Journal of Business Economics
https://doi.org/10.1007/s11573-021-01068-3

ORIGINAL PAPER

Profit uplift modeling for direct marketing campaigns: approaches and applications for online shops

Daniel Baier[1] · **Björn Stöcker**[2]

Accepted: 20 October 2021

Abstract

In order to select "best" customers for a direct marketing campaign, response models are widespread: a sample of customers receives an ad, a catalog, a sample pack, or a discount offer on a test basis. Then, their responses (e.g., website visits, conversions, or revenues) are used to build a predictive model. Finally, this model is applied to all customers in order to select "best" ones for the campaign. However, up to now, only models that reflect website visits, conversions, or revenues have been proposed. In this paper, we discuss the shortcomings of these traditional approaches and propose profit uplift modeling appoaches based on one-stage ordinary regression and random forests as well as two-stage Heckman sample selection and zero-inflated negative binomial regression for parameter estimation. The new approaches demonstrate superiority to the traditional ones when applied to real-world datasets. One dataset reflects recent discount offers of a large online fashion retailer. The other is the well-known Hillstrom dataset that describes two Email campaigns.

Keywords Uplift modeling · Heckman sample selection model · Zero-inflated negative binomial regression · Random forests · Online shops

JEL Classification C01 · C53 · M31 · M37

✉ Daniel Baier
daniel.baier@uni-bayreuth.de

Björn Stöcker
bjoern.stoecker@baur.de

[1] Chair of Marketing and Innovation, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany

[2] Head of CRM, BAUR Versand, Bahnhofstraße 10, 96224 Burgkunstadt, Germany

Published online: 19 November 2021

1 Introduction

Before launching a direct marketing campaign, often, a sample of customers is testwise contacted. Their desired responses (e.g., website visits, conversions, revenues) as well as their past information and buying behavior is used to build a response model. Then, this model is applied to all customers and to select likely responders for the campaign. However, this traditional approach has two shortcomings:

- First, response models focus on likely responders, possibly independent of the contact. This could be a waste of money, e.g. in case of unnecessarily distributed sample packs or discount offers.
- Second, up to now, response models only predict binary outcomes (website visits, conversions) or revenue outcomes, not the more informative profit outcomes at the individual level.

Both shortcomings restrict the usefulness of the traditional approaches for maximizing profit. In this paper, we propose new profit uplift modeling approaches as alternatives: first, uplifts focus—in contrast to responses—on the incremental response due to a treatment using control groups of customers. Second, profit is more difficult to model since this outcome is only observable in a few cases but more closely related to the main objective than website visit, purchase, or revenue. The proposed new approaches in this paper extend findings from the field of binary and revenue uplift modeling (e.g., Radcliffe and Surry 1999, 2011; Kane et al. 2014; Rudaś and Jaroszewicz 2018; Gubela et al. 2020) and from the field of two-stage estimation via sample selection (see, e.g., Heckman 1979) and zero-inflated regression (see, e.g., Lambert 1992; Ridout et al. 2001) as well as one-stage parameter estimation via ordinary regression and random forest. We show that the new approaches are well suited to select "best" customers as targets for direct marketing campaigns and improve profit.

The paper is organized as follows: In Sect. 2, we discuss the traditional approaches and their shortcomings and, in Sect. 3, the new approaches. In Sect. 4, the superiority of the new over the traditional approaches is demonstrated using a new dataset from a major German online retailer (with a sample of $n = 155{,}388$ customers). In Sect. 5, the well-known Hillstrom direct marketing campaign dataset (with a sample of $n = 64{,}000$ customers) is used for the same purpose. The paper closes with conclusions and outlook.

2 Background and related work

Testing and predictive modeling are assumed to be the analytical cornerstones of today's direct marketing (Blattberg et al. 2008). The modeling process usually consists of the following five steps: (1) Define the managerial problem in terms of

a campaign and its intended effects, (2) translate this description to a predictive model with treatment, responses, and potential predictors, (3) sample customers for collecting responses, (4) calibrate and validate the predictive model, (5) apply the model to all customers and select "best" customers according to the predictive model. Typical managerial problems are selecting targets for an acquisition campaign at hand, deciding on customers to receive a catalog or inlay, or identifying promising customers for a customer tier program. Outcomes are the response to the treatment, predictors are customer characteristics (age, gender, income if available) and variables that describe past information and buying behavior in the customer database (see, e.g., Blattberg et al. 2008 for an overview).

