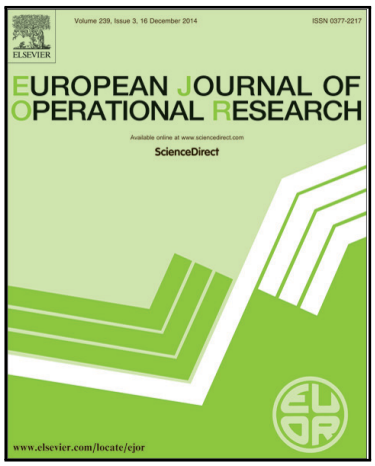

Response Transformation and Profit Decomposition for Revenue Uplift Modeling

Robin M. Gubela , Stefan Lessmann , Szymon Jaroszewicz

PII: S0377-2217(19)30941-5
DOI: https://doi.org/10.1016/j.ejor.2019.11.030
Reference: EOR 16166

To appear in: *European Journal of Operational Research*

Received date: 11 July 2018
Accepted date: 15 November 2019

Please cite this article as: Robin M. Gubela , Stefan Lessmann , Szymon Jaroszewicz , Response Transformation and Profit Decomposition for Revenue Uplift Modeling, *European Journal of Operational Research* (2019), doi: https://doi.org/10.1016/j.ejor.2019.11.030

Highlights

- We introduce uplift modeling for revenue maximization
- A response transformation enables uplift modeling using standard classifiers
- We propose a decomposition of campaign profit for uplift modeling
- We provide a broad empirical evaluation of revenue uplift models in e-commerce
- Our new uplift strategy outperforms several causal machine learning algorithms

Response Transformation and Profit Decomposition for Revenue Uplift Modeling

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

Stefan Lessmann (corresponding author)
Chair of Information Systems
Humboldt-Universität zu Berlin
Spandauer Str. 1, 10178 Berlin
stefan.lessmann@hu-berlin.de

Szymon Jaroszewicz
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
s.jaroszewicz@ipipan.waw.pl

Response Transformation and Profit Decomposition for Revenue Uplift Modeling

Abstract

Uplift models support decision-making in marketing campaign planning. Estimating the causal effect of a marketing treatment, an uplift model facilitates targeting marketing actions to responsive customers and efficient allocation of marketing budget. Research into uplift models focuses on conversion models to maximize incremental sales. The paper introduces uplift models for maximizing incremental revenues. If customers differ in their spending behavior, revenue maximization is a more plausible business objective compared to maximizing conversions. The proposed methodology entails a transformation of the prediction target, customer-level revenues, that facilitates implementing a causal uplift model using standard machine learning algorithms. The distribution of campaign revenues is typically zero-inflated because of many non-buyers. Remedies to this modeling challenge are incorporated in the proposed revenue uplift strategies in the form of two-stage models. Empirical experiments using real-world e-commerce data confirm the merits of the proposed revenue uplift strategy over relevant alternatives, including uplift models for conversion and recently developed causal machine learning algorithms. To quantify the degree to which improved targeting decisions raise return on marketing, the paper develops a decomposition of campaign profit. Applying the decomposition to a digital coupon targeting campaign, the paper provides evidence that revenue uplift modeling, as well as causal machine learning, can improve campaign profit substantially.

Keywords—OR in Marketing, Profit Analytics, Uplift Model, Causal Machine Learning

1 Introduction

Predictive models support decision-making in marketing. Use cases include forecasting usage frequencies in social media (e.g., Ballings & Van den Poel, 2015), response toward advertisement (e.g., Goldfarb & Tucker, 2011), and customer churn (e.g., Verbeke et al., 2012). We distinguish two categories of marketing models, response and uplift models. Response models predict customer behavior in general. The term was coined in the direct marketing literature, where the modeling goal is often to predict how a customer will react to a marketing stimulus (e.g., Baesens et al., 2002). Uplift models also consider a direct marketing setting but estimate the differential change in response behavior due to the marketing action (Devriendt et al., 2018). This way, an uplift model accounts for the causal link between the action and customer response. Causality is crucial to measure the true impact of a marketing campaign, maximize campaign profit, and allocate scarce marketing resources efficiently (e.g., Lo & Pachamanova, 2015). The paper proposes novel methodologies for uplift modeling to support targeting decisions in marketing campaign planning.

We consider promotional campaigns that aim at maximizing sales revenues in e-commerce through issuing digital coupons (e.g., Reimers & Xie, 2019). Coupon targeting is a relevant decision task. For example, US retailers distributed 256.5 billion coupons for consumer packaged goods in 2018, and consumers redeemed over 1.7 billion coupons with a combined face value of $2.7 billion (NCH Marketing Services, 2018). Issuing a coupon equates to a price reduction. In this regard, coupon allocation re-emphasizes the cruciality of using a causal targeting model. Offering a coupon to a customer who would buy at the ordinary price decreases sales margin. Therefore, a targeting model must identify those customers who have no intention to buy but can be persuaded to buy through a discount. Uplift models perform this identification by estimating the differential impact of the coupon on a customer's buying propensity, while a response model predicts the *net* buying propensity of customers, which is less useful for targeting (e.g., Ascarza, 2018).

Prior research on uplift models focuses on applications with a binary prediction target, such as a purchase incident. An uplift model could then estimate the change in customers' purchase probabilities due to a marketing action (e.g., a coupon). We call such models conversion models since they capture whether the action has altered customer behavior, where behavior refers to a binary event such as redeeming a coupon, clicking a link, or buying a product (e.g., Devriendt et al., 2018). Targeting campaigns using a conversion uplift model implies a sales maximization objective in that conversion uplift, by definition, captures incremental customer actions. Due to heterogeneity in customer spending (e.g., Schröder & Hruschka, 2017), maximizing incremental conversions and maximizing incremental profits do not coincide. Considering our application setting, the ideal customer to offer a coupon would be a person who has no buying intention initially, is successfully persuaded to buy by the coupon, and then spends a large amount. Conversely, converting a customer who spends a small amount is less valuable and might have a

negative profit impact if the treatment costs associated with coupon provision exceed incremental revenues. The overarching goal of the paper is to introduce revenue uplift modeling. Similar to a conversion model, which strives to maximize incremental sales, a revenue uplift model aims at maximizing incremental revenues. In applications where customers exhibit heterogeneity in spending or value, revenue uplift modeling represents a more direct approach to maximize campaign profit. It facilitates targeting a campaign to those customers who generate the most substantial incremental revenue. The paper develops approaches to develop revenue uplift models using supervised machine learning (SML) algorithms and tests their effectiveness through empirical experimentation in the scope of e-coupon targeting.

A revenue uplift model estimates the incremental sales revenue on an individual customer level. The continuous prediction target prohibits the use of existing uplift modeling approaches for conversion, where the target variable is binary. A first contribution of the paper is the extension of the binary target variable transformation (Lai et al., 2006), a competitive method to develop conversion uplift models (Devriendt et al., 2018; Kane et al., 2014), to accommodate a continuous target variable. We propose two transformations that facilitate developing a revenue uplift model using SML algorithms for regression or classification. This way, marketers have full flexibility to choose their preferred learning algorithm and use their preferred SML software package to estimate a causal revenue uplift model.

In addition to a change in the scaling level of the prediction target, a second methodological challenge in revenue uplift modeling concerns the revenue distribution, which is typically skewed. Considering our application setting, most visitors to an online shop do not buy and generate zero revenues. Likewise, coupon redemption rates in grocery retailing are only 0.7% in the US (NCH Marketing Services, 2018). More generally, the issue of low response rates has a long tradition in the direct marketing literature (e.g., Magliozzi & Berger, 1993). It is also known that a skewed distribution of the target variable impedes regression modeling, for example, in the context of credit risk models for loss given default prediction (Yao et al., 2017). While remedies to skewed response distributions in the form of zero-inflated regression and hurdle models are well-established in other domains, we are not aware of previous work in uplift modeling that addresses skewed distributions of the target variable through specific modeling solutions. A second contribution of the paper is that it introduces and tests a two-stage modeling framework for revenue uplift modeling.

A third contribution concerns the evaluation of (revenue) uplift models. We propose a novel profit decomposition for marketing campaigns to measure the incremental profit impact of an uplift model. The campaign profit measure is easily interpretable and bridges the gap between model-based forecasts and financial business goals. Incorporating a comprehensive cost model, the profit measure supports different types of marketing campaigns, including, but not limited to, coupon targeting and extends previous decompositions of campaign profit (e.g., Lessmann et al., 2019; Verbeke et al., 2012) to uplift modeling.

Last, capturing the impact of a marketing action on customer behavior, uplift modeling is closely related to the literature on causal inference (e.g., Imbens & Rubin, 2015). Using the terminology of that literature, uplift modeling corresponds to estimating conditional average treatment effects (CATE), where the conditioning is based on customer characteristics (Knaus et al., 2018). Aiming at comparing customers, an uplift model emphasizes individual-level effects, which correspond to the finest level of conditioning. Since the fundamental problem of causal inference (Holland, 1986) renders individualized effects unobservable, a more common term is that of an individualized average treatment effect; sometimes also called personalized treatment effect (Guelman et al., 2015a). In principle, any approach for CATE estimation facilitates uplift modeling. However, we note a subtle difference between an uplift model and a CATE estimator, which arises in a campaign targeting context. Targeting a marketing action aims at an accurate ranking of customers (e.g., Lessmann et al., 2019). A marketer can then approach the top-ranked customers, whereby the ranking criterion is the CATE. First estimating individualized treatment effects and then sorting customers accordingly is a viable approach toward uplift modeling. However, a model that gives biased estimates of the CATE – or foregoes its estimation altogether – can still be a valuable uplift model if the relative order of customers in the ranking is accurate. Uplift modeling strategies such as the class variable transformation, which display appealing results in several benchmarks (Devriendt et al., 2018; Gubela et al., 2019; Kane et al., 2014), and extensions as proposed here, belong to the latter category. In view of recent advancements in the literature on treatment effects and the development of several highly recognized methods such as causal forests (Athey et al., 2019), causal boosting (Powers et al., 2018), causal Bayesian Additive Regression Trees (BART) (Hill, 2011), or the x-learner of Künzel et al. (2019), it is interesting to examine whether these CATE estimators are a suitable vehicle for revenue uplift modeling. Finally, the paper makes an empirical contribution. We obtained a large quasi-experimental data set associated with targeting digital coupons to stimulate online sales from our industry partner Akanoo. Using this data, we assess the effectiveness and profitability of the proposed revenue uplift modeling strategies, recently developed causal machine learning algorithms and several benchmark models in a cross-sectional setup.

