that a unit obtains treatment $T_i = 1$, $T_i = 2$, or no treatment (i.e., $T_i = 0$). To support targeting decisions, the treatments' causal impacts are assessed following $\widehat{\tau}_{i,c}(T_i = 1) = P\left(Y_{i,c}\left(T_i = 1\right)|\boldsymbol{X}_i\right) - P\left(Y_{i,c}\left(T_i = 0\right)|\boldsymbol{X}_i\right)$ and $\widehat{\tau}_{i,c}(T_i = 2) = P\left(Y_{i,c}\left(T_i = 2\right)|\boldsymbol{X}_i\right) - P\left(Y_{i,c}\left(T_i = 0\right)|\boldsymbol{X}_i\right)$ regarding conversion uplift modeling; as well as $\widehat{\tau}_{i,r}(T_i = 1) = E\left(Y_{i,r}\left(T_i = 1\right)|\boldsymbol{X}_i\right) - E\left(Y_{i,r}\left(T_i = 0\right)|\boldsymbol{X}_i\right)$ and $\widehat{\tau}_{i,r}(T_i = 2) = E\left(Y_{i,r}\left(T_i = 2\right)|\boldsymbol{X}_i\right) - E\left(Y_{i,r}\left(T_i = 0\right)|\boldsymbol{X}_i\right)$ regarding revenue uplift modeling. A high ITE implies a high impact of the treatment on the unit's incentivized response. Thus, it is common to assign a unit's treatment with the highest ITE. Formally, the treatment based on a binary outcome $\widehat{\tau}_{i,c}^*(T_i)$ or a continuous outcome $\widehat{\tau}_{i,r}^*(T_i)$ can be selected as follows:

$$\widehat{\tau}_{i,c}^*(T_i) = argmax\left(\widehat{\tau}_{i,c}(T_i = 0), \ldots, \widehat{\tau}_{i,c}(T_i = K)\right) \tag{5}$$

$$\widehat{\tau}_{i,r}^*(T_i) = argmax\left(\widehat{\tau}_{i,r}(T_i = 0), \ldots, \widehat{\tau}_{i,r}(T_i = K)\right) \tag{6}$$

The uplift modeling literature (e.g., Kane et al., 2014) categorizes customers into four groups to identify suitable campaign recipients. *Sure Things* convert while *Lost Causes* do not convert regardless of treatment allocation. *Do-Not-Disturbs* do not convert if they receive treatment, while *Persuadables* convert because of it. The treatment selection of Eqs. (5) and (6) is consistent with the uplift modeling objective to identify Persuadables (Kane et al., 2014). In light of the conditional mean difference between treatment and control predictions, as stated in Eqs. (3) and (4), both Sure Things and Lost Causes have low ITE whereas Do-Not-Disturbs and Persuadables have negative and positive ITE, respectively.[1]

We consider three assumptions of the potential outcome framework of causal inference, expanded to the multiple treatment case (Lopez & Gutman, 2017). First, the *conditional independence assumption* refers to the orthogonality between the potential outcomes per treatment and the treatment assignment depending on features, or formally, $\left\{Y_i\left(T_i = 0\right), \ldots, Y_i(T_i = K)\right\} \perp T_i|\boldsymbol{X}_i$. Second, the *overlap*

[1] The categorization is treatment-specific, so that a customer could be a Lost Cause for one treatment and a Persuadable for another, for example. Eqs. (5) and (6) ensure assigning the treatment with the highest ITE to a customer, for which this customer is most likely a Persuadable (if for any treatment).

assumption implies that the probability of receiving a particular treatment—also known as the generalized propensity score $e(T_i, \boldsymbol{X}_i)$—is positive, larger than zero and smaller than 1, $0 < e(T_i, \boldsymbol{X}_i) < 1$, so that there is a probabilistic chance for a unit to obtain any treatment (Imbens, 2000; Rosenbaum & Rubin, 1983). The third assumption, *stable unit treatment value assumption* (SUTVA), further clarifies that a unit's response is independent of the assignment of a treatment to other customers (Rubin, 1980).

Uplift models that satisfy these assumptions facilitate unbiased ITE estimation per treatment, as determined in Eqs. (3) and (4). The use of randomized treatment data complies with the conditional independence and overlap assumptions, which together are denoted as *strong ignorability* (e.g., Imbens & Wooldridge, 2009). For example, randomized treatment allocation ensures independence between the treatment assignment and customer characteristics (Li et al., 2021). In fact, much prior work in marketing employs randomized treatment data as an output of A/B tests (Haupt et al., 2019); also known as multivariate tests in case of multiple treatments. In marketing communication, such tests are performed to gather experimental data for building and evaluating uplift models (Lo, 2002) and/or to assess the effectiveness of alternative marketing incentives to increase business performance (Zhao & Harinen, 2019).

Considering SUTVA requires the independence of a customer's buying decision from the allocation of a treatment to other customers. SUTVA is violated in case of interference or spillover effects (e.g., Ascarza et al., 2017). To illustrate a SUTVA violation, consider two customers A and B. Customer A and B receive a coupon with 20% discount and 10% discount, respectively. Customer B would normally react positively to the discount and buy. However, knowing about customer A receiving a higher discount, B is deterred and chooses not to buy. Another example of SUTVA violations refers to third-party redemptions. Imagine customer A sends his/her 20% discount to customer B who redeems it. Regarding the latter case, our industry partners who contributed the data for our study prevented third-party redemption through their campaign designs. Specifically, third-person redemptions required major technical efforts. Considering the home and fashion campaign, the probability that a third-party redeemed a discount is particularly low due to the short validity period and the encryption of the newsletter redemption codes. Regarding other forms of SUTVA violations, we cannot rule out potential network effects because online shops cannot control customer-to-customer communications. Accordingly, prior work on uplift modeling does not test possible violations of SUTVA (Li et al., 2021) and assumes the suitability of data from randomized controlled trials for ITE estimation. To our best knowledge, this is also standard practice in other fields beyond marketing (e.g., Imai & Ratkovic, 2013). However, an important peculiarity of uplift modeling is that unbiased ITE estimation is not a primary concern because campaign targeting decisions depend on the relative ranking of ITE estimates across customers (Gubela et al., 2020).

To further secure the cause-and-effect association between a marketing treatment and an outcome, we consider Hill's criteria (Hill, 1965). These criteria help us to gain more confidence in the systematic, non-accidental relationship between the two variables and rule out reasoning other than causation. We root this assessment for the multiple treatment models in the application domain of our study and, thus, consider digital price discounts in newsletter and e-coupon (on-site) campaigns. Table 1 shows the results from this analysis.

3 Related Literature

3.1 Modeling Settings

Literature of the single treatment case analyzes which customers to provide a marketing treatment $T_i = 1$ and focuses on binary outcomes. We denote this setting as *single treatment conversion uplift modeling* or *"ST-Conv"*. Established algorithms for this setting include uplift random forests (Guelman et al., 2015), uplift logistic regression (Lo, 2002), and uplift support vector machines (e.g., Zaniewicz & Jaroszewicz, 2013). We refer to benchmark studies for further ST-Conv approaches (e.g., Kane et al., 2014; Knaus et al., 2021). Customers typically buy products with different prices and varying order quantities and thus differ in their spend behavior. Predicting continuous responses aligns modeling outcomes more closely with business objectives than forecasting binary responses (e.g., Gubela et al., 2020). We refer to the corresponding setting as *single treatment revenue uplift modeling* or *"ST-Rev"*. Regarding the modeling methodology, prior work proposes response transformations (Gubela et al., 2020) and uplift linear regressions (Rudaś & Jaroszewicz, 2018).

The multiple treatment case assumes the prevalence of $K > 1$ marketing stimuli (e.g., different coupon discounts) and thus alters the decision problem of customer targeting. A multiple treatment model predicts, for an individual unit, the most effective treatment. Considering different treatment configurations improves campaign decision-making through more customized treatment allocation compared to single treatment targeting. Literature focuses on binary outcomes (e.g., customer conversions). We denote the setting as *multiple treatment conversion uplift modeling* or *"MT-Conv"*. A popular MT-Conv method is the SMA (Lo & Pachamanova, 2015); a generalization of the two model approach (e.g., Cai et al., 2011) for multiple treatments. Other MT-Conv approaches are information-theoretical decision trees

Table 1 Assessment of Multiple Treatment Models Using Hill's Criteria

#	Criterion	Deduced Question	Assessment
1	Strength	To what degree does the treatment affect the outcome?	Strong association • Price is a critical factor (and often a pain point) for customer purchasing decisions • Discounts mitigate this problem and thus affect the purchasing decision
2	Consistency	Do we observe the association repeatedly for different situations?	Consistent association • Targeted men and women bought (all campaigns) • Targeted customers and prospects bought (all campaigns, except for the home and fashion campaign, which targeted solely existing customers) • Campaigns spanned from a few days (the home and fashion campaign) to over a year (e.g., the hats or books campaigns) • We observe the association for different shops, locations, and products
3	Specificity	How specific is the targeted sample with regard to the outcome?	Specific association • Only people were potentially contacted who already had a basic interest in the shop, brand, and products (customers of home and fashion campaign signed up beforehand; people in the other campaigns visited a particular shop website) • Discounts impact the purchasing decision to a larger extent than several alternative factors (Grewal et al., 1998)
4	Temporality	How long does it take to see the effect?	Short-term association • Discounts are redeemable for a short period and thus used instantly • Once a discount is activated, the redemption takes place during the checkout process in the shop
5	Biological gradient	How does the dose–response function look like?	Positive gradient supports the association • Typically, the customer spend increases by the discount value (see Table 2) • Several discounts together may further increase the customer spend. However, only a single voucher is allowed per transaction in the studied applications
6	Plausibility	Is the association reasonable?	Plausible association • Price discrimination theory of coupons (Narasimhan, 1984) • Association credibly shown in prior studies (e.g., Blattberg et al., 1995)
7	Coherence	Does the association conflict with the outcome's natural history?	Coherent association • People are buying goods for personal well-fare • Increasing number of coupons are used[a]; customers appreciate price discounts • Therefore, the association is in line with the general economic trend
8	Experiment	Does a preventive action decrease the event frequency?	Experimental evidence supports the association • Customers often spend less if they do not receive treatment (see Table 2) • However, customers may perceive some specific treatments as worse than no treatment (see Table 2) • Generally, treatment implies a higher value than no treatment (ceteris paribus)
9	Analogy	Is there slighter evidence from other treatments?	Analogous association suspected from other treatments • Price discounts are generally well-perceived • Non-financial incentives may also somewhat positively affect the decision unless they are not seen as annoying or privacy-intrusive (e.g., Martin & Murphy, 2017)

[a]https://balancingeverything.com/coupon-statistics/

(Rzepakowski & Jaroszewicz, 2012), a multiclass transformation (Olaya et al., 2020), and meta-learners for multiple treatments (Zhao & Harinen, 2019). A recent study benchmarks available MT-Conv approaches (Olaya et al., 2020).

Multiple treatment revenue uplift modeling or *"MT-Rev"* represents a modeling setting with a continuous outcome variable in which the treatment variable has more than two levels (including $T_i = 0$). MT-Rev extends the ST-Rev setting and leverages its advantages for the multiple treatment case. MT-Conv models disregard the basket size and product prices, which typically vary across customers. These models identify likely buyers and assume homogenous spendings. In contrast, an MT-Rev model assigns a customer the treatment with the largest impact on customer spending and, in doing so, captures the heterogeneity of customers' basket values. We suggest using an MT-Rev model for campaigns aiming to increase monetary returns, such as up-selling campaigns (e.g., Netessine et al., 2006).