Let Y_i be the binary ($Y_i \in \{0, 1\}$) or continuous ($Y_i \in \mathbb{R}$) outcome for customer i in the customer sample, $\mathbf{x}_i (\mathbf{x}_i \in \mathbb{R}^m)$ customer i's values for the m predictors, and τ_i the indicator whether customer i received the treatment (=1) or not (=0). Then, the main goal for a traditional response modeling approach would be to predict the following purchase likelihood (in case of binary outcomes) or scalar values (in case of continuous outcomes) as customer scores for selecting targets,

$$\text{Response}_i(\mathbf{x}_i) = E(Y_i|\mathbf{x}_i). \tag{1}$$

The data of the treated customers ($\tau_i = 1$) are used to calibrate the response model (using, e.g., logistic regression or simple regression depending on the scale of the outcome). Then, the whole customer database is used for prediction. The customers with the highest (response) scores are targets for the campaign. However, this response modeling approach has one major shortcoming: it favors customers who respond most likely, but it does not take into account that some of them would also respond if not treated. When the treatment is a discount, a voucher, a catalog, an inlay, or one has to deal with postage, this could result in a waste of money for the company.

Therefore, recently, uplift models have been proposed: responses are again collected from a sample of treated customers (the treatment group), but also from a sample of not treated customers (the control group). An uplift model now predicts the difference in the response of a customer if treated ($\tau_i = 1$) and if not treated ($\tau_i = 0$), the so-called uplift score,

$$\text{Uplift}_i(\mathbf{x}_i) = E(Y_i|\mathbf{x}_i, \tau_i = 1) - E(Y_i|\mathbf{x}_i, \tau_i = 0). \tag{2}$$

Terms like differential response (e.g., Radcliffe and Surry 1999), true lift (Lo 2002), or uplift (Radcliffe and Surry 2011) are used for the same idea. Formula (2) allows to estimate the effect of the treatment and enables the company to select customers where the treatment has the highest impact.

However, when trying to estimate the model parameters, a problem arises from the fact that per customer, only one of these two responses is observable: a customer is part of the treatment group ($\tau_i = 1$) or part of the control group ($\tau_i = 0$), not of both groups. Consequently, an uplift model cannot be estimated directly when using formula (2). Instead, one straightforward idea is to develop two separate models (the so-called two model approach): a first model is derived similar to formula (1), based on the treatment group. This model predicts the outcome

if treated in terms of $\mathbf{x}_i$ for all customers. A second model is derived using the control group. This model predicts the outcome if not treated in terms of $\mathbf{x}_i$ for all customers (see Radcliffe and Surry 1999), the difference between the predictions of the two models is the uplift.

An alternative solution is the so-called interaction model proposed by Lo (2002): an interaction (response) model uses the treatment (τ_i) and interactions between the predictors and the treatment as additional predictors. The interaction model can be calibrated on the treatment and the control group simultaneously. Then, for all customers, predictions for all customers are derived via formula (2) by setting the treatment for all customers to 1 in the first term and 0 in the second term.

Over the years, a remarkably high number of uplift modeling approaches, including algorithms to estimate their parameters, have been proposed. Table 1 gives an overview. As one can easily see, most of them aim at predicting uplifts for binary outcomes, e.g., indicators for a visit, conversion, or purchase. Here, logistic regression or decision trees can be applied to estimate model parameters. However, more recently, also revenue uplift modeling approaches have become popular (Gubela et al. 2020; Rudaś and Jaroszewicz 2018). The main idea behind this new development is that the revenue uplift more closely relates to economic objectives than a website visit or purchase uplift. However, in the next section, we will see that even revenue uplift modeling approaches sort customers suboptimally.