2 Background and Related Work

Marketing decision models for differential response analysis (Radcliffe & Surry, 1999) or incremental value modeling (Chickering & Heckerman, 2000) have a long tradition. A seminal paper by Lo (2002) coined the term true lift, which has later been revised to uplift (Radcliffe & Surry, 2011). Most previous studies explain in detail how an uplift model differs from a classical response model to estimate causal effects (e.g., Lo & Pachamanova, 2015). However, especially earlier papers do not emphasize connections between uplift and the literature on treatment effects. Our review aims at providing a comprehensive overview of previous work in different fields and substantiating how the paper adds to the literature.

We consider previous studies on uplift modeling and machine learning methods for CATE estimation. The focus on machine learning is suitable because marketing data sets are typically large and high-dimensional. Recently developed causal machine learning algorithms highlight such data characteristics as design goals and innovation over more traditional methods for treatment effect estimation (Knaus et al., 2018). This suggests that machine learning-based models are well prepared for marketing applications. Overall, our review of prior work identifies 37 studies that have contributed novel methodology to the field of uplift analytics. We organize these studies along the stages of the knowledge discovery in databases (KDD) process model (Fayyad et al., 1996) to identify their methodological contributions in Table 1. The KDD process identifies modeling steps that occur in every data analytics initiative. Therefore, it provides a suitable framework to systematize the previous interdisciplinary work on uplift modeling. Note that Table 1 excludes recent benchmarking studies related to (conversion) uplift and CATE (Devriendt et al., 2018; Gubela et al., 2019; Knaus et al., 2018). These papers offer empirical insight but do not contribute novel methodology.

To capture a causal link between a marketing action and customer behavior, uplift models require data from two groups, the treatment and the control group (Devriendt et al., 2018). Such data is gathered through randomized trials in previous work, often in the form of A/B tests in e-commerce. The existence of two disjoint sets of data represents a major difference to conventional applications of SML. KDD stages need to address this difference. Some studies have explicitly considered issues related to data source selection (e.g., Diemert et al., 2018). Likewise, handling treatment and control group data requires adjustments of data preparation tasks as examined by a few studies (e.g., Hansen & Bowers, 2008) and necessitates novel measures to assess model quality (e.g., Nassif et al., 2013).

However, Table 1 clarifies that the focus in most previous studies was to develop novel algorithms for the estimation of uplift models (i.e., CATE). Early uplift papers (e.g., Hansotia & Rukstales, 2002b) and recent papers in causal machine learning (e.g., Athey et al., 2019; Powers et al., 2018) consider a tree learning framework and propose splitting criteria for CATE estimation. Their partitioning mechanism makes decision trees especially suitable to process different subsets of data (e.g., treatment/control group observations). However, other popular SML algorithms, such as support vector machines (e.g., Kuusisto et al., 2014) or neural networks (e.g., Shalit et al., 2017), have also been extended to uplift settings.

The focus of this paper is on the data transformation step of the KDD process. This step is interesting because a smart transformation of the input (i.e., covariate) or output (i.e., target variable) space of a modeling data set facilitates the development of an uplift model using standard SML algorithms. From a practitioner's point of view, avoiding the development of a tailor-made learning algorithm has advantages. It facilitates capitalizing on the broad set of available SML algorithms and their scalable implementations in software packages (e.g., Kochura et al., 2017). Tailor-made causal machine learning algorithms such as causal forests (Athey et al., 2019) or causal boosting (Powers et al., 2018) show excellent results but do not offer algorithmic flexibility. A tree learning framework may not be the best choice

for a data set. Being able to choose any learning algorithm for uplift estimation is beneficial. The two-stage revenue uplift models, which we propose below to address zero-inflated revenue distributions, are a good example of the merit of (re-)using SML algorithms for uplift estimation. While the development of a corresponding approach, for example, a zero-inflated causal neural network, is an exciting avenue for research, marketing professionals will appreciate the more straightforward solution to devise a model based on existing, proven SML methodology. Also, implementations of recently developed causal machine learning algorithms for high-performance computing frameworks (e.g., Spark) are not yet available. This may prohibit processing the huge amounts of data that routinely occur in corporate marketing. Last, the process of developing a targeting model utilizing uplift transformations is almost the same as in conventional direct marketing. Providing a gradual transition of established targeting practices to a causal modeling framework and leveraging existing technology, we suggest that uplift transformation approaches are suitable to support campaign planning.

Table 1: Prior work on uplift modeling and machine learning for CATE

Study	Research Focus in terms of KDD Process						
			Uplift Transformation				
	Data Selection	Data Pre-Processing	Conversion Response Transformation	Revenue Response Transformation	Covariates Transformation	Uplift / CATE Estimation	Evaluation
Athey and Imbens (2016)						Causal Trees	
Athey et al. (2019)						Causal Forests	
Cai et al. (2011)						Two-Step Estimation Procedure	
Chickering and Heckerman (2000)						Uplift Tree with Post-Processing Procedure	
Diemert et al. (2018)	x					-	
Guelman et al. (2015a)						Causal Conditional Inference Trees/Forests	
Guelman et al. (2015b)						Uplift Random Forests	
Gutierrez and Gérardy (2017)						-	x
Hahn et al. (2019)						Causal Bayesian Regression Trees	
Hansen and Bowers (2008)		x				-	
Hansotia and Rukstales (2002a)						Incremental Response Trees	
Hansotia and Rukstales (2002b)						Uplift Trees with the $\Delta\Delta P$ splitting criterion	
Hill (2011)						Causal BART	
Imai and Ratkovic (2013)						Uplift Support Vector Machine	
Jaroszewicz and Rzepakowski (2014)						Uplift k-Nearest Neighbors	
Kane et al. (2014)	x					-	
Kuusisto et al. (2014)						Uplift Support Vector Machine	
Künzel et al. (2019)						X-Learner	
Lai et al. (2006)			x			-	
Lechner (2019)						Modified Causal Forests	
Lo (2002)					x	-	
Lo and Pachamanova (2015)						Multiple Treatments Logistic Regression	
Nassif et al. (2013)						-	x
Oprescu et al. (2019)						Orthogonal Causal Random Forests	
Powers et al. (2018)						Causal Boosting	
Radcliffe and Surry (1999)						Uplift Trees	
Radcliffe and Surry (2011)						-	x
Rzepakowski and Jaroszewicz (2012a)						Multiple Treatments Uplift Trees	
Rzepakowski and Jaroszewicz (2012b)						Information Theory-Based Uplift Trees	
Rudaś and Jaroszewicz (2018)				x		-	
Shaar et al. (2016)						Reflective and Pessimistic Uplift Modeling	
Shalit et al. (2017)						Causal Artificial Neural Network	
Sołtys et al. (2015)						Uplift Ensemble Methods	
Su et al. (2012)						Uplift k-Nearest Neighbors	
Taddy et al. (2016)						Causal Bayesian Forests	
Tian et al. (2014)					x	-	
Yamane et al. (2018)						Separate Label Uplift Modeling	
This study				x		-	x

Note: Only a study's primary contribution has been considered for the systematization of prior work.

Table 1 displays that previous literature on uplift transformation is relatively sparse. Lo (2002) and Tian et al. (2014) use (logistic) regression to estimate uplift. To capture causality, they augment their models with interaction terms of a binary treatment/control group affiliation variable with other covariates. We formally introduce this approach in Section 3. Lai's weighted uplift method (LWUM) pursues an alternative strategy and transforms the target variable (Lai et al., 2006). Lai's approach was later refined by Kane et al. (2014). Both regimes, covariate and response transformation, facilitate the use of SML algorithms to model uplift. The above studies have considered binary target variables (e.g., conversion uplift). The focus of this paper is revenue uplift, which involves processing a continuous target variable. Processing continuous target variables in a covariate transformation framework is straightforward and has

been considered by Tian et al. (2014). As will become clear in Section 3, a corresponding extension to support continuous target variables involves major adjustments and offers the response variable transformation approach. Developing response variable transformations for continuous target variables and revenue uplift modeling is thus the focus.

Uplift data mining techniques based on decision trees are an alternative to process continuous response variables. In the empirical study, we consider tree-based models for uplift and CATE estimation as benchmarks. Since none of the tree-based techniques has been employed for modeling revenue uplift in e-commerce, the evaluation broadens the scope of empirical results for causal machine learning methods.

3 Revenue Uplift Modeling

In the following, we review uplift modeling fundamentals and introduce our notation. Thereafter, we elaborate on the target variable and covariate transformations for revenue uplift modeling. Finally, we introduce our two-stage models to address zero-inflated revenue distributions.

3.1 Uplift Modeling Fundamentals

Let $X_i = (x_{i1}, \dots, x_{in}) \in \mathbb{R}^n$ be a vector of size n that represents individual customers. The elements of X_i capture customer characteristics such as demographic or behavioral attributes. Let $Y_{i,c} \in \{0,1\}$ be the binary response in a conversion model, with $Y_{i,c} = 1$ indicating, for example, the purchase of a product. The index $i = 1, \dots, N$ refers to individual customers. Throughout the paper, we use the subscripts c and r to distinguish between a conversion and revenue setting. Further, let $T_i \in \{0,1\}$ be an indicator of a customer's group affiliation, with $T_i = 0$ and $T_i = 1$ indicating control and treatment group customers, respectively. As purchase probabilities equal response expectations, or more formally, $P(Y = 1) = E(Y)$, let $E(Y_i|X_i, T_i = 1)$ and $E(Y_i|X_i, T_i = 0)$ denote the group-specific response expectations. This notation facilitates a formal delineation of response and uplift models. Response models aim at predicting (conversion) model lift:

$$\text{Lift}_{i,c}^{\text{Response}} = E(Y_{i,c}|X_i, T_i = 1)/E(Y_{i,c}) = E(Y_{i,c}|X_i, T_i)/E(Y_{i,c}). \quad (1)$$

The predictions are then used to select customers for which this quantity is the largest. Given that the prior expectation of responding, $E(Y_{i,c})$, is a characteristic of the data, the primary task of a response model is to estimate the posterior expectations $E(Y_{i,c}|X_i, T_i = 1)$. Uplift models predict the change in behavior resulting from a marketing action, which is equivalent to a change in the posterior expectation with and without treatment and thus the causal effect of the treatment (e.g., Imbens & Rubin, 2015):

$$\text{Uplift}_{i,c}^{\text{Indirect}} = E(Y_{i,c}|X_i, T_i = 1) - E(Y_{i,c}|X_i, T_i = 0). \quad (2)$$

Estimated class expectations, also called lift and uplift scores, respectively, facilitate a ranking of customers and targeting top-ranked customers with a campaign (e.g., Devriendt et al., 2018).