Figure 1, which we reproduce from Gubela and Lessmann (2020a), summarizes the four settings and highlights MT-Rev.

3.2 Modeling Techniques

The implementation of MT-Rev requires specialized modeling techniques. CTS (Zhao et al., 2017) is a forest-based method that serves as the first approach in this field. Its split criterion selects the treatment with the maximal expected response value at a node. CTS considers the sum of a treatment's observed outcome and predicted response in a parent node weighted by a regularization parameter to calculate the estimates of the conditional expected mean outcomes in a child node. Algorithm 1 sheds light on the detailed procedure of CTS (see Appendix 1). CTS empirically outperforms several SMA-based MT-Conv models (Zhao et al., 2017). This result partly contrasts with findings from other studies (Olaya et al., 2020; Zhao & Harinen, 2019).

Another specialized MT-Rev method is MTRUF (Gubela & Lessmann, 2020b). Like CTS, MTRUF is a tree-based learner. It considers the predictions of $\mathcal{T} \in \mathbb{Z}^+$ associated individual trees. Their split criteria select the treatment with the highest response heterogeneity based on random covariates per node. In contrast to CTS, the trees split nodes with positive gains, do not require regularization, and consider the generalized propensity score. Employing this score facilitates unbiased ITE estimates in observational studies. MTRUF further contrasts CTS in that it does not remove high data amounts due to treatment matching. Algorithm 2 details MTRUF using pseudocode (see Appendix 1). Recent work reports on MTRUF's performance gains over CTS, SMA-based random forests, and the causal forest for multiple treatments (Gubela & Lessmann, 2020b); the latter of which we describe as follows.

Next to these algorithms, we consider two approaches that have not originally been developed for the proposed MT-Rev setting but can serve as powerful benchmarks. We refer to them as the causal forest for multiple treatments and the SMA for continuous outcomes. The causal forest is a single treatment learner for continuous outcomes (Athey et al., 2019) and is examined in different applications (e.g., Haupt & Lessmann, 2022). The causal forest considers bootstrapping and recursive partitioning, as in random forests (Breiman, 2001). Its split criterion aims to raise the estimates' heterogeneity. To this end, the causal forest creates intermediate outcomes based on gradients of parent node parameters. Customers who share a leave with a feature's target value receive a higher weight than other customers. The ITE is predicted based on the weighted outcomes. Expanding this method to accommodate multiple treatments is straightforward. The multiple treatment causal forest derives a treatment's individual effect per customer based on $K - 1$ combinations of treatment and control groups. The treatment with the largest effect per unit is allocated, as stated in Eq. (6). The causal forest exhibits statistical consistency and asymptotical normality for fixed covariates and is empirically competitive (e.g., Knaus et al., 2021).

Prior work uses classification methods and examines the SMA as a binary response approach (e.g., Zhao & Harinen, 2019). The SMA represents a model-agnostic framework that allows the flexible use of machine learning models for both classification and regression. The SMA is often implemented using random forests (e.g., Olaya et al., 2020), which outperform other base learners like support vector classifiers and k-nearest neighbors (Zhao et al., 2017) and forecast

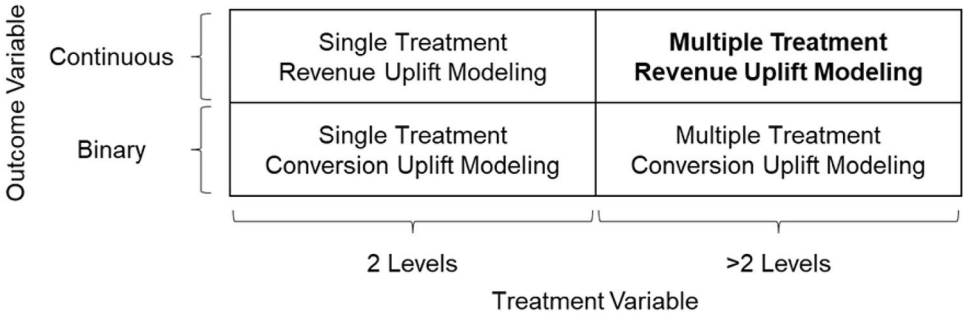

Fig. 1 Modeling Settings

binary and continuous outcomes given observed customer characteristics. The SMA employs $K + 1$ response models on treatment-dependent subsamples (including $T_i = 0$). The expected outcome forecasts from treatment group customers are subtracted by those from control group customers to calculate the individual effect of a treatment. The treatment with the highest ITE relative to other treatments is allocated to a customer. A conceptual drawback of the SMA refers to the poor approximation due to the consideration of separate models on independent data samples.

We implement the four MT-Rev methods using 500 approach-dependent trees and consider a random subsample of $\sqrt{p}$ covariates per candidate split. The causal forest uses the causal tree splitting rule. It performs honest splitting by applying honest risk evaluation for cross-validation with a split alpha parameter of 0.5 to ensure consistent and asymptotically Gaussian estimates. We use a compute server with 256 gigabytes of memory and 40 cores at 3.0 gigahertz to run the models.

4 Experimental Design

We aim to validate the relative utility of the MT-Rev approaches and compare their performances to MT-Conv, ST-Rev, and ST-Conv learners. Reaching this goal requires campaign data, the operationalization of the modeling settings, and evaluation metrics. We shed light on these aspects as follows.

4.1 Campaign Data and Sample Splitting

We examine monetary promotions from international retailers. Prior work identifies higher effectiveness of monetary campaigns as compared to non-monetary ones (Chandon et al., 2000). We received the first data set (*home and fashion*) from a mail-order company that conducted A/B tests for experimental data collection and to identify the most effective treatment. To this end, e-mail newsletters were randomly allocated to newsletter subscribers and existing, active[2] customers with valid e-mail contact. These marketing treatments were sent on the Sunday morning of November 8 in 2020, and responses have been tracked during a consecutive twelve-day period. The newsletters included a link to the retailer's online shop. Apart from customer interactions with the campaign, a customer's surfing behavior during a subsequent shop visit has been monitored. The newsletters of treated customers contained discounts of 10%, 15%, and 20%, which applied to a customer's shopping cart value for home, fashion, and shoe products. Control group customers received a discount-free newsletter. Customers clicking a newsletter's link were forwarded to the online shop. Promotion codes from treated customers who accessed the shop through the link were automatically activated to avoid technical issues with coupon redemption.

We received seven additional data sets from a second industry partner that supports international retailers in onsite targeting initiatives. The companies performed A/B tests by randomly distributing a coupon from several coupon types and a control option (i.e., no coupon) to gather experimental data. A coupon popped up during a customer's online journey after either the third, sixth, or ninth pageview at the seller's online shop. The coupons differed in their face value (e.g., 10, 15, 20) and measurement unit (i.e., euro currency or percentage). They contained promotion codes that required an activation during the checkout and applied to customer basket values. The data sets do not contain person-identifiable information like names or IP addresses. We describe them as follows.

The second data set (*hats*) is based on a hat retailer's campaigns in France, Spain, Italy, the Netherlands, and Germany from December 2018 until January 2020. Treated customers received either a 10€ or 10% coupon by chance. The company is small and offers its products purely through online channels. The third data set (*books1*) refers to a European online bookseller's campaign with 10€, 15€, 10%, 12%, 13%, and 15% coupons that could be redeemed for non-book products due to the fixed book price regulations. The fourth data set (*various products*) refers to coupon campaigns with 5€, 10€, 5%, and 15% discounts from Dutch, French, and German online vendors with specialized, mostly fashion-related product assortments, such as underwear, socks, shirts, blouses, and vines. The retailers are small and medium-sized enterprises with popular brands in their B2C industries and large customer bases.

The fifth and sixth data sets are based on a major international bookstore chain that offers several million products through market-specific offline and online shops (e.g., e-books, movies, music files, and computer games). Specifically, the fifth data set (*books2—AT*) refers to a campaign with 20€ and 15% discounts in Austria from December 2018 until January 2020. The sixth data set (*books2—GER*) is about a campaign with 15€, 12%, 15%, and 17% coupons in Germany during the same period. Lastly, the seventh and eighth data sets refer to a leading shoe seller, which runs thousands of retail stores globally. These data sets differ in time: the campaign from the seventh data set (*shoes1*) lasted from December 2018 until July 2019, while that from the eighth data set (*shoes2*) was carried out from July 2019 until January 2020. Both campaigns took place in Germany and assigned notably high discount values of 20%, 50%, and 75% to random customers. Table 2 provides statistics of the data

[2] Customers who were inactive for over two years were not eligible for being potentially targeted.

Table 2 Overview on Data Sets

	Data Set 1: Home and Fashion	Data Set 2: Hats	Data Set 3: Books1	Data Set 4: Various Products	Data Set 5: Books2 (AT)	Data Set 6: Books2 (GER)	Data Set 7: Shoes1	Data Set 8: Shoes2
Type of campaign	Outbound marketing campaign	Inbound/outbound marketing campaign	Inbound/outbound marketing campaign	Inbound/outbound marketing campaign	Inbound/outbound marketing campaign	Inbound/outbound marketing campaign	Inbound/outbound marketing campaign	Inbound/outbound marketing campaign
Campaign	E-mail newsletter	E-coupon	E-coupon	E-coupon	E-coupon	E-coupon	E-coupon	E-coupon
Market	Germany	France, Spain, Italy, the Netherlands, Germany	Europe	The Netherlands, France, Germany	Austria	Germany	Germany	Germany
Data dimensions	573,958 customers, 81 variables	96,283 customers, 22 variables	71,635 customers, 68 variables	166,536 customers, 40 variables	19,234 customers, 24 variables	195,111 customers, 26 variables	709,347 customers, 23 variables	132,620 customers, 22 variables
Treatments	P10, P15, P20	A10, P10	A10, A15, P10, P12, P13, P15	A5, A10, P5, P15	A20, P15	A15, P12, P15, P17	P20, P50, P75	P20, P50, P75
Number of units	P10: 143,330 (25%) P15: 143,594 (25%) P20: 143,701 (25%) Ctrl: 143,333 (25%)	A10: 940 (1%) P10: 71,219 (74%) Ctrl: 24,124 (25%)	A10: 9,350 (13.1%) A15: 29,348 (41%) P10: 2,232 (3.1%) P12: 333 (0.5%) P13: 3,414 (4.8%) P15: 8,840 (12.3%) Ctrl: 18,118 (25.3%)	A5: 99,314 (58.3%) A10: 12,463 (7.3%) P5: 39,327 (23.1%) P15: 4,928 (2.9%) Ctrl: 10,504 (6.2%)	A20: 6,127 (31.9%) P15: 8,293 (43.1%) Ctrl: 4,814 (25%)	A15: 34,093 (17.5%) P12: 3,516 (1.8%) P15: 81,802 (41.9%) P17: 26,976 (13.8%) Ctrl: 48,724 (25%)	P20: 17,036 (2.4%) P50: 361,072 (50.9%) P75: 153,520 (21.6%) Ctrl: 177,719 (25.1%)	P20: 11,148 (8.4%) P50: 78,090 (55.9%) P75: 10,029 (7.6%) Ctrl: 33,353 (25.1%)
Conversions (ATE in brackets)	P10: 2,030 (485) P15: 2,866 (1,321) P20: 4,015 (2,470) Ctrl: 1,545	A10: 43 (-1,707) P10: 5,439 (3,689) Ctrl: 1,750	A10: 1,124 (-1,311) A15: 5,059 (2,624) P10: 221 (-2,214) P12: 37 (-2,398) P13: 247 (-2,188) P15: 993 (-1,442) Ctrl: 2,435	A5: 16,450 (14,770) A10: 5,373 (3,693) P5: 7,108 (5,428) P15: 1,664 (-16) Ctrl: 1,680	A20: 1,150 (234) P15: 1,744 (828) Ctrl: 916	A15: 6,647 (-3,213) P12: 1,029 (-8,831) P15: 15,071 (5,211) P17: 6,764 (-3,096) Ctrl: 9,860	P20: 1,493 (-10,983) P50: 26,355 (13,879) P75: 9,458 (-3,018) Ctrl: 12,476	P20: 833 (-1,442) P50: 5,297 (3,022) P75: 766 (-1,509) Ctrl: 2,275
Spend p. P. (ATE in brackets)	P10: 2.37€ (0.31€) P15: 3.66€ (1.60€) P20: 5.09€ (3.03€) Ctrl: 2.06€	A10: 7.02€ (1.69€) P10: 5.48€ (0.14€) Ctrl: 5.34€	A10: 0.13€ (-0.06€) A15: 0.25€ (0.06€) P10: 0.11€ (-0.08€) P12: 0.24€ (0.05€) P13: 0.11€ (-0.08€) P15: 0.18€ (-0.01€) Ctrl: 0.19€	A5: 15.80€ (-0.56€) A10: 37.22€ (20.86€) P5: 15.88€ (-0.48€) P15: 25.76€ (9.41€) Ctrl: 16.36€	A20: 5.88€ (0.07€) P15: 6.53€ (0.72€) Ctrl: 5.81€	A15: 5.15€ (-0.96€) P12: 7.18€ (1.07€) P15: 5.51€ (-0.60€) P17: 9.05€ (2.93€) Ctrl: 6.11€	P20: 5.15€ (1.07€) P50: 4.31€ (0.23€) P75: 3.52€ (-0.56€) Ctrl: 4.08€	P20: 4.37€ (0.73€) P50: 3.56€ (-0.08€) P75: 4.19€ (0.55€) Ctrl: 3.64€