Another interesting aspect in Table 1 is that most recent uplift modeling approaches rely on transformed outcomes for parameter estimation. This transformation was introduced for binary outcomes in a seminal paper (Lai 2006) and later extended to continuous outcomes (Gubela et al. 2020; Rudaś and Jaroszewicz 2018). The main idea behind it is to transfer as much information as possible from the observed responses in the two groups into the dependent variable and so being able to directly estimate the uplift model parameters (a so-called direct model). So, e.g., Rudaś and Jaroszewicz (2018)—following the proposal of Lai (2006) for binary outcomes—proposed to estimate their revenue uplift model,

$$\text{Uplift}_i(\mathbf{x}_i) = E(Z_i|\mathbf{x}_i), \tag{3}$$

directly using transformed revenue outcomes,

$$Z_i = \begin{cases} +\frac{1}{q^T} Y_i \;\; if \;\; \tau_i = 1 \;\wedge\; Y_i > 0 \\ 0 \qquad if \;\; Y_i = 0 \\ -\frac{1}{q^C} Y_i \;\; if \;\; \tau_i = 0 \;\wedge\; Y_i > 0 \end{cases}, \tag{4}$$

for parameter estimation. q^T and q^C are the fractions of the treatment group and the control group in the customer sample. Rudaś and Jaroszewicz (2018) discuss in their paper that this weighting facilitates unbiased estimation of the model parameters when relying on linear models. The main idea behind the positive weighting of the observed revenues in the treatment sample and the negative weighting of the observed revenues in the control sample is that so the best possible information is forwarded to parameter estimation. It is assumed that the purchasers in the treatment group generate probably a (low to high) positive revenue uplift. Likewise, it is

Table 1 Uplift modeling approaches

Approach	Outcome	Algorithm	Reference
Differential response analysis: modeling true response	Binary	DT	Radcliffe and Surry (1999)
Incremental value modeling	Binary	DT	Hansotia and Rukstales (2002)
The true lift model	Binary	Logistic regression	Lo (2002)
Influential marketing: a new direct marketing strategy	Binary (transformed)	Association rules, DT, Logistic regression	Lai (2006)
Using control groups to target on predicted lift	Continuous	CART	Radcliffe (2007)
Uplift modelling with significance-based uplift trees	Continuous	CART	Radcliffe and Surry (2011)
DTs for uplift modeling with multiple treatments	Multiple binary	DT	Rzepakowski and Jaroszewicz (2012)
Support vector machines for uplift modeling	Binary (transformed)	SVM	Zaniewicz and Jaroszewicz (2013)
Uplift random forests	Binary (transformed)	Causal conditional inference tree/RF	Guelman et al. (2015)
Mining for truly responsive customers	Binary (transformed)	Logistic regression	Kane et al. (2014)
Ensemble methods for uplift modeling	Binary (transformed)	Ensemble methods	Sołtys et al. (2015)
L_p-support vector machines for uplift modeling	Binary (transformed)	SVM	Zaniewicz and Jaroszewicz (2017)
Revenue uplift modeling	Continuous (transformed)	Linear regression	Rudaś and Jaroszewicz (2018)
Revenue uplift modeling	Continuous (transformed)	Lasso, Ridge, and Theil-Sen regression, MLP, RF	Gubela et al. (2020)
Profit uplift modeling	Continuous (transformed)	OLS, Heckman sample sel., Zero-inflated NB, RF	This paper

Only new approaches are reflected, *CART* classification and regression tree, *DT* decision tree, *MLP* multilayer perceptron, *RF* random forest, *sel.* selection, *SVM* support vector machine

assumed that the purchasers in the control group generate probably a (low to high) negative revenue uplift.

Another major problem with uplift modeling approaches is to validate their predictions at the customer level as for these predictions—as mentioned above—no observations exist. The widespread solution for this problem is to develop so-called Qini curves and calculate the so-called Qini coefficient Q (Radcliffe 2007; Radcliffe and Surry 2011): the customers are sorted according to a descending (uplift) score and partitioned into deciles (or other partitions of the customers) with similar scores. Then, within the deciles, customer responses from the treatment group are averaged as well as customer responses from the control group. The difference of these two means is assumed to be the "observed" uplift in this decile. Figure 1 shows the typical results for such a validation applied to (uplift) scores from a sample dataset.