Equation (2) illustrates a possible strategy to develop a conversion uplift model by estimating two classification models from treatment and control group observations, respectively, using some SML algorithm. Subtraction of the expectations on the two mirrors the idea of uplift modeling, the objective of which is to maximize the number of treatment responders while minimizing control responders based on model predictions. This strategy is known as indirect uplift modeling (e.g., Kane et al., 2014) as it predicts uplift using two independent classifiers. The indirect uplift approach suffers from conceptual drawbacks that often lead to poor model performance (Hansotia & Rukstales, 2002b), which has motivated the development of alternative conversion uplift modeling strategies (see Table 1).

Our research focusses on revenue uplift modeling. We first introduce the most basic strategies, which are revenue response modeling and the indirect revenue uplift approach. In contrast to their analogs in conversion modeling, these differ in that they take a continuous response (revenue) $Y_{i,r} \in \mathbb{R}$ instead of a binary response (conversion) $Y_{i,c}$ into account. Formally, revenue response models predict:

$$\text{Lift}_{i,r}^{\text{Response}} = E(Y_{i,r}|X_i, T_i = 1), \tag{3}$$

whereas indirect revenue uplift models predict uplift by

$$\text{Uplift}_{i,r}^{\text{Indirect}} = E\big(Y_{i,r}\big|X_i, T_i = 1\big) - E\big(Y_{i,r}\big|X_i, T_i = 0\big). \tag{4}$$

Irrespective of a modeling strategy, revenue uplift models are developed from a data set $\{Y_{i,r}, X_i, T_i\}_{i=1}^{N}$. In marketing, such data stems from a previous campaign or may be gathered using A/B tests. More generally, randomized trials are a typical way to obtain modeling data. In our setting, $Y_{i,r}$ represents the total purchase amount of customer i, X_i includes characteristics of this customer, and T_i indicates whether the customer has received a marketing stimulus. We assume conditional independence, which implies that values of T_i have been generated at random independent from customer characteristics. This assumption is critical to mitigating the bias of model estimates (Imbens & Rubin, 2015).

3.2 Revenue Uplift Transformations

3.2.1 Revenue Response Transformation

We propose two response transformation approaches for revenue modeling, which we call continuous response variable transformation with weightings (CRVTW), and revenue discretization transformation (RDT). The approaches aim to transfer the treatment/control group information from the input space (i.e., independent variables) to its output space (i.e., the dependent variable).

CRVTW transforms the response variable in such a way that it captures critical information from treatment and control groups. More specifically, we define $z_{i,rw} \in \mathbb{R}$ as follows:

$$z_{i,rw} = \begin{cases} +\frac{1}{q_\tau} Y_{i,r} & \text{if } T_i = 1 \ \wedge \ Y_{i,r} > 0 \\ -\frac{1}{q_\varsigma} Y_{i,r} & \text{if } T_i = 0 \ \wedge \ Y_{i,r} > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

with $q_\tau = N_\tau/N$ and $q_\varsigma = N_\varsigma/N$ as the fractions of treatment or control group customers relative to the whole population. The weightings facilitate unbiased estimation. For formal proofs on statistical properties, we refer to Rudaś and Jaroszewicz (2018). The transformed response variable is positive if the customer responded positively to the incentive and provided a purchase value of $Y_{i,r}$. For purchasers of the control group, $Y_{i,r}$ corresponds to the negative purchase volume. The remaining cases include treatment and control group customers without purchase, and we define $z_{i,rw}$ to be zero in these cases. We denote the number of observations in total and in the treatment and control group by N, N_τ and N_ς, respectively.

CRVTW draws inspiration from Lai's method for conversion uplift modeling (Lai et al., 2006), which, in the light of recent benchmarking results (e.g., Devriendt et al., 2018; Kane et al., 2014), can be considered a promising modeling strategy for conversion uplift. Equation (5) extends this approach to continuous response variables. A standard regression model suffices to model $z_{i,rw}$, which itself possesses all relevant information to capture uplift. Using some regression method to learn a functional relationship between $z_{i,rw}$ and X_i provides a model that distinguishes purchasing customers from the treatment and control groups while also capturing the dependency between covariates and purchase amounts. We assume that targeting customers according to the predictions of such an uplift model (i.e., soliciting customers in descending order of scores from a prediction of $z_{i,rw}$) leads to a target group of persuadable customers who spend a lot if converted. Following the same logic, we assume scores from predicting $z_{i,rw}$ for customers who are ready to buy without solicitation and thus not in need of a marketing investment to be negative, resulting in them being ignored in targeted marketing actions. CRVTW represents an uplift transformation and facilitates the development of a revenue uplift model through estimating (6) using some SML algorithm for regression:

$$\text{Uplift}_{i,r}^{\text{CRVTW}} = E\left(z_{i,rw} \middle| X_i\right). \quad (6)$$

The second transformation we propose, RDT, is based on Bodapati and Gupta (2004), who recommend a discretization of continuous responses in direct marketing models. In the context of response modeling, Bodapati and Gupta (2004) demonstrate that the discretization decreases bias and increases accuracy. In contrast to forecasting the annual number of product purchases individually, the managerial challenge is to predict whether this number exceeds a pre-defined threshold to determine the value of a discretization function d(y), which they define as follows:

$$d(y) = \begin{cases} 0 & \text{if } y \in (0, y_{\text{threshold}}] \\ 1 & \text{if } y \in (y_{\text{threshold}}, \infty) \end{cases} \quad (7)$$

with $y_{threshold}$ as the pre-defined value of the absolute number of purchases as a typical example of a marketing application. From here, supervised classifiers can be built on the resulting function. Note that the discretization can be based on a revenue variable as well.

We extend (7) for revenue uplift modeling. To that end, we propose to first obtain $z_{i,rw}$ through applying CRVTW and to then convert $z_{i,rw}$ into a dichotomous response variable $z_{i,rg}$ through (8). Rather than predicting $z_{i,rw}$ with regression models, our second response transformation creates a binary target variable, which we model using SML algorithms for classification. Formally, RDT extends the CRVTW transformation with the following discretization scheme:

$$z_{i,rg} = \begin{cases} 0 & \text{if } z_{i,rw} \in (-\infty, 0] \\ 1 & \text{if } z_{i,rw} \in (0, \infty) \end{cases} \tag{8}$$

with $z_{i,rg} \in \{0,1\}$. A crucial difference between (8) and the discretization of Bodapati and Gupta (2004) is that the response variable has been pre-transformed and that negative numbers are captured in $z_{i,rg}$ because of control group purchasers. This points out that $z_{i,rg}$ embodies information related to the treatment and control group, which is imperative for uplift modeling. In summary, RDT estimates uplift as:

$$\text{Uplift}_{i,r}^{RDT} = E(z_{i,rg}|X_i). \tag{9}$$

The reason why we set the threshold to zero in (8) is related to the analytical objective in the context of uplift modeling. The definition of a "failure" is $z_{i,rg} = 0$ and includes non-incentivized purchasers ($z_{i,rw} = -\frac{1}{q_c} Y_{i,r}$), incentivized non-purchasers ($z_{i,rw} = 0$), and non-incentivized non-purchasers ($z_{i,rw} = 0$). In contrast, the definition of "success" relates to customers who have purchased a product with the causal connection to the marketing action ($z_{i,rw} = \frac{1}{q_t} Y_{i,r}$).

3.2.2 Covariates Transformation

Table 1 indicates alternative strategies that entail transformations of the data input space. Although not the focus of this paper, covariates transformation approaches also provide a framework to perform revenue uplift modeling using SML algorithms. The interaction term method (ITM) by Lo (2002) and the treatment-covariates interactions approach (TCIA) by Tian et al. (2014) both consider dichotomous response variables. However, extensions to continuous responses for accommodating revenues are straightforward.

The ITM approach uses the binary treatment/control group affiliation variable T_i and augments the prediction model with an interaction term $X_i \cdot T_i$ to forecast revenue uplift as:

$$\text{Uplift}_{i,r}^{ITM} = f(X_i, T_i, X_i \cdot T_i). \tag{10}$$

Lo (2002) builds two equivalent uplift models on the treatment and control samples, whereby the covariate space in each sample was extended with the interaction term. To calculate uplift, the scores of the two

models are subtracted. While Lo (2002) uses logistic regressions (LogR), his approach can accommodate arbitrary SML algorithms. Reviewing (10), it is apparent that the additional interaction term modifies the predictions of treated observations in contrast to control observations.

3.3 Two-Stage Revenue Uplift Models

We introduce novel models to account for zero-inflated revenue distributions. The purest form of a zero-inflated model is a two-stage model where a classification model predicts whether the response will be zero (e.g., to identify purchase incidents), and a regression model, estimated from non-zero responses, predicts response values (e.g., a buyer's purchase volume). More formally, a classification model estimates $E(Y_{i,c}) = f_c(X_i)$ and a regression model predicts $E\big(Y_{i,r}[Y_{i,r} > 0]\big) = f_r(X_i)$.

We multiply the separate model forecasts to calculate an expected purchase amount. For hurdle models, decomposing the log-likelihood into two independent parts and maximizing them separately is a valid approach as the estimated probabilities of both parts can be formulated as an additive term (e.g., Hofstetter et al., 2016). The same holds for zero-inflated Poisson regressions (Lambert, 1992). Using linear regression models for the second stages in our two-stage models, we do not see a structural difference to hurdle models and estimate the component models for purchase incident and value independently. The occurrence of two separate, sequential customer decisions in our e-commerce context further reinforces this argumentation.

The methodological setup of two-stage models differs across revenue uplift strategies due to varying response types. More specifically, the transformation of CRVTW prepares a response variable with positive (negative) values for buyers from treatment (control) groups and zero values for non-purchasers (see (5)). We apply a classification model to identify buyers and a regression model on the transformed revenue response. RDT outputs a binary response variable (see (8)), and can be considered a special case of a two-stage model. We fit a classification model on the transformed response and a regression model on the revenue variable to predict purchase volumes. In terms of RDT, we further introduce the synthetic minority oversampling technique (SMOTE), which is a popular approach to address imbalanced class distributions through creating synthetic minority class examples (Chawla et al., 2002). We build a SMOTE model on the discretized revenue response and use a classification model for prediction.

For the indirect revenue uplift strategy, following (4), we consider a two-stage model on the treatment sample and a separate two-stage model on the control sample. For each of these models, we use a classification model to forecast purchase incidents and a regression model to predict purchase volumes, and multiply the obtained predictions to derive model scores. We then subtract the scores of the control group models from the scores of the treatment group models. ITM-based two-stage models consider an additional interaction term for the second stage regression models on the treatment and control samples (see (10)).