P5 = 5%, P10 = 10%, P12 = 12%, P13 = 13%, P15 = 15%, P17 = 17%, P20 = 20%, P50 = 50%, P75 = 75%, A5 = 5€, A10 = 10€, A15 = 15€, A20 = 20€, Ctrl = Control group

sets after pre-processing, which we clarify in Appendix 2. For brevity, "A" and "P" denote absolute and percentage discounts, respectively.

Different data set characteristics may affect the performance of the multiple treatment models. Based on a theoretical analysis with simulated data, Fernández-Loría and Provost (2022) suggest that ITE estimation is challenging for small treatment groups. In practice, collecting experimental data for specific treatments may be costly, leading to treatment/control group imbalances (Haupt et al., 2019). Processing too few observations per treatment may negatively influence the training of the multiple treatment models and their predictions. As Gubela and Lessmann (2021) observe, a low average treatment effect (ATE) may further exacerbate this issue. Typically, the more treatments are available, the smaller are the data samples per treatment. The lower the uplift signal per treatment, the more difficult it is for an uplift model to identify the few high-value customers that buy because of the particular treatment. Thus, low ATE per treatment may imply a limited predictive performance of multiple treatment models.

To prepare the experiments, we draw ten bootstrap folds per data set and randomly split each fold into a 70% training and a 30% test partition. We report averages across the bootstrap folds per data set to increase the results' robustness.

4.2 Operationalization of the Settings

One category of baseline approaches against which we compare the performance of the multiple treatment methods refers to single treatment (conversion/revenue) uplift models. Given campaign data on multiple treatments, how can one compare all four settings' relative utility, including those of the single treatment settings? Several approaches exist to operationalize the settings (Gubela & Lessmann, 2020a). As we show below, the first two of them enable the application of single treatment learners, while the third facilitates using multiple treatment learners.

Specifically, the first approach, *treatment joining approach*, merges the treatments into one set and compares its gains against the control group. A unit's treatment value is 1 if this unit received any treatment, otherwise 0. An uplift model predicts a binary response from the adapted treatment set and control group samples (e.g., Zaniewicz & Jaroszewicz, 2017). Extending this approach toward ST-Rev only requires a continuous response. We call the second approach *treatment dropping approach*. Its idea is to choose an arbitrary treatment and dismiss the other treatments. A single treatment model forecasts a binary outcome based on the selected treatment and control group (e.g., Kane et al., 2014). An extension toward ST-Rev models is straightforward.

Both approaches reduce the multiple treatment decision problem to that of single treatments and facilitate using corresponding models. The downside of these approaches is that they disregard differences in the effectiveness of individual treatments. The treatment dropping approach further implies a loss of data in the magnitude of $K-1$ treatments. Thus, we prioritize the treatment joining approach over the treatment dropping approach to implement the single treatment settings. We stress that we do not criticize prior work using these approaches to empirically validate new single treatment learners in light of the scarcity of open-access campaign data.

Finally, MT-Conv research uses the third approach, which we refer to as *multiple treatment approach*, to compare each treatment's effectiveness against the control group without treatment modification. A treatment's effect on a unit is measured and the most impactful treatment is allocated.

Figure 2 illustrates the three approaches for conversion and revenue uplift modeling. Suppose $K=3$ treatments. In terms of the treatment dropping approach, we randomly remove treatments $T=1$ and $T=2$ for illustration.

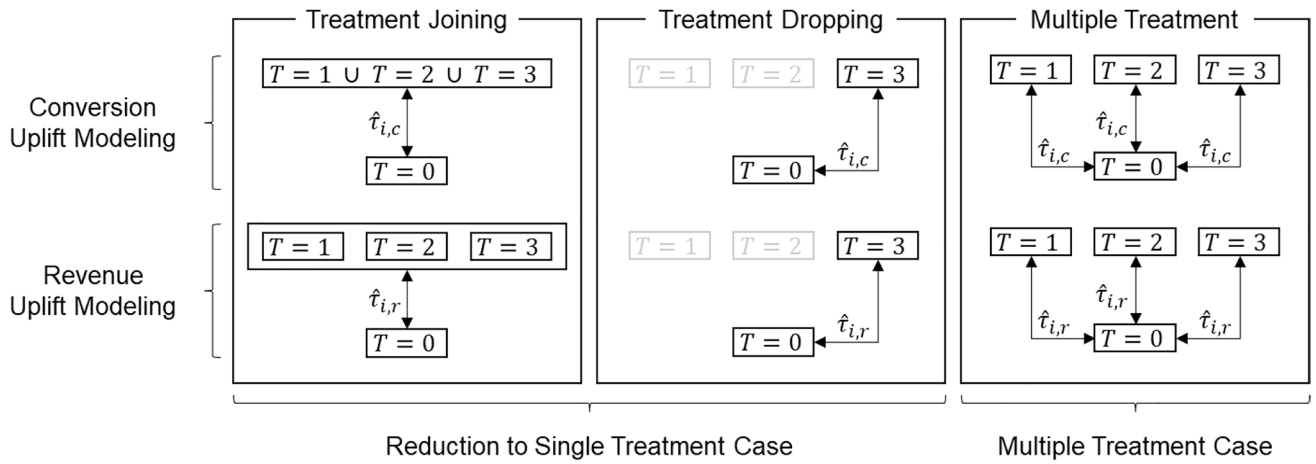

Fig. 2 Approaches to Operationalize the Modeling Settings

4.3 Evaluation Metrics

We measure the campaign return on marketing using Qini curves (e.g., Zhao & Harinen, 2019) and the recently proposed expected response metric (Zhao et al., 2017). These metrics depict the state-of-the-art to evaluate the performance of multiple treatment learners (Olaya et al., 2020). Uplift models predict the ITE per treatment and test set unit. The treatment with the highest forecasted ITE is assigned to a customer. Regarding Qini curves, we create matrices for the treated and control customers from the test set, which contains the ITE estimate, spend, and profit. Next, we sort the customers in decreasing order of the ITE estimates. This step ensures that more receptive customers (those with higher ITE estimates) get a higher rank than less receptive customers (those with lower ITE estimates), following the efficient resource allocation paradigm of uplift modeling. For the first customer decile (i.e., 10% of the population), we then take the sum of the revenue (profit) of treated customers minus the weighted sum of the revenue (profit) of control customers. The control-dependent weighting divides the number of treated customers by the number of control customers (Radcliffe, 2007). We re-iterate the procedure consecutively for the remaining deciles. Qini curves reduce the multiple treatment problem to a binary distinction between the treated and not treated clients. This is common practice in evaluating multiple treatment uplift models (see Olaya et al., 2020) and facilitates intuitive comparisons of uplift model performance. However, given that the two-dimensional curves do not elucidate the different treatments per decile, their interpretability in terms of treatment choice is limited. Assigning treatment to no customer (decile 0) or all customers (decile 10) does not require a targeting policy to discriminate between more and less relevant customers.

In contrast to Qini curves, the expected response metric performs treatment matching per customer. To this end, it checks if a customer's model-predicted treatment with the highest ITE matches the observed treatment from the historical data to gain confidence in the predictions. Only if this condition holds, the corresponding customer's expected outcome will be assessed, which typically causes a significant loss of data. The observed outcome from matched customers is divided by their treatment probability and sample size to calculate an unbiased expected response per person. We refer to Zhao et al. (2017) for proofs.

5 Results

We report the results as follows. We first analyze the effectiveness of the MT-Rev approaches. Then, we investigate the best MT-Rev method against relevant MT-Conv models and single treatment learners.

5.1 Analysis of MT-Rev Approaches

Figure 3 illustrates the MT-Rev learners' campaign return on marketing in terms of the expected response per customer and the incremental revenue/profit from targeting (i.e., the Qini metric) across the ten deciles for the data sets 1–4. These data sets refer to the home and fashion campaign, the books1 campaign, the hats campaign, and the various products campaign. Regarding the home and fashion data set, we assess a customer's profit tracked during a twelve-day period after the campaign. Figure 7 in Appendix 3 provides further results for a six-week period. The curves and shaded areas represent the models' mean results and standard errors across bootstrap folds. The grey line refers to random targeting.

Figure 3 reveals the following insights. First, the MT-Rev approaches generate higher campaign return on marketing than targeting none, all, or randomly. We remind readers that targeting all customers yields the same incremental revenue across models, whereas the expected spend response differs in decile 10 due to the treatment matching procedure. Most of the model curves steeply increase on the first deciles, which indicates that the models target a few high-value customers. This finding applies to both metrics and stresses the relevance of these models for campaigns with limited budgets (e.g., Ascarza, 2018).

Second, we observe approach-related performance differences. MTRUF and CTS are the most effective methods regarding both metrics. They generally outperform the SMA and causal forest for continuous outcomes. CTS is superior on the home and fashion data set. MTRUF is the predominant MT-Rev learner in terms of the books1 data set. CTS increases the incremental revenue from targeting more than the other techniques regarding the hats and various products data sets. We explain the strong performance of these models by referring to their methodological designs. Both MTRUF and CTS are specialized MT-Rev learners that construct trees by distinguishing between the expected value per treatment.