In both diagrams, the customers are sorted according to descending uplift predictions from left to right and grouped in deciles. In the right diagram, the calculated average ("observed") uplift per decile is given, as discussed above. In the left diagram, from decile to decile, the average cumulative uplift is plotted, which means that for the first decile, the values in the left and right diagram are identical, but from then, aggregated values up to the current decile are given in the left diagram. The last value (here: 0.045) of this so-called Qini curve reflects the uplift across all ten deciles (all customers of the treatment and the control group) which is identical for all scorings based on the data. For comparisons, also the Qini curve for a random sorting (the random uplift model) is plotted in the left diagram. Its incremental uplift curve connects the zero point with the average uplift across all customers, the last value (0.045). The quality of an uplift model is judged by its ability to sort customer deciles according to decreasing "observed" uplifts (in the right diagram) but—similar to ROC curves—also by calculating the area between the Qini curve and the line for the random model (in the left diagram). Here the length of the x-axis

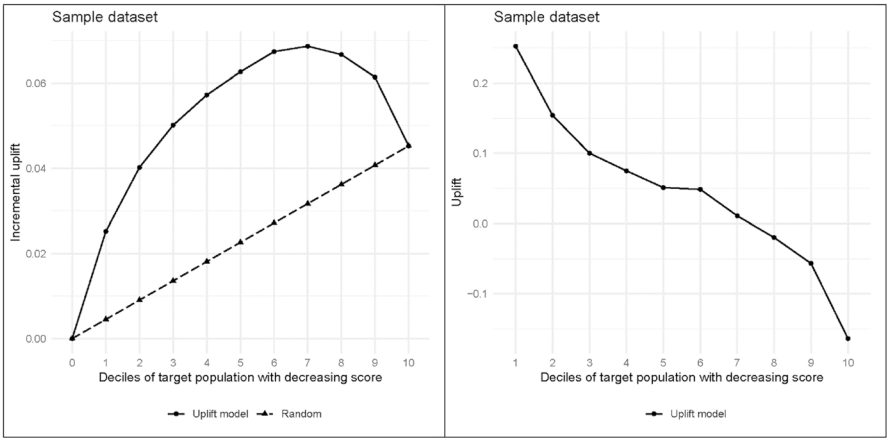

Fig. 1 Qini curve (left) and mean uplifts (right) for a sample dataset. The area between the Qini curve of an uplift model and a random model is the Qini coefficient Q (here: Q = 0.0296), which can be used for model selection

is assumed to be 1. In Fig. 1, this value—the so-called Qini coefficient Q—is 0.0296 and could serve for comparisons with other uplift models (other sortings of customers according to their predicted scores). The random model has Q = 0, the maximum is data-dependent. It should be noted that these two diagrams can be generated for uplift models with binary response outcomes but also for uplift models with continuous outcomes (as in revenue uplift modeling or our new profit uplift modeling approach discussed in the following section).

3 Profit uplift modeling approaches for online shops

3.1 Potential usefulness of profit uplift modeling approaches

As already discussed, most uplift modeling approaches reflect binary outcomes. Only recently, continuous outcomes have received more interest, e.g., in the papers by Rudaś and Jaroszewicz (2018) as well as Gubela et al. (2020). This is surprising since, from the beginning of the development of uplift modeling approaches, also datasets with continuous outcomes have been made available. So, e.g., the famous Hillstrom dataset (Radcliffe 2008)—which is often seen as the standard dataset in uplift modeling and has been used in many papers when uplift models were introduced or compared—contains as binary outcomes the website visits (= 1: yes, = 0: no) and the purchase information (= 1: yes, = 0: no) but also the revenue generated by this purchase (spend in $). However, maybe since the share of purchasers in this dataset (0.9% of the customers) and the revenue uplift were very low, and, additionally, the revenues concentrate on very few purchasers, this dataset did not stimulate the scientific community to develop continuous outcome uplift models. Even in the newer and methodologically advanced paper Rudaś and Jaroszewicz (2018), this dataset is only used as a basis for a simulation at the end of the paper. The main methodological progress in revenue uplift modeling in their paper was demonstrated by using synthetic data. However, recently, Gubela et al. (2020) have demonstrated in their paper with large real-world datasets (nearly 3 million sessions from visits at 25 European online shops) that revenue uplift modeling approaches provide further insights.