4 Campaign Profit for Uplift

To measure the incremental profit impact of an uplift model, we develop a profit decomposition. The novel measure details costs related to coupon targeting. Extensions to other marketing settings are straightforward. Hansotia and Rukstales (2002b) develop a break-even decision rule that is estimated by expected profit subtracted by costs per contact. In contrast to their rule, we decompose campaign profit for uplift into expected costs associated with the marketing campaign and differential gain in revenue, define profit as a function of lift, distinguish different types of costs, and facilitate flexibility in the choice of arbitrary direct and indirect uplift models. To derive this profitability, we start from marketing campaign profit, Ω, which we define as follows (e.g., Lessmann et al., 2019):

$$\Omega^{\text{Lift}} = \text{N} * \tau * (\pi_+ * \gamma(\tau) * \delta - \varepsilon), \tag{11}$$

with N referring to the total population of customers, τ the share of targeted customers, π_+ the fraction of customers who respond to the campaign, $\gamma(\tau)$ the lift index, which is given as $\gamma(\tau) = \pi_\tau/\pi_+$ where π_τ indicates the share of positive reactions within the campaign target group (e.g., Neslin et al., 2006), and δ and ε denote revenue and costs, respectively. We reformulate campaign profit as $\Omega^{\text{Lift}} = \text{N} * \tau * (\pi_+ * \pi_\tau/\pi_+ * \delta - \varepsilon) = \text{N} * \tau * (\pi_\tau * \delta - \varepsilon) = \text{N}_\tau * \pi_\tau * \delta - \text{N}_\tau * \varepsilon$ with $\text{N}_\tau = \text{N} * \tau$ for simplification.

Martens and Provost (2011) express campaign profit as a function of the lift measure. To generalize their measure to uplift modeling, we develop a profit equation that captures incrementality. We therefore modify the term $\text{N}_\tau * \pi_\tau * \delta$ by $\text{N}_\tau * \pi_\tau * \delta_\tau - \text{N}_\varsigma * \pi_\varsigma * \delta_\varsigma$, with $\text{N}_\varsigma, \pi_\varsigma$ as control group equivalents to the quantities N_τ, π_τ and δ_τ and δ_ς as the average revenue in the treatment and control groups, respectively. With this, we add necessary control group information to the revenue side of the equation. Hence, we express campaign profit for uplift as $\Omega^{\text{Uplift}} = (\text{N}_\tau * \pi_\tau * \delta_\tau - \text{N}_\varsigma * \pi_\varsigma * \delta_\varsigma) - \text{N}_\tau * \varepsilon$.

The term $\text{N}_\tau * \varepsilon$ on the cost side implies that only contact costs occur. Campaigns may require a more comprehensive cost model. We therefore generalize uplift campaign profit as:

$$\Omega^{\text{Uplift}} = (\text{N}_\tau * \pi_\tau * \delta_\tau - \text{N}_\varsigma * \pi_\varsigma * \delta_\varsigma) - \varepsilon_{\text{treatment}} \tag{12}$$

with $\varepsilon_{\text{treatment}}$ reflecting all treatment-related costs of a targeting campaign. First, $\varepsilon_{\text{treatment}}$ embodies contact costs, $\varepsilon_{\text{contact}}$. They occur whenever a customer receives a treatment and depend on the number of targeted customers. Contact costs are thus a function of the target fraction and c_{unit} states the constant unit costs for performing a promotional action. This suggests $\varepsilon_{\text{contact}}(\tau) = \text{N}_\tau * \varepsilon_{\text{unit}}$. Depending on the area of application, unit costs vary from zero or near-zero (e.g., automated e-couponing) to several euros per transaction (e.g., outbound call campaigns with personal customer support).

In addition to contact costs, a second component of $\varepsilon_{\text{treatment}}$ in (12) refers to the costs of the incentive, which the campaign offers solicited customers. Such costs are especially relevant for couponing where a discount is offered to persuade customers to buy. They occur as soon as a treated customer accepts the

marketing offer. In the following, we focus on campaigns that offer a relative discount and define $\rho \in \{0.05, \ldots, 0.95\} \subseteq \mathbb{R}$ to be the corresponding price reduction. Typically, promotion values are in intervals of five (i.e., 5%, 10%, etc.). The choice of a relative discount has been suggested by Akanoo. It does not constrain the campaign profit formulation and can be exchanged with an absolute discount value or other cost factors when needed. Discounts limit revenue to the extent of the financial value of the promotional incentive. Hence, we describe incentive costs as $\varepsilon_{\text{incentive}}(\tau, \delta_{\pi_+}, \rho) = \rho * \sum_{i=1}^{N_\tau} \delta_{\pi_+,i}$. In contrast to contact costs, incentive costs depend on the individual customer's reaction and shopping basket size.

Based on contact and incentive costs, we modify campaign profit from (12) as follows:

$$\begin{aligned} \Omega^{\text{Uplift}} &= (N_\tau * \pi_\tau * \delta_\tau - N_\varsigma * \pi_\varsigma * \delta_\varsigma) - \varepsilon_{\text{treatment}} \\ &= (N_\tau * \pi_\tau * \delta_\tau - N_\varsigma * \pi_\varsigma * \delta_\varsigma) - (\varepsilon_{\text{contact}}(\tau) + \varepsilon_{\text{incentive}}(\tau, \delta_{\pi_+}, \rho)) \\ &= (N_\tau * \pi_\tau * \delta_\tau - N_\varsigma * \pi_\varsigma * \delta_\varsigma) - N_\tau * \varepsilon_{\text{unit}} - \rho * \sum_{i=1}^{N_\tau} \delta_{\pi_+,i} \end{aligned} \quad (13)$$

Equation (13) presents the fundament for marketing investment decisions on a decile-level. We assess revenue uplift models in terms of (13) and other metrics and compare their financial impacts to benchmarks.

5 Experimental Setup

5.1 Campaign and Data

Akanoo, an online marketing agency headquartered in Germany, provided the data for this study. The marketing campaign is based on a real-time targeting process where uplift models score visitors of an online shop according to their browsing behavior. Targeted customers are offered a ten percent discount off their shopping basket (i.e., $\rho = 0.1$). This setting displays typical characteristics of targeted marketing in that the coupon should only be offered to persuadable customers to avoid a waste of marketing resources.

Akanoo gathers data as follows. A random forest-based propensity model pre-screens shop visitors through predicting purchase expectations (Baumann et al., 2018). Visitors with a high likelihood of buying are not eligible for a discount. The remaining visitors are randomly distributed into a treatment and control group at a ratio of 3:1. Treatment group visitors are offered the discount, which they can use in the current browsing session. Purchases and basket values of treatment and control group visitors are recorded and available in our data. Although the filtering mechanism, which Akanoo imposes to avoid soliciting likely buyers, only excludes a small fraction of visitors from the chance of receiving a coupon, it creates a quasi-experimental setting (e.g., Armstrong & Patnaik, 2009). While a truly randomized trial was preferable to estimate the causal effect of the digital coupon, the focus of this paper is to examine the relative

effectiveness of alternative uplift modeling strategies. The data facilitates a comparison of different models on equal ground, which confirms its suitability for the study.

Comprising 2,951,313 observations and 60 variables, the data facilitates a large-scale empirical analysis. Since the data comes from twenty-five European online shops, it also provides a broad view across different product assortments and visitor patterns. Each observation represents an individual customer session, that is, a customer journey from shop access until leave. Covariates regard the time, views, baskets, prior visits and technical characteristics and split into numeric and categorical variables. Detailed information on the data set and independent variables are available in the online appendix (see Figure A1.1 and Table A1.1).

Uplift statistics such as the signal-to-noise ratio clarify campaign effectiveness (Kane et al., 2014). Such measures compare responders from the treatment/control groups to determine the initial (conversion) uplift in the data. We add revenue uplift statistics and report the corresponding information in Table 2.

Table 2: Descriptive uplift statistics of experimental data

Group Affiliation	Share (%)	Customer Sessions	Purchasers	Conversion Rate (%)	Conversion Uplift (%)	Revenue (€)*	Revenue Per Person (€)*	Revenue Uplift (€)*
Treatment	74.9	2,210,190	162,570	7.35	0.16	178,216	0.08	0.05
Control	25.1	741,123	53,340	7.19		58,149	0.03	
Total	*100*	*2,951,313*	*215,910*	-	-	*236,365*	-	-

***** For business-related reasons, the data contains normalized revenue.

Conversion uplift has also been expressed as the average treatment effect in previous studies (Guelman et al., 2015a). With a conversion and (normalized) revenue uplift of 0.16% and 0.05€, respectively, the overall uplift signal is rather low and not significant (p-value of Pearson's X^2 test <0.000).

5.2 Uplift Models and Learning Algorithms

Table 3 summarizes the strategies to develop revenue uplift models, their category and formalization, as well as the underlying learning algorithm. We focus on the two approaches for response transformations. The other strategies emerge as extensions of their conversion uplift equivalents (e.g., Kane et al., 2014).

Table 3: Revenue uplift strategies

Revenue Uplift Strategy	Category	Model Formalization	Learning Algorithm
Continuous Response Variable Transformation with Weightings (CRVTW)	Response Transformation	(5) and (6)	Regression
Revenue Discretization Transformation (RDT)	Response Transformation	(8) and (9)	Classification
Interaction Term Method (ITM)	Covariates Transformation	(10)	Regression
Indirect Revenue Uplift Strategy (INDIRECT)	Two Model Approach	(4)	Regression

Implementing a conversion or revenue uplift strategy requires a classification or regression learning algorithm. As RDT models a binary response variable (see (8)), implementing a RDT revenue uplift model

requires an underlying classification algorithm. We implement all other revenue uplift strategies using SML algorithms for regression. For classification, we consider tree-based algorithms (e.g., random forest, gradient boosting classifiers (GBC), and extremely randomized trees (ERT)), k-nearest neighbors (KNN), support vector classifiers (SVC), logistic regressions, linear discriminant analysis with shrinkage (LDA_WS) and quadratic discriminant analysis (QDA). For regression, we use random forest regressors (RFR), support vector regressors (SVR), neural networks – multi-layer perceptron regressors (MLP) – and specific algorithms like the least absolute shrinkage and selection operator (LASSO) for least angle regression (LARS), abbreviated as LASSO LARS (LL), and Theil-Sen regressors (TSR). Table A1.2 and Table A1.3 in the online appendix provide details on the algorithms' meta-parameters, candidate settings, and the number of related models for the classification and regression tasks. These numbers emerge from estimating a candidate model for every combination of meta-parameter settings per learning algorithm during model selection (e.g., Lessmann et al., 2015). Hastie et al. (2009) provide a detailed description of the employed methods.

We also consider parametric models to implement revenue uplift strategies. The one-stage revenue uplift models are logistic regression models, linear discriminant analysis (LDA) classifiers without shrinkage, and ordinary least squares (OLS) regression models. Hence, we apply LogR and LDA with the RDT strategy to solve a classification task and OLS with the remaining strategies to solve a regression task. For all strategies, we consider LogR and LDA as classification models for the first stage and OLS regression models for the second stage. For RDT, we further consider SMOTE. Table 4 summarizes the parametric models.