A third observation refers to the variability of model results on the hats data set. The MT-Rev methods' standard errors across bootstrap folds are exceptionally high regarding the expected response metric. This uncertainty may be a result of the data set characteristics. The hats campaign allocated the 10€ coupon to only 1% of customers, whereas the 10% coupon was assigned to 74% of customers. Besides its small size, which may explain the forecasts' uncertainty (Olaya et al., 2020), the 10€ coupon also implied notably low conversion rates in both the training and test partitions per bootstrap fold (each with 28,885 customers). For example, only eighteen and

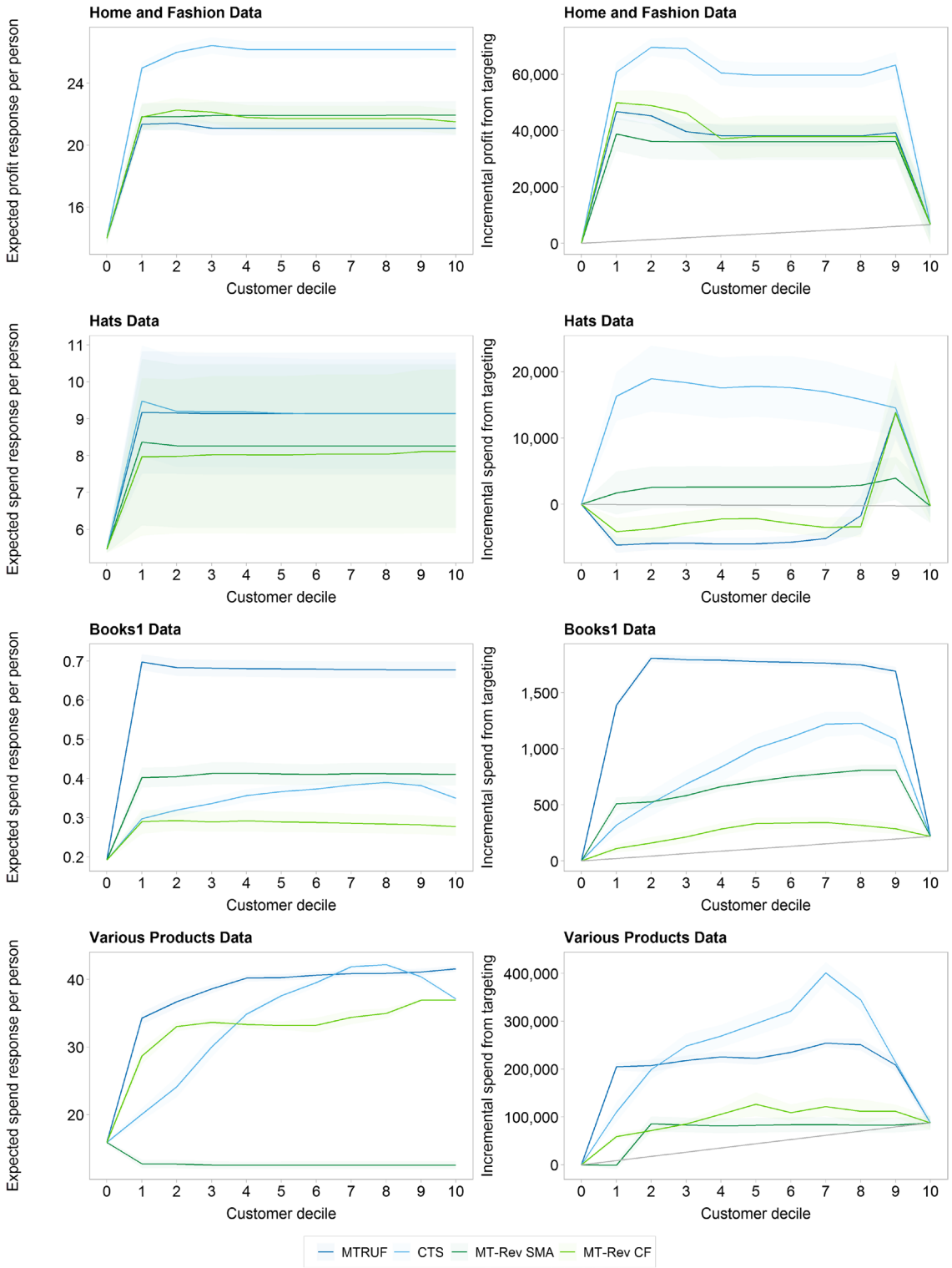

Fig. 3 MT-Rev Analysis for Data Sets 1–4

nine customers bought with a 10€ discount regarding the fourth bootstrap fold of the training and test partition, respectively. These low figures indicate the challenge of the multiple treatment models to capture likely

(high-value) buyers due to the 10€ discount and propose the treatment that matches the observed treatment, as required by the expected response metric.[3]

Fourth, several models partly perform inferior to no or random targeting. Specifically, MTRUF and the causal forest yield lower performance than random targeting regarding the first eight deciles of the hats data set. We interpret this result by referring to our previous discussion of the hats campaign. Excitingly, both models enormously increase the incremental revenue on the ninth decile. They outperform random targeting by several magnitudes and achieve a similar level of the campaign return on marketing as CTS. To this end, MTRUF and the causal forest assign treatment to expected high-value customers after targeting 80% of less valuable customers. This action is suboptimal from a business perspective. Moreover, the SMA model yields a lower expected spend per person on the various products data set than no targeting. The 10€ and 15% coupons have much higher spend ATE than the 5€ and 5% coupons. With its treatment-dependent random forests, SMA may find it particularly challenging to match these treatments effectively.

We repeat the steps of the previous analysis for the data sets 5–8. These data sets refer to two international companies and have structural differences. The bookstore chain conducted campaigns in Austria (books2—AT) and Germany (books2—GER). The shoe retailer conducted campaigns during different periods. Specifically, the first campaign took place from December 2018 until July 2019 (shoes1), and the second campaign was carried out from July 2019 until January 2020 (shoes2). Figure 4 shows the corresponding results for the employed causal models (colored curves) and the random targeting policy (grey line).

We make the following observations based on Figure 4. First, targeting customers based on the propositions of the MT-Rev approaches allows marketers to realize remarkable campaign return on marketing. This finding applies across data sets, evaluation metrics, and customer deciles and supports our claim of the MT-Rev setting's advantage for target marketing. Akin to prior results for the data sets 1–4, we observe steep increases of many model curves on the first few deciles for the data sets 5–8. To this end, the models effectively identify customer subgroups that likely generate substantial revenue because of a specific treatment.

Second, the specialized MT-Rev approaches outperform SMA and the causal forest across data sets and on all but one decile. Regarding the books data sets, MTRUF raises more expected spend response per person but lower incremental revenue than CTS. This observation holds for both the Austrian and the German market. CTS is generally more effective than the other models on the shoes data sets. An exception refers to the shoes2 data set, for which MTRUF yields the highest expected spend response per person.

Third, the causal forest performs worse than random targeting for the books2 campaign in Germany and the shoes2 campaign. Its model curves fall below the lines showing random targeting for deciles 1–5 (shoes2) and deciles 1–8 (books2—GER). Here, the performance of the MT-Rev causal forest much differs from the other MT-Rev approaches.

5.2 Analysis of MT-Rev vs. Other Settings

In the following, we investigate the extent to which the MT-Rev models perform differently than the MT-Conv, ST-Rev, and ST-Conv models. We consider the best MT-Rev model per data set and evaluation metric from the previous analysis and echo prior experiments regarding the eight data sets and two metrics. We employ SMA-based random forests (abbreviated to "SMA") and the causal forest ("CF") for each additional modeling setting and data set. We reiterate that a valuable property of the separate model approach is its flexibility to accommodate classification and regression models as conversion uplift models ("Conv") and revenue uplift models ("Rev"), respectively. Much prior work uses random forests (e.g., Olaya et al., 2020), which adapt to both response scales. Considering the single treatment settings ("ST"), the SMA can be seen as the two model approach (e.g., Cai et al., 2011). The causal forest algorithm has been recently developed as a single treatment learner (Athey et al., 2019). As a benchmark approach, it serves for both single treatment settings and the multiple treatment conversion setting ("MT-Conv").

Figure 5 displays the models' performances in terms of the expected response metric and the Qini metric for the data sets 1–4. The curves and their shades depict the predicted averages and standard errors, respectively; the grey line represents the results from random targeting.

The following findings emerge from Figure 5. First, the MT-Rev model achieves significantly more expected responses per person and significantly higher incremental value from targeting than each of the benchmark approaches (including random targeting). Table 3 provides the corrected p-values based on several different post hoc tests; all of which are based on García et al. (2010) (see Appendix 4). The MT-Rev model is either CTS (home and fashion, hats), MTRUF (books1), or each of these approaches for one metric (various products). The gains of the MT-Rev models are especially pronounced in terms of the expected response metric. To this end, the MT-Rev model consistently outperforms the second-best model per data set and customer decile. A prominent example refers

[3] Few buyers existed among the customers with a match of the estimated and the observed treatment. For example, the treatments matched for only 4,206 buyers (i.e., 1.5% of the population) and 7,115 buyers (i.e., 2.5% of the population) regarding SMA and CTS, respectively. This observation may also explain the variability of the results as only few persons had positive expected responses for model assessment.