This superiority of a continuous outcome uplift modeling approach can also be seen when reflecting the assumed behavior of a small sample of customers as shown in Table 2. Here, for 12 customers, their potential outcomes (purchases, revenues, and profits) are given in case of a direct marketing campaign with a discount offer of d = 20% and a profit margin of m = 30%. Profits are calculated for customers in the treatment group as 10% (= m–d) and for the control group as 30% (= m) of the revenue.

One can easily see that the 12 customers reflect a typical behavior: they show—on average—a purchase outcome uplift (8%) when offered a discount, they generate a higher revenue when a discount is offered (+ 69 €), but it is not useful to offer the discount to all customers since the profit uplift—on average—is negative (– 9 €). Only five customers (1, 2, 3, 4, and 5) show a profit uplift, which means that only these five customers should be offered the discount. The customer sorting according

Table 2 Sample of customers with potential purchase, revenue, profit if treated (offered a discount of 20% at a margin of 30%) and if not treated (no discount offer)

Cus-tomer	Purchase			Revenue			Profit		
	If treated	If not tr.	Uplift	If treated	If not tr.	Uplift	If treated	If not tr.	Uplift
1	1	0	1	160 €	0 €	160 €	16 €	0 €	16 €
2	1	1	0	300 €	60 €	240 €	30 €	18 €	12 €
3	1	0	1	40 €	0 €	40 €	4 €	0 €	4 €
4	1	0	1	30 €	0 €	30 €	3 €	0 €	3 €
5	1	0	1	20 €	0 €	20 €	2 €	0 €	2 €
6	1	1	0	70 €	40 €	30 €	7 €	12 €	–5 €
7	0	1	–1	0 €	20 €	–20 €	0 €	6 €	–6 €
8	0	1	–1	0 €	40 €	–40 €	0 €	12 €	–12 €
9	0	1	–1	0 €	60 €	–60 €	0 €	18 €	–18 €
10	1	1	0	400 €	200 €	200 €	40 €	60 €	–20 €
11	1	1	0	500 €	250 €	250 €	50 €	75 €	–25 €
12	1	1	0	270 €	290 €	–20 €	27 €	87 €	–60 €
Mean	75%	67%	8%	149 €	80 €	69 €	15 €	24 €	–9 €

Interpretation: customer 1 generates a revenue of 160 € if treated and 0 € if not treated. The treatment generates a revenue uplift of 160 € (=160–0 €) and a profit uplift of 16 € (=160 € * (30–20%) – 0 € * 30%). Please note that this perfect information is not observable in practice since a customer can only be part of the treatment group (if treated) or the control group (if not treated)

to the purchase outcome and the revenue outcome differs: the three customers with the highest revenue uplift show a purchase uplift of 0. However, as also can be seen in Table 2, both sortings considerably differ from the sorting according to the profit uplift: if the customers were targeted according to their revenue uplift, customers with a positive profit uplift but also with a negative profit uplift would receive a discount offer. It should be noted that this difference in sorting heavily relies on the ability of discounts to generate additional revenues but also on the fact that in online shops, high discounts are widespread but would lead to losses if granted to all customers. Moreover, it should be mentioned that Table 2 reflects an ideal situation in-so-far that from each customer, two observations are available—the outcomes with and without treatment—which in reality is not possible.