Table 4: Parametric revenue uplift models with one- and two-stage base learners

	RDT	CRVTW	ITM	INDIRECT
One-Stage	LogR LDA	OLS	OLS	OLS
Two-Stage	LogR, OLS LogR (SMOTE)	LogR, OLS	LogR, OLS	LogR, OLS
	LDA, OLS LDA (SMOTE)	LDA, OLS	LDA, OLS	LDA, OLS

5.3 Meta-Parameter Tuning and Performance Evaluation

We randomly partition the data into a training set (40% or 1,180,525 observations), a validation set (30% or 885,394 observations), and a hold-out test set (30% or 885,394 observations). Each of these sets includes browser sessions from the treatment and control group. We use the training set to develop uplift models using the learning algorithms of Table A1.2 and Table A1.3 (see online appendix). To illustrate this, consider, for example, the first learner from Table A1.3, which is a LL model. To implement an indirect revenue uplift strategy according to (4), we estimate one LL model from the treatment group observations of the training set and another regression model from the control group observations of the training set. We then score observations in the validation set with both regression models and calculate the differ-

ences between the two models' scores. This difference represents an uplift score, which we use for assessment. We repeat the process for the other revenue uplift strategies (Table A1.3) and develop corresponding uplift models based on LL regression. Next, we utilize another learning algorithm and repeat; considering different meta-parameter settings whenever a learner exhibits meta-parameters.

The validation set facilitates identifying the best combination of a learning algorithm and meta-parameter configuration per uplift strategy. For example, we may find the best ITM revenue uplift model to come from a ridge regression (Ridge) learner with a regularization strength of 0.01 (see Table A1.3). We use this specification and re-estimate the best ITM revenue uplift model from the union of the training and validation set (covering 70% of the whole data) to obtain our final ITM revenue uplift model, which enters subsequent comparisons to other uplift strategies and benchmarks on the test set. To ensure the robustness of results, we repeat the random partitioning of the data into training, validation, and test set – and, thus, all intermediate steps described above for one random partitioning – ten times. As we do not perform meta-parameter optimization for the analysis of one-stage vs. two-stage models, we thereby perform model training on both the training and validation sets and predict on the hold-out test set.

Due to the fundamental problem of causal inference (Holland, 1986), which points to the impossibility of simultaneously targeting and not targeting the same customer, our analyses require uplift-specific performance measures. The Qini coefficient and Qini curves, generalizations of the Gini coefficient and gain charts for uplift models, have been developed for this purpose (Radcliffe, 2007). Our focus on revenue uplift modeling implies an extension of such measures to a revenue setting. To this end, the Revenue Qini coefficient Q_r considers the incremental gain in revenue, derived by calculating the decile-wise difference in the revenue metric between treatment and control group observations. Q_r is defined as the area between the respective Revenue Qini curve, which illustrates the performance of a revenue uplift model for each of the ten targeted deciles, and a diagonal line representing random targeting (Radcliffe & Surry, 2011).

To emphasize the flexibility in the choice of financial performance indicators for practitioners, we use varying expressions of Q_r and Revenue Qini curves for different parts of our empirical study. More specifically, these comprise revenue per person and total revenue, normalized revenue for comparisons across experimental settings, and a scaled version of Q_r to dissolve the dependency on the sample size (Radcliffe & Surry, 2011). Note that from a view of corporate practice, Revenue Qini curves better prepare operational decision-making compared to Q_r as they summarize model performance on a more granular level (i.e., per targeted decile). Therefore, we add different measures of Revenue Qini curves. For more detailed descriptions of Qini-related performance measures, we refer to the online appendix (see Subsection A1.2). Next to Q_r and Revenue Qini curves, we assess campaign effectiveness in terms of campaign profit for uplift as derived in Section 4 (all in euros) and model computation times (in seconds).

6 Empirical Analysis of Uplift Transformations for Revenue Uplift Models

We split the evaluation of the proposed revenue uplift transformations into three parts. Revenue uplift transformations require an underlying learning algorithm. The first part clarifies, for each transformation, whether this algorithm should be a standard SML algorithm or a two-stage model. We restrict this part to parametric models. Drawing upon corresponding results, part two assesses several configurations of our revenue uplift transformations using a rich set of SML algorithms, either in a one- or a two-stage regime, and together with meta-parameter tuning. The second analysis identifies the best way to implement revenue uplift modeling for the available data. In the third part, we empirically substantiate our claim that revenue uplift models give higher campaign revenue than uplift models for conversion. For better readability, the online appendix provides a list of all acronyms used in the empirical analysis.

6.1 Two-Stage Models for Zero-Inflated Revenue Distributions

We report empirical results obtained from instantiating revenue uplift transformations using one- or two-stage models in Table 5. We assess different models in terms of the Revenue Qini coefficient Q_r on the test set. Higher Qini values indicate better performance. In Table 5, we concentrate on parametric models and use OLS regression, LogR and LDA to develop uplift models. This way, we circumvent the need to tune algorithmic meta-parameters and simplify the analysis. The one-stage approach can be considered the standard way to develop an uplift model, whereas the two-step approach addresses the skewed distribution of the revenue target variable. Bold font highlights the best model.

Table 5: Revenue Qini coefficient of parametric revenue uplift models with one- and two-stage base learners

		Revenue Qini Coefficient	
Strategy	**Model**	**One-Stage**	**Two-Stage**
RDT	LogR	**0.021**	
	LogR (SMOTE)		0.011
	Two-Stage (LogR, OLS)		0.020
	LDA	0.017	
	LDA (SMOTE)		0.007
	Two-Stage (LDA, OLS)		0.010
CRVTW	OLS	**0.159**	
	Two-Stage (LogR, OLS)		0.149
	Two-Stage (LDA, OLS)		0.144
ITM	OLS	**0.136**	
	Two-Stage (LogR, OLS)		0.118
	Two-Stage (LDA, OLS)		0.049
INDIRECT	OLS	0.146	
	Two-Stage (LogR, OLS)		0.178
	Two-Stage (LDA, OLS)		**0.187**

Table 5 puts the two-stage approach into perspective. In most cases, it does not improve results over the one-stage approach. For RDT, simple one-stage implementations using LogR and LDA perform much

better than their SMOTE and two-stage model counterparts. This is also true for ITM and CRVTW, where the OLS implementation outperforms two-stage alternatives, independent of whether the latter is implemented using LogR or LDA. Only the indirect revenue uplift strategy benefits from addressing skewed revenue distributions using a two-stage model, whereby the LDA-based implementation is superior. The online appendix offers a more detailed view of alternative models' performances in terms of cumulative incremental normalized revenue per person along targeted deciles using Revenue Qini curves (see Figure A2.1). Corresponding results agree with Table 5 to a large extent. Based on these results, we develop RDT, CRVTW, and ITM uplift models using standard SML algorithms in subsequent experiments. For the indirect uplift model, we pursue a two-stage approach.

6.2 Analysis of Revenue Uplift Strategies and Learning Algorithms

This part of the analysis aims at evaluating the performance of the proposed revenue uplift strategies. The comparison also considers the interaction between a revenue uplift strategy and the underlying SML algorithm. We assess revenue uplift strategies in terms of Q_r, which we express as the (normalized) revenue per customer from test set observations, averaged over ten cross-validation iterations. Table 6 and Table 7 depict the empirical results for classification and regression algorithms, respectively. Table 7 also includes revenue response models. In both tables, we highlight the best, second-best, and third-best models using bold, italic face, and an underscore, respectively.

Table 6: Revenue Qini coefficient of classification-based RDT uplift models across learning algorithms

Revenue Strategy	Classification Algorithm							
	ERT	GBC	KNN	LDA_WS	LogR	QDA	RFC	SVC
RDT	**0.745**	0.555	0.462	0.376	0.461	0.266	*0.634*	0.242

Table 7: Revenue Qini coefficient of regression-based revenue uplift models across learning algorithms

Revenue Strategy	Regression Algorithm					
	LL	MLP	RFR	Ridge	SVR	TSR
CRVTW	-0.136	0.053	0.282	0.218	0.156	0.257
ITM	-0.049	0.041	-0.139	**0.572**	-0.114	0.283
Indirect Approach*	0.351	0.026	0.038	0.257	0.108	0.024
Response Modeling	-0.138	0.071	0.389	*0.473*	0.225	0.304

*According to Table 5, we implement this approach using a two-stage model (INDIRECT_TS). We use LDA for the first stage and consider several regression algorithms for the second stage, which undergo meta-parameter tuning.

RDT gives the best results together with an extremely randomized tree base learner. The top-three classifiers for RDT ground on tree-learning, which supports the popularity of decision trees in previous work (see Table 1). A tree-based learner, random forest regression, provides the best result for CRVTW, whereas ITM and the indirect approach work best when implemented using ridge regression and LASSO LARS, respectively. Interestingly, a revenue response model outperforms several uplift strategies in terms of Q_r although it disregards the causal effect of the treatment on customer behavior. We attribute this result to the data with relatively small revenue uplift (see Table 2). However, based on the results, we

emphasize that a careful selection of a revenue uplift strategy is crucial. While the response model outperforms several uplift model configurations, the best regression-based uplift model, ITM based on ridge regression, outperforms the best response model with a twenty-one percent margin.

Comparing the performance of the best classifiers with the best regressors, we observe the highest performance for the RDT-based extremely randomized tree, which is 30.2% better than the ITM-based ridge regression. RDT also contributes the second-best and fourth-best result across Table 6 and Table 7 using, respectively, random forest and gradient boosting implementation. CRVTW, on the other hand, never achieves a top rank. This is remarkable since RDT is based on a discretization of the continuous target variable that emerges from the CRVTW uplift transformation. Substantially better performance of RDT supports the results of Bodapati and Gupta (2004) and further extends these to uplift settings.

Table 6 and Table 7 offer an aggregated view on model performance. Interested readers find more detailed results in the online appendix, where we report the distribution of model performance across cross-validation iterations (see Figure A2.2). Related results indicate that RDT provides higher uplift and less variation compared to other uplift strategies. Overall, we observe tree-based RDT algorithms to outperform their competitors. We further secure this conclusion with a decile-wise analysis of alternative revenue uplift strategies using Revenue Qini curves. Corresponding results are available in the online appendix (see Figure A2.3) and confirm the RDT model to outperform other models on most deciles. We also observe that CRVTW performs inferior to other strategies, such as a simple revenue response model.