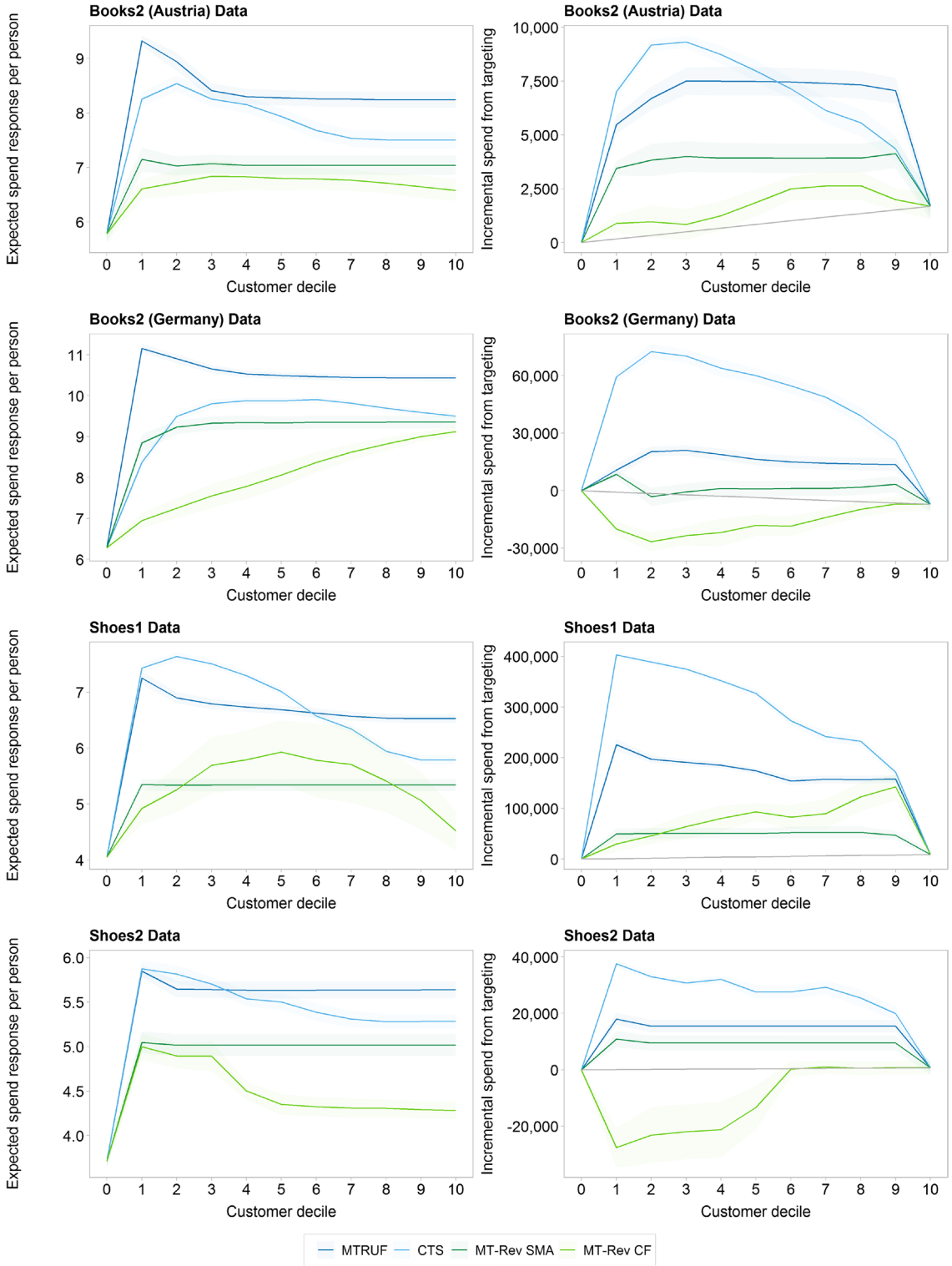

Fig. 4 MT-Rev Analysis for Data Sets 5–8

to the hats campaign. Despite its somewhat volatile estimates, CTS yields an expected spend response of over 9€ per customer on average. In contrast, the expected spend response per person of the second-best model, the MT-Conv SMA, is below 6€. Another example refers to the books1 campaign. MTRUF provides an expected (scaled) spend response per person of about 0.70€ across deciles. In contrast, the runner-up model, again the MT-Conv SMA, generates only at most about 0.55€

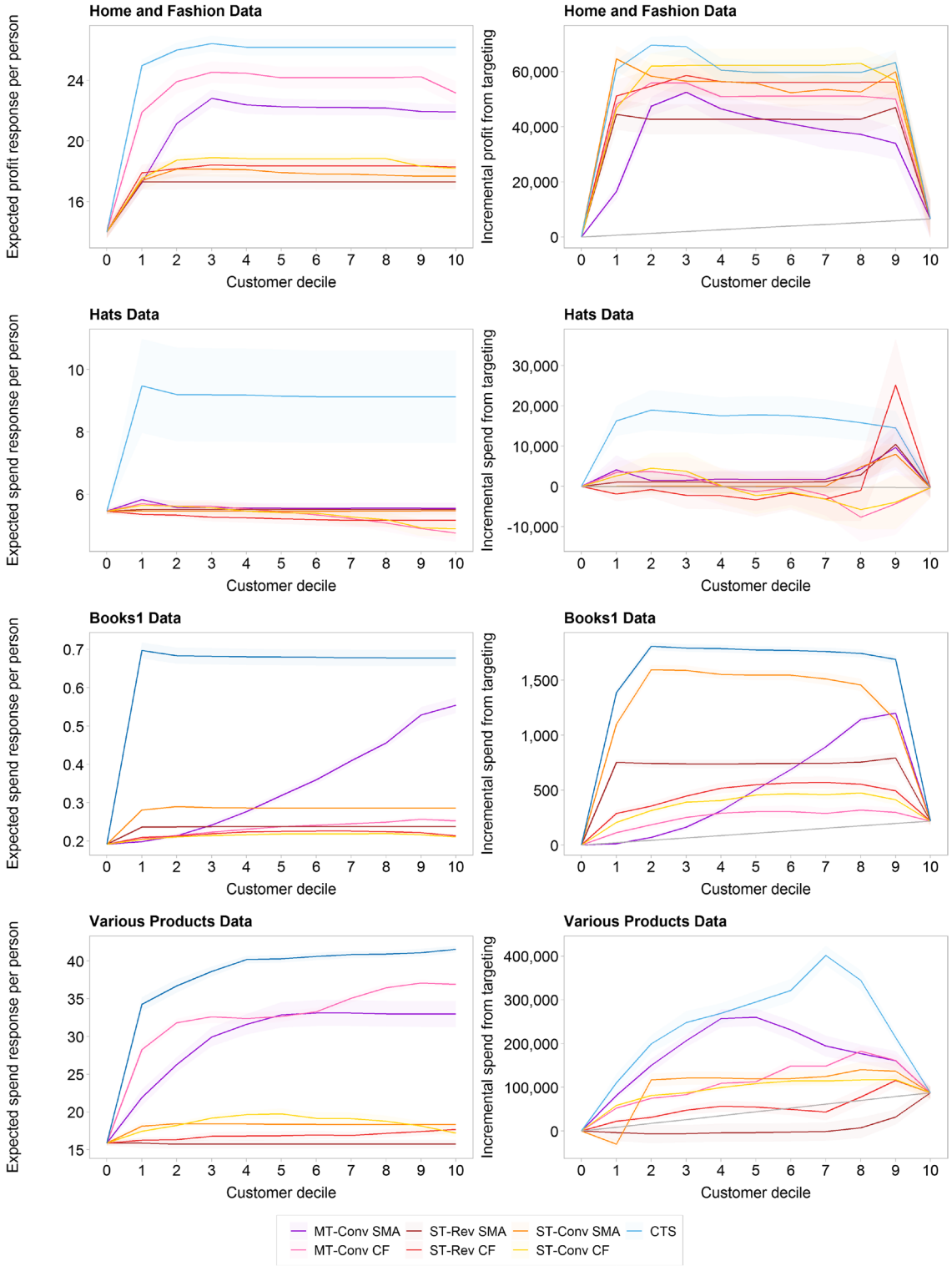

Fig. 5 Analysis of MT-Rev vs. Other Settings for Data Sets 1–4

when targeting all customers and gradually less campaign return on marketing when targeting fewer customers. These results underline the viability of the MT-Rev approaches for campaign practices.

Second, the performance of the benchmark models remarkably varies per data set and evaluation metric. For example, these models deliver a high expected profit response per customer on the home and fashion data set. The single treatment approaches substantially raise the incremental profit from targeting over random targeting and partly even outperform CTS for deciles 4–9. The extensive, meaningful data set and the positive conversion ATE may explain these results. On the other hand, the benchmark models do not provide enhancements of the campaign return on marketing on the hats data set. This observation applies to both metrics. The performance of most benchmark models stagnates across deciles. An exception is that most models' Qini curves peak at the ninth decile, which further underlines their little practical advantage for the hats campaign. Some of the models even perform inferior to no or random targeting, which also applies to the various products data set in a few instances. In this case, the comparably few covariates and the small size of the control group may impede the predictions of several single treatment models.

Third, the MT-Conv models often considerably outperform the single treatment learners regarding the expected response metric. Regarding the home and fashion data set, for example, the MT-Conv causal forest efficiently assigns high-profit and low-profit customers to the first and last deciles, respectively. This model yields an expected profit of about 24€ per customer, which contrasts the best single treatment learner with a difference of over 5€ per customer. Considering the Qini metric, however, the advantages of MT-Conv models over the most competitive ST-Rev and ST-Conv models are limited. An example is the ST-Conv SMA on the books1 data set, which increases the incremental revenue to a higher degree than the MT-Conv models.

We continue analyzing the best MT-Rev model against the benchmark approaches for the data sets 5–8. Figure 6 visualizes the results. Recall that the campaign return on marketing from random targeting is illustrated again by the grey line.

The subsequent insights emerge. First, the MT-Rev models significantly raise both the expected spend response per person and the incremental revenue compared to the individual benchmark models and the random targeting baseline.[4] Table 4 details the statistical results based on post hoc tests (see Appendix 4). Neither a single MT-Conv model nor a single treatment model is as competitive as the MT-Rev model on any data set or customer decile. For example, CTS achieves more than twice as much incremental revenue from targeting than the best benchmark model per data set. Regarding the shoes2 campaign, it even realizes on average about three times higher incremental revenue than the second-best model. The MT-Rev models also contribute much higher campaign return on marketing than the benchmark approaches when targeting the 10% of customers with the highest predicted ITE. Corresponding improvements on the expected response metric range from about 20% to over 50% across the data sets. These findings align with previous results and reinforce the usefulness of the MT-Rev approaches for target marketing initiatives.

Second, the benchmark models contribute different levels of the campaign return on marketing per data set. Considering the books campaign in Austria, these models increase the expected spend response per customer and incremental revenue over no or random targeting. Regarding the books campaign in Germany, several models yield inferior results to no or random targeting, especially the causal forest models for single treatments. In terms of both shoes data sets, most benchmark models raise the expected spend response per person. However, the causal forest models for single treatments (shoes1) and, in addition to them, the causal forest for multiple treatments and binary outcomes (shoes2) incur negative incremental revenue from targeting. We remind that the causal forest has been initially developed for the single treatment case, which may explain its limited gains here.

Third, we identify remarkable performance differences between the benchmark models. The MT-Conv models are more effective than the ST-Rev and ST-Conv models regarding the expected response metric. In particular, the MT-Conv SMA performs several magnitudes better than the MT-Conv causal forest and the single treatment models on the books data sets. In terms of the shoes campaigns, the MT-Conv causal forest outperforms the other approaches across most of the deciles. Its closest competitor is the MT-Conv SMA (shoes2) and a single treatment learner (shoes1). A final observation refers to the Qini metric. The multiple treatment models yield higher incremental revenue than several single treatment learners. However, the ST-Rev SMA shows superior performance than both MT-Conv models across the data sets. The gradients of its curves are exceptionally high on the first decile. This result supports the benefit of revenue uplift modeling for campaigns with budget constraints.

[4] We note a single exception: The performance of the best MT-Rev model is statistically insignificant to that of the ST-Rev SMA model for the Qini metric and the data sets 5–8. We refer the reader to Appendix 4 for related details.