3.2 Profit uplift modeling approaches in detail

After demonstrating the potential usefulness of profit uplift modeling approaches, now, they are discussed in detail. The main idea is to use formulae (2) (as a two model or an interaction model approach) or (3) and (4) (as a direct approach) for modeling continuous outcomes but to replace the observed revenue by derived profits and the revenue uplift predictions by profit uplift predictions. We follow Blattberg et al. (2008) as in Sect. 2 and discuss the five steps of the predictive modeling process now in detail:

1. Define the managerial problem: online shops have a huge variety of potential offerings that could motivate their customers to purchase and/or to spend more. So, e.g., discount offerings are widespread. Gubela et al. (2020) mention in their e-commerce datasets discounts of 10% to stimulate a purchase during a website visit. Depending on the branch or the product group, the discounts offered to customers per mail, inlays, or newsletter could even be higher. So, e.g., in fashion online shops, discount offerings of 20% are quite common. Moreover, in furniture online shops, even discounts up to 50% and more are frequent. Alternative purchase stimuli are, e.g., vouchers, attached gifts, bonus programs, tombolas, and raffles. However, since profit margins for online shops are typically low (e.g., between 5 and 15%, sometimes up to 40%), these discounts, vouchers, and gifts are double-edged swords: they could generate more revenue but at the same time reduce profit at the customer level dramatically. Consequently, a scoring system is needed that relates offerings, (past) information, and shopping behavior to profit uplift.
2. Translate the managerial problem to a predictive model: as Blattberg et al. (2008, p.250) discuss in their overview, widespread and useful predictors for binary and continuous response outcomes (e.g., visit, purchase, revenue, profit) in database marketing response are

    - customer characteristics (socio-demographics, lifestyle, psychographics),
    - previous behavior (purchases and responses to previous marketing efforts, typically described using recency, frequency, and monetary value (RFM) variables), and
    - previous marketing (efforts targeted at the customer, including catalogs, Emails, discounts).

Similar variables have been used in the uplift modeling literature. So, e.g., the Hillstrom dataset embodies as customer characteristics the living environment (rural, suburban, urban), as previous behavior recency (time since last purchase), history (money spend in the last year), mens and womens (indicators for product categories bought in the last year), and newbie (indicates a first purchase in the last year), and as previous marketing the used shopping channels. Additionally, nowadays, for online shops, variables that describe the online information behavior are tracked and used, e.g., the duration and recency of shop visits or the number of page views (see, e.g., Gubela et al. 2020). These predictors should also be used in our profit uplift modeling approaches, if available.

As the outcome of our predictive model—in contrast to the already published revenue uplift modeling approaches—we define for the first time in literature the profit outcome (in case of a two model or interaction modeling approach as in formula (2) or the profit uplift outcome (in case of a direct modeling approach as in formulae 3 and 4) at the customer level. The calculation of the profit outcomes from the revenue outcomes depends on the treatment offered to the customer (e.g. discount, bonus, vouchers, attached gifts, bonus programs, tombolas, or raffles) and the margin. In the control group, it only depends on the margin. Especially the calculation of the latter is a critical point since, in online shopping, the clear allocation of item-related costs to a purchase

is difficult as besides the supply costs also return, damage, loss, and other aspects would have to be taken into account. Nevertheless, we follow the argumentation by Blattberg et al. (2008) and Gubela et al. (2020) in their determination of cut-off points for customers where average values across product categories or shops were used (and are available in online shops).

3. Sample customers for collecting responses: as usual in uplift modeling, the dependency of outside-effects can be reduced if the treatment and the control groups are random samples out of the customer base, ideally balanced with respect to selected predictors (e.g., recency, frequency, monetary value). Moreover, since responses to direct marketing campaigns typically are rather low (e.g., 0.9% purchasers in the Hillstrom dataset), the drawing of large samples is necessary to develop stable models.
4. Calibrate and validate the predictive model: the small percentage of purchasers in the treatment group and the control group reduces the number of applicable models and parameter estimation algorithms considerably. In fact, the revenue- or profit-generating response can be seen—simplified—as a two-stage process that should be modeled: in the first stage (few) customers decide to purchase items (being treated or not): we have a traditional response model with a binary outcome. In the second stage, only the profit-generating behavior of the purchasers is modeled. The predictors in both model stages could be the same or different ones. If we use formulae (3) and (4) for this purpose (the direct modeling approach), the "observed" negative and positive profit outcomes have to be transformed to "normal shape" by a Box-Cox-transformation. If we use formula (2) for this purpose (the two model or an interaction model approach), the non-negative profit outcomes have to be modeled directly. Here, besides the case of "normal shaped" (with or without Box-Cox-Transformation), profit data could be interpreted as being count data in "negative binomial shape" (after transformation to Millicent and rounding). For both two-stage modeling cases, well-known parameter estimation procedures exist:

    - In the first case with profit outcomes of the purchasers in "normal shape", Heckman's sample selection model (Heckman 1979)—also called Tobit-2 model—can be applied (Toomet and Henningsen 2008). This model can be described by two equations:

$$\begin{aligned} Y_i^{S*} &= \beta^{S*\prime} X_i + \varepsilon_i^{S*} \\ Y_i^{O*} &= \beta^{o*\prime} X_i + \varepsilon_i^{o*} \end{aligned} \quad with \quad \begin{pmatrix} \varepsilon_i^{S*} \\ \varepsilon_i^{o*} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix} \right), \tag{5}$$

      where Y_i^{S*} represents the selection tendency (here: purchasing tendency) for individual i and Y_i^{O*} the latent (profit) outcome. We observe the binary outcome Y_i^S and—for the selected cases (the purchasers)—the continuous outcome Y_i^O as follows,

$$Y_i^S = \begin{cases} 0 & if\ Y_i^{S*} < 0 \\ 1 & \text{otherwise,} \end{cases}$$
$$Y_i^O = \begin{cases} 0 & if\ Y_i^S = 0 \\ Y_i^{O*} & \text{otherwise.} \end{cases} \tag{6}$$

The conditional regression estimation proposed by Heckman (1979) applies the so-called Heckman correction (inverse of Mill's ratio) to eliminate the sample selection effect.

- For the second case, with profit outcomes converted to count data and in "negative binomial shape", count models can be applied. Here, again, the (few) purchasers induce many zeros in the count outcomes, which can be reflected again by a two-stage model, using the so-called hurdle models by Mullahy (1986), the so-called zero-inflated Poisson regression model by Lambert (1992) or—as used in our paper—the flexible so-called zero-inflated negative binomial regression model by Ridout et al. (2001). In all cases, Poisson or negative binomial regression models are used as second stage models and again combined with a selection model for non-negative counts.

As a third and fourth alternative to this two-stage modeling approaches, a simple (one-stage) regression model (OLS) and a (one-stage) random forest regression model (e.g., CART by Breiman 2001) can be applied as direct models according to formula (3) and (4). Random forests are known in machine learning for being very robust against unbalanced data with few purchasers and consequently few positive profit outcomes.

As usual in predictive modeling, a partitioning of the data into train and (holdout) test data is needed to control the predictive validity (here: with respect to profit Qini coefficients). Also, a partitioning of the train data in calibration and validation data to tune hyperparameters of the algorithms is widespread. According to many authors in the uplift modeling literature (e.g., Devriendt et al. 2018), here, especially the preprocessing of the predictors (using, e.g., variable selection or principal components analysis) and a selection of not too much (say 5 to 15 according to Devriendt et al. 2018) predictors are important for calibration and validation.

5. Apply the model to all customers and select "best" customers: the calibrated, validated, and tested profit uplift model is used to score the customers and to select profitable ones for the direct marketing campaign. Since the predicted score, the profit uplift per customer is informative, a concentration on customers with scores larger than 0 could be a standard strategy. The mean uplift curve as in Fig. 1 is well suited for this selection process.

In the following two sections, the discussed new profit uplift modeling approaches (based on OLS, Heckman's sample selection model, random forest, and zero-inflated negative binomial regression) are applied to demonstrate their usefulness. The results are compared to revenue response and revenue uplift as well as profit response modeling approaches.