To summarize observed results and provide a holistic picture of model performance, we perform a multi-criteria evaluation. This evaluation includes the Qini curve values of important lower deciles due to fixed budget constraints, the Qini curve value as an average across deciles and a weighted Qini curve value based on the procedure proposed by Ling and Li (1998), which weights decile-wise results in light of their marketing importance. Formally, $Q_{wr} = \left(0.9 * Q_{1,r} + 0.8 * Q_{2,r} + \cdots + 0.1 * Q_{9,r}\right) / \sum_d Q_{d,r}$ with wr indicating weighted incremental (normalized) revenue and $d = 0, 1, \ldots, 9$ denotes an index of the targeted deciles. Contrary to Q_r, which takes random targeting into account but disregards decile-specific model performance, the Qini curve-related measures do not consider random targeting but focus on different aspects of incremental, cumulative model performance. Therefore, they are more suitable to guide operational decision-making. Table 8 reports the results of the multi-criteria evaluation and average ranks calculated across performance measures. For each measure, we highlight the three best models as before.

Table 8: Multi-criteria evaluation of revenue uplift models

Revenue Model	Top 10% Qini Curve	Top 30% Qini Curve	Averaged Qini Curve	Weighted Qini Curve	Average Rank
RDT (ERT)	**1,470**	**1,505**	**1,599**	**6,806**	1
CRVTW (RFR)	468	892	1,119	4,226	4
ITM (Ridge)	1,034	*1,431*	*1,424*	*6,313*	2

INDIRECT_TS (Ridge)	310	804	1,094	3,928	4
RESPONSE (Ridge)	*1,191*	1,191	1,323	5,563	3

Table 8 provides strong evidence in favor of the proposed RDT revenue uplift model. It performs best on all dimensions and achieves the first rank. Due to being ranked second in both the top 30% Qini curve, the averaged Qini curve, and the weighted Qini curve, ITM is the second-best strategy, followed by the response model-based ridge regression with three third-best ranks and a second-best rank on the top 10% Qini curve. Given substantial differences between the simple revenue response model and RDT as well as ITM in terms of three out of four assessment criteria, Table 8 makes a strong case for revenue uplift modeling and the cruciality of a causal approach to target marketing actions. Empirically, revenue response modeling does not prove to be a viable alternative. CRVTW and the indirect approach do not achieve a good rank on any performance dimension and share the last position in the multi-criteria evaluation.

6.3 Revenue Versus Conversion Uplift Models

We complete the analysis of revenue uplift transformation with a comparison of selected strategies to classical conversion uplift models in Table 9. In line with Table 8, we chose RDT and ITM, as the best revenue uplift strategy based on classification and regression models, respectively, for the comparison. For conversion uplift models, we draw on previous results of Kane et al. (2014) that identify ITM and LWUM as competitive modeling strategies for conversion uplift. These approaches require an underlying base learning algorithm for classification. We consider the algorithms that also facilitate the development of RDT uplift models (see Table A1.2 in the online appendix). As in the revenue uplift case, we report results only for the best conversion model, which we determine through comparing alternative classification algorithms and meta-parameters on the validation data. For completeness, the table also reports results for an ordinary response model. We mark the most effective model for each decile in italic font and highlight the overall three highest values across models and deciles in bold.

Table 9: Comparison of revenue and conversion uplift modeling

Strategy	Focus	Algorithm	Incremental Normalized Revenue Per Decile									
			Dec. 1	Dec. 2	Dec. 3	Dec. 4	Dec. 5	Dec. 6	Dec. 7	Dec. 8	Dec. 9	Dec. 10
RDT	Revenue	ERT	*1,470*	1,200	*1,505*	1,637	1,396	1,668	***1,907***	***1,892***	1,717	1,683
ITM	Revenue	Ridge	1,034	*1,618*	1,431	*1,638*	1,468	1,280	1,397	1,436	1,512	1,683
	Conversion	SVC	272	994	1,208	1,181	1,399	1,533	1,717	1,603	1,670	1,683
		KNN	412	748	913	986	1,212	1,490	1,489	1,843	*1,840*	1,683
LWUM	Conversion	RFC	1,369	1,278	1,332	1,386	*1,643*	*1,793*	**1,868**	1,829	1,817	1,683
		KNN	453	829	952	1,288	1,397	1,534	1,613	1,788	1,839	1,683

Response Modeling	Conver-sion	LogR	1,340	1,224	823	1,372	1,273	1,330	1,351	1,385	1,509	1,683
		RFC	1,456	965	1,354	1,330	1,393	1,576	1,726	1,724	1,665	1,683

The first finding of Table 9 is that the two best results with the highest incremental (normalized) revenue come from RDT, which confirms previous findings and the suitability of the RDT modeling strategy. For the data under study, we observe large campaigns targeting 70% of the customer base to give the highest normalized revenues. We attribute this pattern to the specific data used in this study. In general, marketers prefer targeting a small fraction of customers. Therefore, it is appealing to observe revenue uplift strategies to also provide the best results for each of the first four deciles. Overall, Table 9 confirms revenue uplift modeling strategies such as RDT to increase campaign revenue to a larger extent than traditional uplift models for conversion. From a marketing view, this finding might not come as a surprise. It stresses that customers differ in their spending. A targeting model should take this heterogeneity into account. However, we reiterate that much prior literature on machine learning-based targeting models does not consider uplift models at all and that the few uplift modeling studies predominantly use conversion modeling.

7 Analysis of RDT Uplift Strategy Against Causal Machine Learning

Previous analysis has identified RDT as the most suitable revenue uplift strategy for the available e-commerce data. We have also observed this strategy to perform best when implemented with an ERT base learning algorithm and obtained a set of optimal meta-parameters for the base learner. To set the performance of this specific configuration, our best revenue uplift model, into context, we compare it to several recently proposed causal machine learning algorithms. The causal learning algorithms include causal trees and their successor causal forests (Athey & Imbens, 2016; Athey et al., 2019), causal boosting (Powers et al., 2018), causal BART (Hill & Su, 2013; Hill, 2011), and an x-learner (Künzel et al., 2019), which we implement using a random forest regressor. These benchmarks facilitate estimating customer-level CATE for continuous target variables and thus revenue uplift modeling.

Software to apply the causal machine learning algorithms are publicly available. We detail the employed packages, our hardware, and how we specified the causal machine learning algorithms in the online appendix (see Subsection A3.1). Each of the benchmarks has shown strong performance in previous evaluations (see, e.g., Knaus et al., 2018, as well as the above studies, which have introduced the algorithms). However, we are not aware of a previous evaluation in an uplift modeling context. Marketing data sets are typically high-dimensional and include many observations. To shed light on the scalability of available software libraries for causal machine learning algorithms, subsequent analysis reports computation time alongside predictive performance. In addition to Qini scores and runtime measurements, we use our proposed incremental profit measure, as derived in (13), to assess RDT and causal machine learning benchmarks. Corresponding results clarify the degree to which revenue uplift models add to the bottom line and whether differences in the predictive performance of alternative models are managerially meaningful.

Given that the specific configuration of RDT emerges from Section 6, using the same data for the following comparison would give RDT an unfair advantage. Therefore, we use a fresh set of data for the comparison. The new data comes from three online shops that were not part of the original (cross-shop) data set and are new clients of Akanoo. Thus, the additional data was also captured in a different period and includes an extended set of covariates. Descriptive statistics and some more detailed information on the new data are available in the online appendix (see Table A3.1). Subsequently, we refer to the new data sets using the acronyms BAT for books and toys, FA for fashion A, and FB for fashion B. We apply the RDT uplift model as specified in Section 6 without further tuning to the new data sets. Considering structural and temporal differences between the cross-shop and the new data, the following evaluation may be considered an out-of-universe test of RDT.

Table 10 reports empirical results, which we obtain from a random subset of 100,000 observations from the BAT, FA and FB data sets. More specifically, we partition sampled observations into ten bins of 10,000 observations, stratify each bin into 70% training and 30% test data, and repeat the estimation of a model on the training data and assessment on the test data for each partition. In Table 10, we express Q_r as scaled (non-normalized) revenue. Although the subsampling ensures the number of observations to be the same for each shop, we discourage comparing Revenue Qini scores across data sets because the purchase volume and revenue uplift vary substantially across data sets, as shown in the online appendix (see Table A3.1). Table 10 also reports computation times per data set, which we measure in seconds. We use bold and italic font to highlight the best and second-best model per evaluation criterion and data set, respectively. All values represent averages across ten partitions.

Table 10: Performance and computation times of RDT ERT and causal machine learning algorithms

	Revenue Qini Coefficient			Computation Times		
RDT ERT / Causal Machine Learning Algorithm	BAT Data	FA Data	FB Data	BAT Data	FA Data	FB Data
RDT ERT	**1.20**	**0.12**	**0.39**	6.27	*3.51*	4.41
Causal Tree	0.15	0.02	0.21	**0.59**	**0.54**	**0.51**
Causal Forest	0.25	0.07	*0.32*	*3.50*	3.70	*3.21*
Causal Boosting	0.18	0.04	0.18	1,512.00	393.06	207.33
Causal BART	0.24	*0.10*	0.17	130.21	105.64	96.11
X-Learner RF	*0.28*	0.04	0.27	185.97	182.35	181.80
Mean	*0.38*	*0.07*	*0.26*	*306.42*	*114.80*	*82.23*

Before elaborating on our interpretation of Table 10, we note that a more detailed view on model performance is again available in the online appendix in the form of Revenue Qini curves (see Figure A3.1). Considering the aggregated (over deciles) Qini scores of Table 10, we find RDT to deliver the overall best revenue uplift models. It excels on the BAT data set where Qini scores are 4.29 times better than

those of the second-best model (x-learner random forest). For example, the observed Revenue Qini coefficient of 1.20 implies that the summed revenue difference between treatment and control group customers across deciles is larger than the number of test set customers. RDT also performs consistently better than causal machine learning benchmarks on FA and FB. Although performance gains are less substantial compared to BAT, improvements of 20.0% (FA) and 21.9% (FB) in terms of Q_r over the strongest benchmark are still sizeable and confirm the promising performance of RDT. In appraising these results, it is essential to recall that we do not tune meta-parameters of the causal machine learning algorithms. Their application mimics how we use RDT using pre-specified meta-parameter settings. A tuning of meta-parameters is likely to alter empirical results for all techniques in the comparison. Therefore, we expect tuning to mainly change the level of results but not the relative order of competing algorithms.

Comparing the causal machine learners to each other, Table 10 confirms that causal forests outperform their predecessor, the causal tree, across all data sets. Overall, the causal tree appears less suitable for uplift modeling as it typically produces the lowest Revenue Qini scores. However, it is important to note that the causal tree is by far the most efficient algorithm in the comparison. Estimating a model on a 10,000 observation sample requires only half a second on average. Overall, Table 10 indicates that causal forests might be the best causal machine learning algorithm for the focal data. Relative to other causal learners, its Qini scores are consistently high and the algorithm requires only a few seconds to estimate a model. The x-learner performs roughly as good as causal forests but is much slower. Contrary to recent literature that confirms the empirical effectiveness of causal boosting and causal BART (Dorie et al., 2019; Wendling et al., 2018), our results are less in favor of these methods. Especially causal boosting does not perform well in our comparison. Its Qini scores are consistently lower than those of the best causal machine learner and the algorithm needs by far the most time to estimate a model.