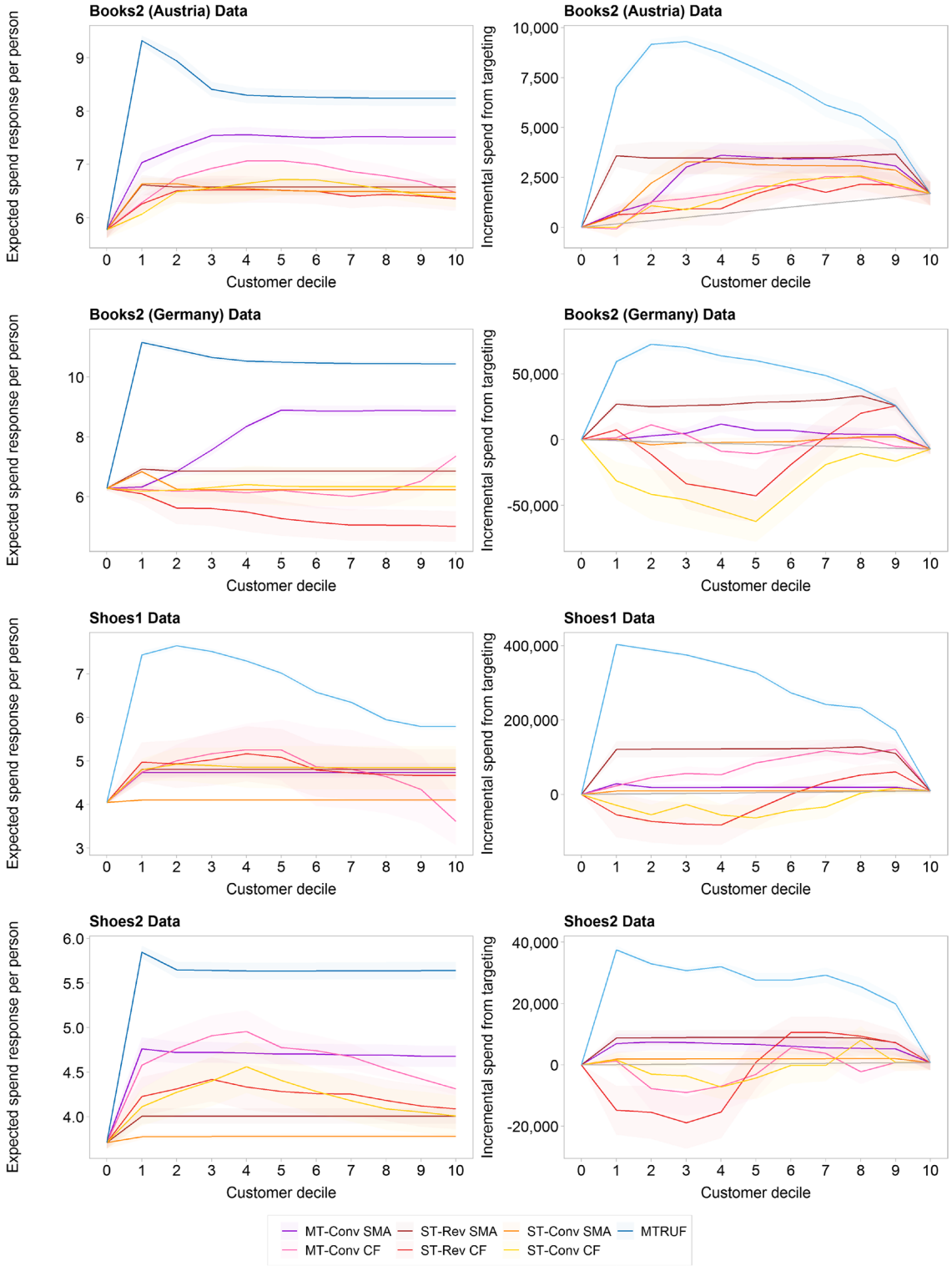

Fig. 6 Analysis of MT-Rev vs. Other Settings for Data Sets 5–8

6 Conclusion, Limitations, and Future Research

The paper introduced multiple treatment revenue uplift modeling to improve the decision-making in target marketing campaigns. In contrast to causal models focusing on binary responses, MT-Rev approaches estimate continuous responses for multiple treatments, which, as we argue, better aligns modeling outcomes with relevant campaign objectives. To contribute to the literature, we conceptualized the proposed MT-Rev setting as in (Gubela & Lessmann, 2020a) and consolidated models in this field, which we assessed by carrying out a large-scale benchmark study. The analysis considered eight monetary marketing campaigns from leading B2C retailers, covering a broad range of e-commerce markets. The campaigns were conducted in different European countries during varying periods within the last two years. The data sets varied in their sizes and attributes as well as in their number, distributions, and the ATE of the marketing treatments. We examined the contributions of the MT-Rev approaches in terms of the campaign return on marketing and their relative utility compared to several baseline practices and MT-Conv and single treatment model benchmarks. The models predicted different responses (i.e., customer revenue, profits, purchases). We used the business-related Qini and the expected response metrics to validate model performance.

We first verified the usefulness to adopt the MT-Rev approaches for customer targeting. Our results demonstrate that these approaches yield higher campaign return on marketing than baseline heuristics, which target randomly, no customer, or all customers for both metrics and across data sets. Few exceptions refer to specific combinations of a model, data set, and decile. Most MT-Rev models even captured the few high-value customers in the first customer deciles, which underlines their suitability for campaign targeting. Among these approaches, we found the specialized MT-Rev methods CTS and MTRUF particularly powerful across data sets and for most of the deciles. Except for the expected response metric and the hats data set, the models exhibited low standard errors across bootstrap samples, which increased our confidence in the predicted results.

Next, we compared the performance of the best MT-Rev learner per data set and evaluation metric against challenging MT-Conv, ST-Rev, and ST-Conv benchmarks. This analysis allowed us to conclude that the MT-Rev model significantly outperformed the approaches from the alternative settings in terms of the campaign return on marketing. The best MT-Rev model outperformed the best benchmark model on all but one data set and across deciles. It yielded substantially higher expected responses per person and incremental value from targeting. While the campaign return on marketing of the benchmark models varied per data set and metric, the MT-Conv models delivered higher expected responses than the single treatment approaches. These advantages were less pronounced in terms of the Qini metric.

We acknowledge the following limitations that provide opportunities for future research. While the data sets have different treatment and response distributions, analyzing additional data characteristics may deliver further insights into the MT-Rev approaches' relative benefits. For example, future work using simulated data may examine different strengths of treatment effects and noise levels. Such analysis facilitates assessing a model's performance in terms of the mean squared error between the true (i.e., specified) and the predicted treatment effect. The recently developed precision in the estimation of heterogeneous effects (PEHE) measure for multiple treatments (Schwab et al., 2019) represents a suitable metric for this purpose. Furthermore, artificial data may also help to examine the performance of the multiple treatment models for customers with similar (uniform) bias and the location of the predictions in relation to the decision boundary, as suggested by Fernández-Loría and Provost (2022). In addition, most of our data sets are based on e-coupon campaigns from the retailing industry. The similarity of these data sets may influence the generalization of our results. Studying other marketing campaigns from other industries thus remains a valuable path for future research. Moreover, we investigate the utility of a monetary treatment's different configurations. Assessing the effectiveness of alternative monetary campaigns (e.g., an e-coupon vs. a physical coupon) or a mix of monetary and non-monetary campaigns (e.g., an e-coupon vs. an advertising call) might be a promising research direction toward omnichannel marketing. Offline and online applications could serve as another attribute of distinction. Beyond that, extending the meta-learners proposed by Zhao and Harinen (2019) to handle regression problems and examining the relative performance of the recently proposed variance reduced treatment selection (VARTS) algorithm (Saito et al., 2020) may be exciting avenues for future research. Experiments with VARTS may be particularly treasured given that it produces unbiased estimates under lower variance compared to CTS, leading to performance gains over CTS, as recently observed for small data samples with high noise levels (Saito et al., 2020). A future analysis may also include powerful uplift algorithms and transformations as additional single treatment benchmarks. Extending single treatment transformations to accommodate multiple treatments may be yet another fruitful approach for future research. A final idea refers to the optimization of the timing and frequency to assign an individual customer a personalized marketing communication from among several alternatives.

Appendix 1. Original MT-Rev Algorithms

Algorithm 1 Contextual Treatment Selection (CTS)

Input: B bootstrap samples, $\mathbb{T}$ trees, $\mathcal{X}$ random covariates to support node splitting, $\mathcal{S}_{min}$ as the minimal number of treatment-dependent units per leaf, n_reg for regularization

Output: Treatment with the highest predicted expected response

Divide the sample into training and test data
for $\mathbb{t} = 1$ to $\mathbb{T}$ **do**
  for $b = 1$ to B **do**
    Draw B bootstrap samples with replacement from the training data (proportional per treatment)
    Select $\mathcal{X}$ random covariates from the covariate space $\boldsymbol{X}$
    Calculate the gain $G(\mathcal{S})$ to split a node based on CTS' splitting criterion (including n_reg)
    repeat
      Split a parent node with the highest (non-negative) gain into two child nodes
    until $G(\mathcal{S}) < 0$ **or** a node contains less than $\mathcal{S}_{min}$ **or** units in a node have equal response values
  end for
end for
Get CTS' response prediction as the average from $\mathbb{T}$ predictions
Allocate the treatment to a unit, which has the highest estimated expected response

Algorithm 2 Multiple Treatment Revenue Uplift Forest (MTRUF)

Input: B bootstrap samples, $\mathcal{T}$ multiple treatment revenue uplift trees, $\mathcal{X}$ random covariates to support node splitting, $\mathcal{S}_{min}$ as the minimal number of treatment-dependent units per leaf, $\mathcal{D}_{max}$ as a tree's maximal depth, $e(t, \boldsymbol{X}_i)$ as the (generalized) propensity score

Output: Predicted outcome heterogeneities between different treatments

Randomly draw B bootstrap samples with replacement
for $b = 1$ to B **do**
  Divide the sample into training and test data
  for $t = 1$ to $\mathcal{T}$ (tree development) **do**
    Randomly sample training data with replacement
    Select $\mathcal{X}$ random covariates from the covariate space $\boldsymbol{X}$
    Calculate the gain $G(\mathcal{S})$ to split a node based on the splitting criterion
    repeat
      Split a parent node with the highest (positive) gain into two child nodes
    until $G(\mathcal{S}) \leq 0$ **or** a node contains less than $\mathcal{S}_{min}$ **or** a tree is grown to $\mathcal{D}_{max}$
  end for
  Get MTRUF's response prediction as the average from $\mathcal{T}$ predictions weighted by $e(t, \boldsymbol{X}_i)$
end for
Average predictions across B bootstrap samples and calculate standard errors
Allocate the treatment to a unit, which has the highest outcome heterogeneity

Appendix 2. Data Pre-Processing

We pre-process the data sets as follows. For the home and fashion data set, we first exclude long-term inactive customers. We further drop 1,706 units for which the e-mail delivery failed due to technical issues. We create a dummy if a unit received a newsletter on November 8 or 9. The average duration of related units opening and clicking the newsletter after the reception is 318.4 and 375.4 minutes. On average, the newsletter has been opened and clicked 0.37 and 0.06 times, respectively. These numbers seem plausible, according to our industry collaborator. Moreover, we drop 11 customers with negative spend. We calculate the profit by considering the spend subtracted by the product of a discount's redemption indicator and the discount value relative to the basket value. The data has neither constant nor missing values. However, since it has been extracted from different databases, some covariates have "-1" labels, which we replace by zeros (e.g., a unit has no prior visit, purchase, etc.). We further create an indicator that documents if we imputed a value. We transform several hardware-related factor variables into dummies, such as a customer's device and operating system. Furthermore, we analyze variable inflation factors (VIF) for dimensionality reduction. After two iterations, we identify many variables with collinearity issues and keep 62 variables with VIF scores of less than 10. Lastly, we drop 54 customers with high negative values in several, primarily duration-related variables. The final sample has 573,958 units and 81 variables, measuring a customer's conversion, spend, and profit in a two-day, twelve-day, and six-week post-campaign period. It also includes a customer's activity in the last session regarding the monetary value, basket interactions, logons, page impressions, and many duration-related variables (e.g., time spent on product pages, detail views, account pages, and different stages during the checkout process). Moreover, customer behavior across sessions within the four pre-campaign weeks are available in the form of the session frequency per device, the number of logons and page impressions, and the accumulated time spent on specific shop websites.