While Knaus et al. (2018) also raise concerns related to the scalability of some causal machine learning algorithms, we caution against over-emphasizing runtime results of Table 10. We find it crucial to distinguish between an algorithm's scalability and an algorithm's implementation. The sharing of codes that emerge from research via public repositories is a valuable development. Experiments like that of Table 10 would not have been possible had the authors of the causal machine learners not released their codes to the public. One cannot expect such codes to be fully optimized and production-ready. For example, despite many empirical successes of the gradient boosting machine, it took years before highly efficient implementations of that algorithm became available (Chen & Guestrin, 2016). Researchers and practitioners interested in experimenting with causal machine learning algorithms on large data sets can still benefit from the runtime comparison, which might offer some guidance on the choice of technique. To conclude the discussion of computation times, we note that the proposed RDT approach displays competitive results. Although being less efficient than causal trees and causal forests, the observed runtimes do not raise concern against RDT. Again, its efficiency depends on the underlying learning algorithm. We choose ERT because it gave the best empirical performance in Section 6. However, RDT can easily ac-

commodate other classifiers, including high-performance implementations of, e.g., random forest or gradient boosting. Therefore, we are confident that scalability is not a problem for RDT.

To give a clearer view of the business impact of using alternative revenue uplift models, we use our profit decomposition to measure the financial return of e-couponing campaigns and subsequently report campaign profit for uplift among RDT and causal machine learning algorithms. Contact costs can be neglected in our setting because digital coupons can be issued at zero costs, for example, by inserting a pop-up window in a browser session. However, we consider the incentive costs associated with the discount value. We assume that treated buyers have activated the e-coupon as part of the checkout process. As before, profit is not normalized as we assess the effectiveness of the marketing campaigns on independent online shops. This allows us to assess the bottom-line impact of an uplift model for each of the three online shops. We calculate campaign profit using (13) and report corresponding results in Table 11. To this end, Table 11 presents the (absolute) performance of RDT ERT and the second-best model per data set, which we choose based on averaged profit across deciles. Table 11 further displays the relative profit increase of RDT ERT compared to the respective runner-up. Results are again averages across data partitions. We highlight the highest profit values per data set and decile in bold font. Interested readers find an extended version of Table 11 in the online appendix (see Table A3.2), where we report incremental campaign profit for all causal machine learning algorithms.

Table 11: Incremental campaign profit of RDT ERT and competitive causal machine learning algorithms

Data	RDT ERT / Causal Machine Learning Algorithm	Incremental Campaign Profit Per Decile									
		Dec. 1	Dec. 2	Dec. 3	Dec. 4	Dec. 5	Dec. 6	Dec. 7	Dec. 8	Dec. 9	Dec. 10
BAT	RDT ERT	**5,668**	**8,109**	**9,115**	**10,202**	**11,275**	**12,266**	**13,262**	**14,144**	**14,896**	15,135
	X-Learner RF	3,100	5,130	6,271	7,375	8,630	9,615	10,831	11,577	12,956	15,135
	Profit increase of RDT ERT (%)	*82.8%*	*58.1%*	*45.4%*	*38.3%*	*30.6%*	*27.6%*	*22.4%*	*22.2%*	*15.0%*	*0.0%*
FA	RDT ERT	548	**938**	1,161	**1,443**	1,553	**1,739**	**1,985**	**2,143**	**2,301**	2,400
	Causal BART	**751**	924	**1,166**	1,387	**1,564**	1,728	1,808	2,038	2,140	2,400
	Profit increase of RDT ERT (%)	*-27.0%*	*1.5%*	*-0.4%*	*4.0%*	*-0.7%*	*0.6%*	*9.8%*	*5.2%*	*7.5%*	*0.0%*
FB	RDT ERT	1,223	2,452	**3,414**	**4,150**	**4,837**	**5,401**	**5,856**	**6,223**	**6,383**	6,683
	Causal Forest	**1,484**	**2,719**	3,184	3,896	4,352	4,900	5,471	5,976	6,304	6,683
	Profit increase of RDT ERT (%)	*-17.6%*	*-9.8%*	*7.2%*	*6.5%*	*11.1%*	*10.2%*	*7.0%*	*4.1%*	*1.3%*	*0.0%*

Table 11 showcases the incremental campaign profit for each targeted decile. To appreciate these values, consider the first value of RDT from the BAT data. There, using RDT would lead to a profit increase of 5,668€. From an operational viewpoint, a marketer would target the top ten percent of the highest scored

customers by the proposed revenue strategy, which equates to 300 customers. We arrive at this number by considering the sampled customer population per data set (i.e., 100,000 observations) and dividing them by the number of ten data partitions. We cut each partition into a 30% test fold and, given this example, target customers from the first decile. Thus, RDT would target 300 customers, each of them providing an average incremental profit of 18.89€. Recall that treatment costs are already recognized in our e-couponing settings.

Considering the example of BAT, the actual campaign provided an incremental client revenue, or revenue uplift, of 9.76€ per person (see Table A3.1). More specifically, the average revenue per shop visitor has been 14.24€ and 4.48€ from the treatment and control group, respectively. Regarding the incentive costs of a ten percent discount relative to a buyer's average purchase volume, we obtain an incremental revenue of treated customers of 12.82€. Consequently, the actual incremental campaign profit per person in the BAT campaign was 8.34€ (i.e., 12.82€ minus 4.48€). Compared to the targeting model used for this campaign, our analysis suggests that targeting e-coupons using RDT boosts profit per customer to 18.89€, which equates to a relative increase of 126.6%. The fact that the total incremental profit of 5,668€ (see Table 11) is an average across predictions on ten disjoint data partitions indicates that the result is robust.

We ascertain that incremental campaign profit differs across data sets. The profit of targeting the whole customer population is 15,135€, 6,683€, and 2,400€ for the BAT, FB, and FA data set, respectively. We argue that this is due to the lower revenue per person and revenue uplift on the FA data set and refer to a variation in customer spending behavior. The lower revenue per person from the treatment group in FA is an essential aspect regarding its positive relation with campaign profit for uplift. Contrary to the FA data, the BAT and FB data have a 2.8 and 1.9 higher revenue per treated customer, and a 2.7 and 1.9 higher initial revenue uplift, respectively. The FA data might also have a higher degree of structural complexities, which complicates the identification of behavioral patterns.

Previous results evidence RDT as a promising strategy. However, we also observe sizeable profit increases over the actual campaign from the causal machine learning algorithms. More specifically, we identify the x-learner random forest, causal BART and causal forest as competitive models for the BAT, FA and FB data sets, respectively. The x-learner achieves a 24% higher incremental profit per person than the actual campaign for the first decile of the BAT data. Across targeted deciles of the FA data set, there is a fierce competition between RDT and causal BART, which is reflected by the small difference of incremental profits for several deciles. Apart from this, causal BART outperforms RDT with a 27% higher incremental profit on the first decile. Results on FB show a similar tendency. The causal forest achieves a 17.6% and 9.8% better performance than RDT on the first and second deciles, respectively.

Overall, Table 11 re-emphasizes previous empirical results in favor of RDT, which provides the overall highest profit on most of the targeted deciles for each of the data sets. Averaging over deciles, RDT is 34.2%, 0.1%, and 2,0% better than the runner-up for BAT, FA and FB, respectively. Its substantial finan-

cial achievements are particularly prevalent on BAT, where its relative profit increase ranges between 15.0% (ninth decile) and 82.2% (first decile). Thus, RDT becomes even more important for campaign management as the most considerable relative difference of 82.2% relates to targeting the smallest fraction of customers.

8 Summary, Implications and Limitations

Much prior work used conversion uplift models for targeting marketing campaigns. We introduced new target variable transformations to enable revenue uplift modeling. Assuming a marketing analyst to aim at maximizing the profit of targeted marketing actions, the proposed revenue uplift models offer a more direct and natural way to pursue this goal. Unlike conversion models, they account for heterogeneity in customer spending and target customers to maximize incremental revenue. Several empirical experiments demonstrated the effectiveness of the new RDT approach. Across large amounts of real-world e-commerce data, it performed consistently better than alternative strategies for response, conversion uplift, and revenue uplift modeling in terms of business-oriented performance metrics. Comparisons to powerful causal machine learning algorithms further support this view. Examining the business impact of observed performance differences using a new profit decomposition, we found RDT to provide sizeable improvements in profit compared to causal machine learning in an e-couponing context.

The empirical results have several implications for academia and corporate practice. From an academic perspective, the proposed target variable transformations extend existing approaches to develop causal uplift models to continuous responses. Causal inference is relevant in many disciplines. While the paper concentrates on targeting decisions in marketing, the RDT approach could be used in other scientific domains to process continuous responses. Medical applications might be a good example. Examining the differential impact of a treatment (e.g., a new medication plan) on a continuous outcome variable (e.g., recovery time) is an exemplary application setting. Such settings are well studied in the literature and typically approached using causal models for treatment effect estimation. Our comparison to cutting-edge causal machine learning algorithms indicates that uplift transformations like RDT could extend the set of modeling tools in medical and other domains in a valuable way. A related point concerns the distribution of responses, which might introduce modeling challenges when transiting from discrete to continuous outcome variables. We have shown a possible approach to address zero-inflated response distributions using two-stage models. Our results in a marketing context do not evidence two-stage models to be useful. Whether modeling tasks in other domains lead to the same conclusion is a question for future research. In general, the flexibility to implement a causal model using any regression/classification algorithm or more specialized techniques such as hurdle models is a fundamental advantage of uplift transformations.

From a practitioner's point of view, new ways (i.e., RDT) to solve known problems (i.e., campaign planning) also have value. More importantly, the delineation between conversion and revenue uplift modeling is, to the best of our knowledge, originally proposed here. Marketing campaigns differ in their objective and may aim at lead generation, market growth, profit maximization, to name only a few. Uplift models are well-established in corporate practice. Our paper raises awareness for the point that different marketing objectives, such as revenue/profit maximization, are easier to accommodate in a model that predicts a continuous outcome variable. We suggest and empirically compare concrete options to implement this concept using the proposed revenue uplift transformation or causal machine learning. We consider the algorithmic flexibility of the former a major advantage for marketing practice since it builds upon existing technology. Considering the case of the RDT approach, any software package capable of running, e.g., a logistic regression can be employed to build a revenue uplift model. Causal machine learning is a viable alternative, and recent developments in the field, such as causal forests, enjoy much recognition. Our paper is one of the first to test corresponding techniques in a real-world marketing setting. While much more evidence in e-commerce and other marketing applications will be needed to have trust in their effectiveness in marketing, our results make the first step in this direction and offer guidance for research and development initiatives aiming at exploring causal machine learning in industry. Independent of a specific modeling method, practitioners may also benefit from the profit decomposition. We design it as a new measure that better captures business goals than existing performance metrics for uplift modeling. The decomposition supports different cost models and campaign types. Measuring a causal model's incremental profit impact, it bridges the gap between results from a data analytics model and financial implications in an easily interpretable manner.

Our study has several limitations that open ways for future research. First, the employed marketing data set does not originate from a proper randomized trial so that we cannot rule out the effect of confounders. For example, spillover effects from parallel campaigns may have affected the customer behavior that we observe in the data. However, we deem such risk as low because Akanoo only displays one marketing incentive per customer session for the periods over which the data was gathered. Consequently, for all recorded sessions in our data set, we can be sure that Akanoo has not enacted other campaigns in parallel. Still, as Akanoo acts in a consulting role to different online shops, there is a chance that the shops themselves have run further campaigns or commissioned further service providers. While many of the shops in our data set are relatively small and may not have the resources to carry out advanced campaigns themselves, we acknowledge this possible shortcoming. Furthermore, we consider a cross-sectional design. Real-world marketing campaigns are often executed over a longer time and possibly with some adjustments. The cross-sectional setup may thus appear unrealistic. We strongly support the use of longitudinal data for an evaluation of (revenue) uplift modeling in future research, which was less suitable here. We obtained data from a relatively short period of two months. This would leave us with roughly sixty observations to measure campaign uplift at a daily interval in a longitudinal setting. A panel setup was also not

suitable because many of the shoppers in our data are new customers or customers that were not re-identified (e.g., because of using different devices). Only few customers re-appear several times so that a panel design would have implied a substantial loss of data. In summary, while the cross-sectional design has limitations, we are confident that it represented the most suitable choice to leverage the large number of observations across different shops and product assortments, which are available in the focal data set. Future research revisiting the models considered here and testing their performance in experimental designs that represent ongoing marketing campaigns is highly beneficial to further improve confidence and trust in novel revenue uplift models and causal machine learning.

Acknowledgment

We are grateful to Fabian Gebert and colleagues from Akanoo for organizing access to the data for the empirical study. Valuable help with the empirical experiments from Tillmann Radmer is also acknowledged and much appreciated. We also thank four anonymous reviewers who provided several insightful comments to improve earlier versions of the paper. Finally, we very much appreciate the time and efforts of the editor, Prof. Robert Graham Dyson, in handling our paper.

References

Armstrong, J. S., & Patnaik, S. (2009). Using quasi-experimental data to develop empirical generalizations for persuasive advertising. *Journal of Advertising Research*, *49*(2), 170-175.

Ascarza, E. (2018). Retention futility: Targeting high risk customers might be ineffective. *Journal of Marketing Research*, *55*(1), 80-98.

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148-1178.

Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, *138*(1), 191-211.

Ballings, M., & Van den Poel, D. (2015). CRM in social media: Predicting increases in Facebook usage frequency. *European Journal of Operational Research*, *244*(1), 248-260.

Baumann, A., Haupt, J., Gebert, F., & Lessmann, S. (2018). The price of privacy: An evaluation of the economic value of collecting clickstream data. *Business & Information Systems Engineering*, *61*(4), 413-431.

Bodapati, A., & Gupta, S. (2004). A direct approach to predicting discretized response in target marketing. *Journal of Marketing Research*, *41*(1), 73-85.

Cai, T., Tian, L., Wong, P. H., & Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, *12*(2), 270-282.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In B. Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen & R. Rastogi (Eds.). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, ACM, pp. 785-794.

Chickering, D. M., & Heckerman, D. (2000). A decision theoretic approach to targeted advertising. In C. Boutilier & M. Goldszmidt (Eds.). *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI'00)*, Morgan Kaufmann, pp. 82-88.

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.

Diemert, E., Betlei, A., Renaudin, C., & Massih-Reza, A. (2018). A large scale benchmark for uplift modeling. *Proceedings of the AdKDD and TargetAd Workshop*, London, UK.

Dorie, V., Hill, J., Shalit, U., Scott, M., & Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, *34*(1), 43-68.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.

Goldfarb, A., & Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, *30*(3), 389-404.

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.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2015a). A decision support framework to implement optimal personalized marketing interventions. *Decision Support Systems*, *72*, 24-32.

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2015b). Uplift random forests. *Cybernetics and Systems*, *46*(3-4), 230-248.

Gutierrez, P., & Gérardy, J.-Y. (2017). Causal inference and uplift modelling: A review of the literature. *International Conference on Predictive Applications and APIs*, pp. 1-13.

Hahn, P. R., Murray, J. S., & Carvalho, C. (2019). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Preprint arXiv:1706.09523*.

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, *23*(2), 219-236.

Hansotia, B., & Rukstales, B. (2002a). Direct marketing for multichannel retailers: Issues, challenges and solutions. *Journal of Database Marketing & Customer Strategy Management*, *9*(3), 259-266.

Hansotia, B., & Rukstales, B. (2002b). Incremental value modeling. *Journal of Interactive Marketing*, *16*(3), 35-46.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Hill, J. L., & Su, Y.-S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, *7*(3), 1386-1420.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217-240.

Hofstetter, H., Dusseldorp, E., Zeileis, A., & Schuller, A. A. (2016). Modeling caries experience: Advantages of the use of the hurdle model. *Caries Research*, *50*(6), 517-526.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945-960.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443-470.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.

Jaroszewicz, S., & Rzepakowski, P. (2014). Uplift modeling with survival data. *ACM SIGKDD Workshop on Health Informatics (HI-KDD'14)*, New York, USA.

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

Knaus, M. C., Lechner, M., & Strittmatter, A. (2018). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *IZA Discussion Paper No. 12039*.

Kochura, Y., Stirenko, S., Alienin, O., Novotarskiy, M., & Gordienko, Y. (2017). Performance analysis of open source machine learning frameworks for various parameters in single-threaded and multi-threaded modes. In N. Shakhovska & V. Stepashko (Eds.). *Advances in Intelligent Systems and Computing II*, Springer, pp. 243-256.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156-4165.

Kuusisto, F., Costa, V. S., Nassif, H., Burnside, E., Page, D., & Shavlik, J. (2014). Support vector machines for differential prediction. In T. Calders, F. Esposito, E. Hüllermeier & R. Meo (Eds.). *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD'14)*, Part II, Springer, pp. 50-65.

Lai, Y.-T., Wang, K., Ling, D., Shi, H., & Zhang, J. (2006). Direct marketing when there are voluntary buyers. *Proceedings of the 6th International Conference on Data Mining (ICDM'06)*, IEEE Computer Society, pp. 922-927.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1-14.

Lechner, M. (2019). Modified causal forests for estimating heterogeneous causal effects. *IZA Discussion Paper No. 12040*.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124-136.

Lessmann, S., Haupt, J., Coussement, K., & De Bock, K. W. (2019). Targeting customers for profit: An ensemble learning framework to support marketing decision-making. *Information Sciences*. https://doi.org/10.1016/j.ins.2019.05.027.

Ling, C. X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. In R. Agrawal, P. E. Stolorz & G. Piatetsky-Shapiro (Eds.). *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98)*, AAAI Press, pp. 73-79.

Lo, V. S. Y. (2002). The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter, 4*(2), 78-86.

Lo, V. S. Y., & 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.

Magliozzi, T. L., & Berger, P. D. (1993). List segmentation strategies in direct marketing. *Omega, 21*(1), 61-72.

Martens, D., & Provost, F. (2011). Pseudo-social network targeting from consumer transaction data. *NYU Working Paper CeDER-11-05*. https://ssrn.com/abstract=1934670.

Nassif, H., Kuusisto, F., Burnside, E. S., & Shavlik, J. W. (2013). Uplift modeling with ROC: An SRL case study. In G. Zaverucha, V. S. Costa & A. Paes (Eds.). *Late Breaking Papers of the 23rd International Conference on Inductive Logic Programming (ILP'13)*, pp. 40-45.

NCH Marketing Services (2018). Year-end coupon facts at a glance. https://www.nchmarketing.com/2018-year-end-coupon-facts-at-a-glance.aspx. Accessed 14 August 2019.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204-211.

Oprescu, M., Syrgkanis, V., & Wu, Z. S. (2019). Orthogonal random forest for causal inference. *Preprint arXiv:1806.03467*.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., & Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine, 37*(11), 1767-1787.

Radcliffe, N. J. (2007). Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal, 1*, 14-21.

Radcliffe, N. J., & Surry, P. D. (1999). Differential response analysis: Modeling true responses by isolating the effect of a single action. *Credit Scoring and Credit Control IV*, Edinburgh, Scotland.

Radcliffe, N. J., & Surry, P. D. (2011). Real-world uplift modelling with significance-based uplift trees. *Portrait Technical Report, TR-2011-1*.

Reimers, I., & Xie, C. (2019). Do coupons expand or cannibalize revenue? Evidence from an e-market. *Management Science, 65*(1), 286-300.

Rudaś, K., & Jaroszewicz, S. (2018). Linear regression for uplift modeling. *Data Mining and Knowledge Discovery, 32*(5), 1275-1305.

Rzepakowski, P., & Jaroszewicz, S. (2012a). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems, 32*(2), 303-327.

Rzepakowski, P., & Jaroszewicz, S. (2012b). Uplift modeling in direct marketing. *Journal of Telecommunications and Information Technology, 2*, 43-50.

Schröder, N., & Hruschka, H. (2017). Comparing alternatives to account for unobserved heterogeneity in direct marketing models. *Decision Support Systems, 103*, 24-33.

Shaar, A., Abdessalem, T., & Segard, O. (2016). Pessimistic uplift modeling. *22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'16)*, San Francisco, USA.

Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*, Sydney, Australia.

Sołtys, M., Jaroszewicz, S., & Rzepakowski, P. (2015). Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery, 29*(6), 1531-1559.

Su, X., Kang, J., Fan, J., Levine, R. A., & Yan, X. (2012). Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research, 13*, 2955-2994.

Taddy, M., Gardner, M., Chen, L., & Draper, D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics, 34*(4), 661-672.

Tian, L., Alizadeh, A. A., Gentles, A. J., & Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, *109*(508), 1517-1532.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211-229.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., & Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, *37*(23), 3309-3324.

Yamane, I., Yger, F., Atif, J., & Sugiyama, M. (2018). Uplift modeling from separate labels. *Advances in Neural Information Processing Systems 31 (NIPS'18)*, Montréal, Canada, pp. 9927-9937.

Yao, X., Crook, J., & Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research*, *263*(2), 679-689.