Regarding the remaining data sets, we first remove customers who received a non-monetary incentive and clean several campaign value labels to clarify the campaign's treatments after clarification with our industry collaborator. We drop single variables with a significant share of missing values or solely constant values. Next, we correct data types where appropriate (e.g., we convert character variables to integers). We extract time data from the epoch timestamp of a customer's last session and last conversion. We create dummy variables based on these variables and based on a customer's current channel and initial screen type. Dummies with over 97% zeros are dropped to facilitate model predictions. We further transform customer spend from cents into euros by dividing corresponding values by 100 and round to two-digit numbers. The pre-processed data sets with similar numbers of variables have 96,283 records and 22 variables (hats), 19,234 records and 24 variables (books2—AT), 195,111 records and 26 variables (books2—GER), 709,347 records and 23 variables (shoes1), 132,620 records and 22 variables (shoes2). These variables capture different aspects of a customer's activity during the actual online session at a corresponding shop. Meta-variables identify which treatment a customer received (if any), customer purchases, and spend. Further exemplary variables refer to the use of a shopping cart, the channel to access the shop (e.g., direct access per URL, through an e-mail link, or affiliate marketing), interactions with the coupon pop-up window (e.g., confirming the call for action, closing the pop-up window, dropping-off), if the client used a mobile device and if this person is known.

The pre-processed books1 and various products data sets measure additional customer characteristics. These refer to the current session (e.g., which page type the customer visited before, the session's daytime/nighttime and whether the shop visit is close to Christmas) and past sessions (e.g., purchases in the preceding week/month/year and session counts in the prior week). Moreover, further device-related data (e.g., the operating system) and location-based data are included. To this end, the books1 data set contains 71,635 records and 68 variables, while the various products data set has 166,536 and 40 variables. The customer spend on the books1 data set has been scaled by our partner for confidentiality reasons by dividing its value per customer by the variable's average.

Appendix 3. Additional Empirical Results

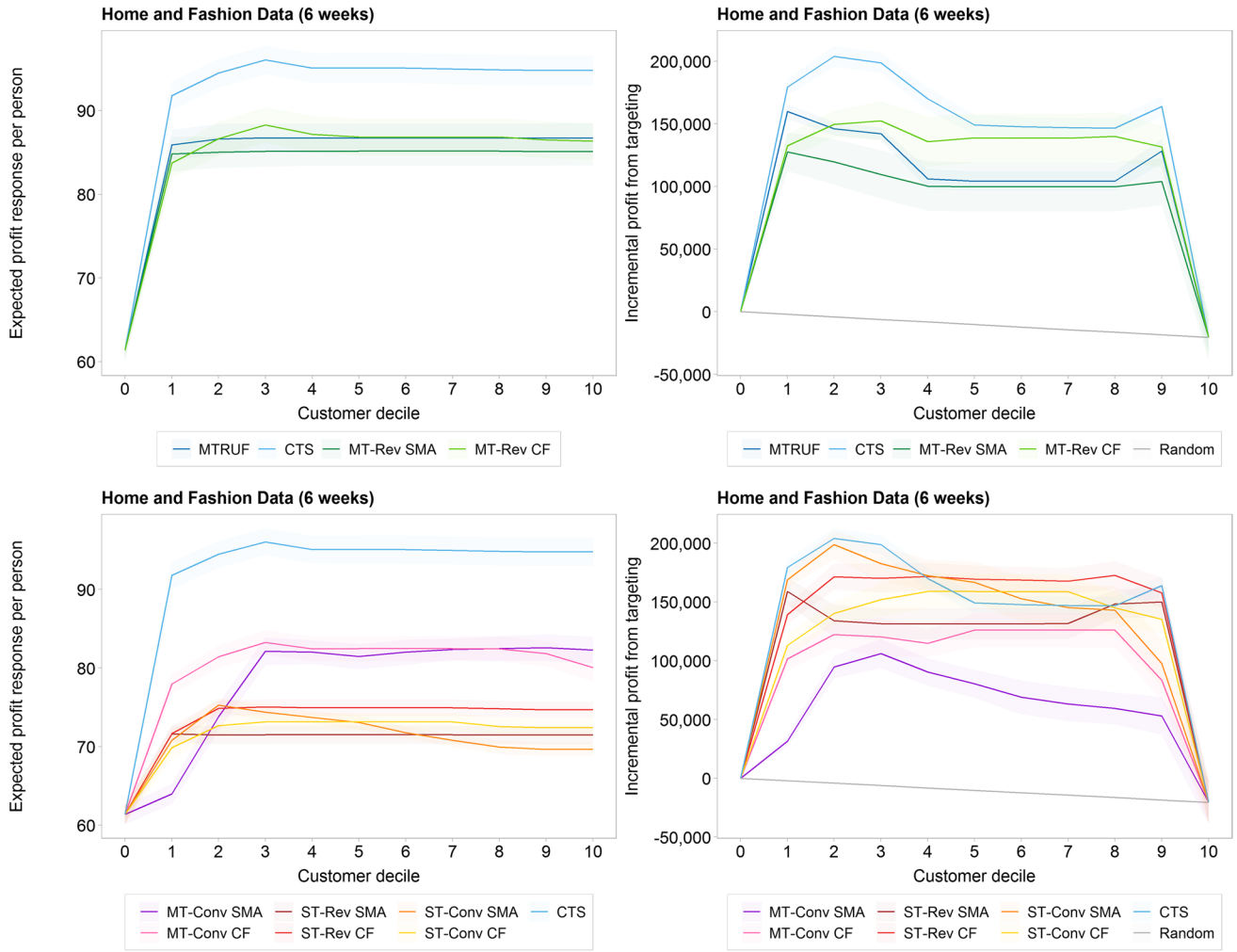

Fig. 7 Model Analysis for a Six-Week Post-Campaign Period

Appendix 4. Multiple Hypothesis Testing

Table 3 Post Hoc Tests of MTRUF/CTS vs. Each Benchmark for Data Sets 1–4

Metric	Model	Corrected p-Values			
		Holland	Finner	Rom	Li
Expected response	MTRUF/CTS	*NA*	*NA*	*NA*	*NA*
	MT-Conv SMA	0.0074	0.0054	0.0074	0.0054
	MT-Conv CF	0.0074	0.0045	0.0074	0.0037
	ST-Rev SMA	0	0	0	0
	ST-Rev CF	0	0	0	0
	ST-Conv SMA	0	0	0	0
	ST-Conv CF	0	0	0	0
Qini	MTRUF/CTS	*NA*	*NA*	*NA*	*NA*
	MT-Conv SMA	0	0	0	0
	MT-Conv CF	0	0	0	0
	ST-Rev SMA	0	0	0	0
	ST-Rev CF	0	0	0	0
	ST-Conv SMA	0.0004	0.0004	0.0004	0.0004
	ST-Conv CF	0	0	0	0
	Random Targeting	0	0	0	0

We first conducted a Friedman Aligned Ranks test as a non-parametric omnibus test per metric to check for any statistically significant difference in the campaign return on marketing between the models. We confirm that this is true based on the test results, which are $T = 154.14$, $df = 6$, $p < 2.2 \times 10^{-16}$ (expected response) and $T = 101.51$, $df = 7$, $p < 2.2 \times 10^{-16}$ (Qini metric). Also note that corrected p-values below 1×10^{-4} are shown as 0.

Table 4 Post Hoc Tests of MTRUF/CTS vs. Each Benchmark for Data Sets 5–8

Metric	Model	Corrected p-Values			
		Holland	Finner	Rom	Li
Expected response	MTRUF/CTS	*NA*	*NA*	*NA*	*NA*
	MT-Conv SMA	0.0023	0.0023	0.0023	0.0023
	MT-Conv CF	0	0	0	0
	ST-Rev SMA	0	0	0	0
	ST-Rev CF	0	0	0	0
	ST-Conv SMA	0	0	0	0
	ST-Conv CF	0	0	0	0
Qini	MTRUF/CTS	*NA*	*NA*	*NA*	*NA*
	MT-Conv SMA	0	0	0	0
	MT-Conv CF	0	0	0	0
	ST-Rev SMA	0.3696	0.3696	0.3696	0.3696
	ST-Rev CF	0	0	0	0
	ST-Conv SMA	0	0	0	0
	ST-Conv CF	0	0	0	0
	Random Targeting	0	0	0	0

We first conducted a Friedman Aligned Ranks test as a non-parametric omnibus test per metric to check for any statistically significant difference in the campaign return on marketing between the models. We confirm that this is true based on the test results, which are $T = 159.45$, $df = 6$, $p < 2.2 \times 10^{-16}$ (expected response) and $T = 160.51$, $df = 7$, $p < 2.2 \times 10^{-16}$ (Qini metric). Also note that corrected p-values below 1×10^{-4} are shown as 0

Declarations

Conflict of Interest None.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abbasi, A., Zahedi, F. M., Zeng, D., Chen, Y., Chen, H., & Nunamaker, J. F., Jr. (2015). Enhancing predictive analytics for anti-phishing by exploiting website genre information. *Journal of Management Information Systems, 31*(4), 109–157. https://doi.org/10.1080/07421222.2014.1001260

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

Ascarza, E., Ebbes, P., Netzer, O., & Danielson, M. (2017). Beyond the target customer: Social effects of customer relationship management campaigns. *Journal of Marketing Research, 54*(3), 347–363. https://doi.org/10.1509/jmr.15.0442

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

Blattberg, R. C., Briesch, R., & Fox, E. J. (1995). How promotions work. *Marketing Science, 14*(3), G122–G132. https://doi.org/10.1287/mksc.14.3.G122

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

Cai, T., Tian, L., Wong, P. H., & Wei, J. L. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics, 12*(2), 270–282. https://doi.org/10.1093/biostatistics/kxq060

Chandon, P., Wansink, B., & Laurent, G. (2000). A benefit congruency framework of sales promotion effectiveness. *Journal of Marketing, 64*(4), 65–81. https://doi.org/10.1509/jmkg.64.4.65.18071

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188. https://doi.org/10.2307/41703503

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

Devriendt, F., Van Belle, J., Guns, T., & Verbeke, W. (2020). Learning to rank for uplift modeling. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/TKDE.2020.3048510

Fernández-Loría, C., & Provost, F. (2022). Causal classification: Treatment effect estimation vs. outcome prediction. *Journal of Machine Learning Research, 23*(59), 1–35. https://www.jmlr.org/papers/v23/19-480.html

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences, 180*(10), 2044–2064. https://doi.org/10.1016/j.ins.2009.12.010

Grewal, D., Krishnan, R., Baker, J., & Borin, N. (1998). The effect of store name, brand name and price discounts on consumers' evaluations and purchase intentions. *Journal of Retailing, 74*(3), 331–352. https://doi.org/10.1016/S0022-4359(99)80099-2

Gubela, R. M., Bequé, A., Gebert, F., & Lessmann, S. (2019). Conversion uplift in e-commerce: A systematic benchmark of modeling strategies. *International Journal of Information Technology & Decision Making, 18*(3), 747–791. https://doi.org/10.1142/S0219622019500172

Gubela, R. M., & Lessmann, S. (2020a). Interpretable multiple treatment revenue uplift modeling. *Proceedings of the 26th Americas Conference on Information Systems (AMCIS'20)*, AIS, 18. https://aisel.aisnet.org/amcis2020/data_science_analytics_for_decision_support/data_science_analytics_for_decision_support/18/. Accessed 10 March 2022.

Gubela, R. M., & Lessmann, S. (2020b). Uplift forest for multiple treatments and continuous outcomes. *Proceedings of the 41st International Conference on Information Systems (ICIS'2020b)*, AIS, 17. https://aisel.aisnet.org/icis2020/digital_commerce/digital_commerce/17/. Accessed 10 March 2022.

Gubela, R. M., & Lessmann, S. (2021). Uplift modeling with value-driven evaluation metrics. *Decision Support Systems, 150*, 113648. https://doi.org/10.1016/j.dss.2021.113648

Gubela, R. M., Lessmann, S., & Jaroszewicz, S. (2020). Response transformation and profit decomposition for revenue uplift modeling. *European Journal of Operational Research, 283*(2), 647–661. https://doi.org/10.1016/j.ejor.2019.11.030

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2015). Uplift random forests. *Cybernetics and Systems, 46*(3–4), 230–248. https://doi.org/10.1080/01969722.2015.1012892

Haupt, J., Jacob, D., Gubela, R. M., & Lessmann, S. (2019). Affordable uplift: Supervised randomization in controlled experiments. *Proceedings of the 40th International Conference on Information Systems (ICIS'19)*, AIS, 24. https://aisel.aisnet.org/icis2019/data_science/data_science/24. Accessed 10 March 2022.

Haupt, J., & Lessmann, S. (2022). Targeting customers under response-dependent costs. *European Journal of Operational Research, 297*(1), 369–379. https://doi.org/10.1016/j.ejor.2021.05.045

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58*(5), 295–300. https://doi.org/10.1177/003591576505800503

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics, 7*(1), 443–470. https://doi.org/10.1214/12-AOAS593

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika, 87*(3), 706–710. https://doi.org/10.1093/biomet/87.3.706

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5–86. https://doi.org/10.1257/jel.47.1.5

Ioannou, A., Lycett, M., & Marshan, A. (2022). The role of mindfulness in mitigating the negative consequences of technostress. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10239-0

Kane, K., Lo, V. S., & 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

Kauten, C., Gupta, A., Qin, X., & Richey, G. (2021). Predicting blood donors using machine learning techniques. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10149-1

Khan, R., Lewis, M., & Singh, V. (2009). Dynamic customer management and the value of one-to-one marketing. *Marketing Science, 28*(6), 1063–1079. https://doi.org/10.1287/mksc.1090.0497

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal, 24*(1), 134–161. https://doi.org/10.1093/ectj/utaa014

Lee, M., & Lee, J. (2012). The impact of information security failure on customer behaviors: A study on a large-scale hacking incident on the internet. *Information Systems Frontiers, 14*, 375–393. https://doi.org/10.1007/s10796-010-9253-1

Li, J., Zhang, W., Liu, L., Yu, K., Le, T. D., & Liu, J. (2021). A general framework for causal classification. *International Journal of Data Science and Analytics, 11*, 127–139. https://doi.org/10.1007/s41060-021-00249-1

Lo, V. S. (2002). The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter, 4*(2), 78–86. https://doi.org/10.1145/772862.772872

Lo, V. S., & 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), 79–95. https://doi.org/10.1057/jma.2015.5

Lopez, M. J., & Gutman, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science, 32*(3), 432–454. https://doi.org/10.1214/17-STS612

Luo, X., Lu, X., & Li, J. (2019). When and how to leverage e-commerce cart targeting: The relative and moderated effects of scarcity and price incentives with a two-stage field experiment and causal forest optimization. *Information Systems Research, 30*(4), 1203–1227. https://doi.org/10.1287/isre.2019.0859

Luong, T. T., Sivarajah, U., & Weerakkody, V. (2021). Do agile managed information systems projects fail due to a lack of emotional intelligence? *Information Systems Frontiers, 23*, 415–433. https://doi.org/10.1007/s10796-019-09962-6

Martens, D., Provost, F., Clark, J., & Fortuny, EJd. (2016). Mining massive fine-grained behavior data to improve predictive analytics. *MIS Quarterly, 40*(4), 869–888. https://doi.org/10.25300/MISQ/2016/40.4.04

Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. *Journal of the Academy of Marketing Science, 45*(2), 135–155. https://doi.org/10.1007/s11747-016-0495-4

Mendon, S., Dutta, P., Behl, A., & Lessmann, S. (2021). A hybrid approach of machine learning and lexicons to sentiment analysis: Enhanced insights from twitter data of natural disasters. *Information Systems Frontiers, 23*, 1145–1168. https://doi.org/10.1007/s10796-021-10107-x

Mikalef, P., Pappas, I. O., Krogstie, J., & Pavlou, P. A. (2020). Big data and business analytics: A research agenda for realizing business value. *Information & Management, 57*(1), 103237. https://doi.org/10.1016/j.jfca.2019.103237

Montaguti, E., Neslin, S. A., & Valentini, S. (2016). Can marketing campaigns induce multichannel buying and more profitable customers? A Field Experiment. *Marketing Science, 35*(2), 201–217. https://doi.org/10.1287/mksc.2015.0923

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press.

Narasimhan, C. (1984). A price discrimination theory of coupons. *Marketing Science, 3*(2), 128–147. https://doi.org/10.1287/mksc.3.2.128

Netessine, S., Savin, S., & Xiao, W. (2006). Revenue management through dynamic cross selling in e-commerce retailing. *Operations Research, 54*(5), 893–913. https://doi.org/10.1287/opre.1060.0296

Olaya, D., Coussement, K., & Verbeke, W. (2020). A survey and benchmarking study of multitreatment uplift modeling. *Data Mining and Knowledge Discovery, 34*, 273–308. https://doi.org/10.1007/s10618-019-00670-y

Pappas, I. O., Mikalef, P., Giannakos, M. N., Krogstie, J., & Lekakos, G. (2018). Big data and business analytics ecosystems: Paving the way towards digital transformation and sustainable societies. *Information Systems and E-Business Management, 16*, 479–491. https://doi.org/10.1007/s10257-018-0377-z

Radcliffe, N. (2007). Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal, 1*, 14–21. https://www.research.ed.ac.uk/en/publications/using-control-groups-to-target-on-predicted-lift-building-and-ass. Accessed 10 March 2022.

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

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association, 75*(371), 591–593. https://doi.org/10.2307/2287653

Rudaś, K., & Jaroszewicz, S. (2018). Linear regression for uplift modeling. *Data Mining and Knowledge Discovery, 32*(5), 1275–1305. https://doi.org/10.1007/s10618-018-0576-8

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

Saito, Y., Sakata, H., & 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 (SDM)*, SIAM, 406–414. https://doi.org/10.1137/1.9781611976236.46.

Sawant, N., Namballa, C. B., Sadagopan, N., & Nassif, H. (2018). Contextual multi-armed bandits for causal marketing. *Preprint*. https://doi.org/10.48550/arXiv.1810.01859.

Schwab, P., Linhardt, L., & Karlen, W. (2019). Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *Preprint*. https://doi.org/10.48550/arXiv.1810.00656.

Smiti, S., & Soui, M. (2020). Bankruptcy prediction using deep learning approach based on borderline SMOTE. *Information Systems Frontiers, 22*, 1067–1083. https://doi.org/10.1007/s10796-020-10031-6

Tafti, A., & Shmueli, G. (2020). Beyond overall treatment effects: Leveraging covariates in randomized experiments guided by causal structure. *Information Systems Research, 31*(4), 1183–1199. https://doi.org/10.1287/isre.2020.0938

Wang, X., Sun, J., Wang, Y., & Liu, Y. (2021). Deepen electronic health record diffusion beyond breadth: Game changers and decision drivers. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-020-10093-6

Wedel, M., & Kannan, P. (2016). Marketing analytics for data-rich environments. *Journal of Marketing, 80*(6), 97–121. https://doi.org/10.1509/jm.15.0413

Wu, J., Zhao, H., & Chen, H. (2021). Coupons or free shipping? Effects of price promotion strategies on online review ratings. *Information Systems Research, 32*(2), 633–652. https://doi.org/10.1287/isre.2020.0987

Zaniewicz, Ł., & Jaroszewicz, S. (2013). Support vector machines for uplift modeling. *Proceedings of the 13th IEEE International*

Conference on Data Mining Workshops, IEEE, 131–138. https://doi.org/10.1109/ICDMW.2013.23.

Zaniewicz, Ł, & Jaroszewicz, S. (2017). Lp-support vector machines for uplift modeling. *Knowledge and Information Systems, 53*(1), 269–296. https://doi.org/10.1007/s10115-017-1040-6

Zhang, J., & Wedel, M. (2009). The effectiveness of customized promotions in online and offline stores. *Journal of Marketing Research, 46*(2), 190–206. https://doi.org/10.1509/jmkr.46.2.190

Zhao, Y., Fang, X., & Simchi-Levi, D. (2017). Uplift modeling with multiple treatments and general response types. *Proceedings of the 2017 SIAM International Conference on Data Mining*, SIAM, 588–596. https://doi.org/10.1137/1.9781611974973.66.

Zhao, Z., & Harinen, T. (2019). Uplift modeling for multiple treatments with cost optimization. *Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA'19)*, IEEE, 422–431. https://doi.org/10.1109/DSAA.2019.00057.

Robin M. Gubela holds a PhD in Information Systems from Humboldt-University of Berlin. In his research, he develops personalization strategies, machine learning algorithms and business-centric evaluation metrics for improved one-to-one marketing. Robin authored papers that have been published in high-ranked journals, such as the European Journal of Operational Research and Decision Support Systems (Google Scholar, ResearchGate). He also presented papers at top academic conferences, such as the International Conference on Information Systems. Robin serves as a reviewer for recognized journals and conferences. He taught in the fields of machine learning, e-business, and digital marketing and led data science, digitalization, and transformation projects at large companies from various industries.

Stefan Lessmann completed his PhD and habilitation at the University of Hamburg in 2007and 2012, respectively. He then joined the Humboldt-University of Berlin in 2014, where he heads the Chair of Information Systems. He serves as an associate editor for several international journals and department editor of Business and Information System Engineering (BISE). Stefan has secured substantial amounts of research funding and published several papers in leading international journals and conferences (Google Scholar, ResearchGate). His research concerns machine learning and artificial intelligence (MLAI) methodologies and their use cases in managerial decision support. Stefan specializes on MLAI applications in the broad scope of marketing and risk analytics. Stefan actively participates in knowledge transfer and consulting projects with industry partners; from start-up companies to global players and not-for-profit organizations.

Björn Stöcker is Head of CRM at BAUR and responsible for direct marketing. BAUR is one of the ten largest online shops in Germany with a focus on fashion, shoes, and furniture. Since joining BAUR in 2008, Björn has focused on personalization, customer relationship management, and sales strategy. With his team, he plans, analyzes anddevelops BAUR's direct marketing campaigns. Björn obtained a PhD from the University of Bayreuth. In his dissertation, Björn researched new approaches in CRM online fashion retail. Resulting papers have been published, for example, in the Journal of Business Economics. A main focus was the optimal customer selection using uplift modeling. Besides his job, Björn engages in teaching different classes about applied e-commerce, sales controlling, and business intelligence.