4 Application to direct marketing campaigns of a German online shop

4.1 Company, campaigns, descriptive uplift statistics, and preprocessing of the data

The data for the first application was provided by one of the pioneers in the mail order business in Germany, the BAUR group, since 1997 a major member of the OTTO Group. The website www.baur.de is one of the ten largest online shops in Germany. Clear customer and service orientation, high-quality standards, and a constantly up-to-date range of items in the fashion, shoe and furniture product range are assumed to be the key success factors (see Baier et al. 2019). The company mainly focuses on customers aged 40–55 and offers well-known brands as well as exclusive fashion branded by BAUR. The online shop—which represents about 90% of the business volume—is supported by catalogs that focus on seasonal or special fashion topics. Like many other online shops, scoring systems are used to select customers for direct marketing campaigns. The development of an effective scoring system is an ongoing central challenge for this company. Therefore, on a regular basis, tests are performed: random samples of customers are divided into treatment and control groups according to balanced designs. Then, the customers of the treatment groups are offered discounts (e.g., by mail), and the purchases of the customers of both groups are tracked in the follow-up (two) weeks and used to refine the scoring system.

The provided data reflects two recent tests. Altogether 155,388 selected customers were divided up into treatment and control groups. The customers in the treatment groups received a 20% discount offer for the next order; the purchases of both groups were tracked in the follow-up weeks. Table 3 reflects the descriptive uplift statistics of these two tests. As one can easily see, the sampling resulted in equally large treatment and control groups. It should be mentioned that for both tests, the samples were selected randomly out of the company's customer base (without overlap) and that the dividing up of the two samples into treatment and control groups was performed in a balanced manner with respect to pre-defined variables that describe the customers' past information and buying behavior, e.g., their purchase volume in the last two years, their usage of the website, as well as the recency of their visits and purchases.

A closer look into Table 3 and the dataset shows that the two tested campaigns were very successful with respect to purchasing rates as well as revenue per purchase and revenue per customer: whereas only 6.75% of the customers in the control groups purchased in the two weeks after the campaign, 11.76% in the treatment groups did so. The purchasers in the treatment groups bought on average items worth 183.04 €, whereas in the control groups, the bought items per purchaser were only worth 156.32 € on average. This difference is even more striking when taking all customers in the two samples into account (21.53 € per customer in the treatment groups vs. 10.55 € in the control groups). In the treatment groups, 25% of the purchasers bought items worth less than 50 € and 25% bought

Table 3 Descriptive uplift statistics of the BAUR dataset

Group	Share (%)	Customers	Purchasers	Purch rate (%)	Purch. upl. (%)	Revenue/ purch. (€)	Revenue/ cust. (€)	Revenue/ purch. upl. (€)	Profit/ purch (€)	Profit/ cust (€)	Profit uplift/ cust.(€)
Treatment	49.97	77,648	9,133	11.76	5.02	183.04	21.53	10.98	18.30	2.15	− 1.01
Control	50.03	77,740	5,244	6.75		156.32	10.55		46.90	3.17	
Total		155,388	14,377	9.25		173.92	16.04		28.73	2.66	

With margin m = 30% and discount d = 20%, disguised

items worth more than 208 € with a median at 104.90 €. In the control groups, 25% of the purchasers bought items worth less than 40 € and 25% bought items worth more than 179,95 € with a median at 84,99 €. 193 (20) customers in the treatment groups and 90 (3) in the control groups bought items worth more than 1.000 € (2.000 €) with a maximum at 4,863.87 € in the treatment groups and a maximum at 2,873.97 € in the control groups.

However, Table 3 also shows a major problem with discount offers. Assuming a (disguised) margin of m = 30% and a discount of d = 20%, the profit per purchaser and the profit per customer in the treatment groups (10% of the revenue) is clearly lower than in the control groups (30% of the revenue). This results in an overall profit per purchase uplift of the tests of − 1.01 €: offering the discount to all customers in the company's customer base seems to increase the overall revenue, but it would decrease the overall profit. So, a concentration on customers with positive uplift predictions and the development of a predictive scoring system is necessary.

The provided data from the two tests were randomly partitioned into a train set (~ 50% or 77,617 customers) and a holdout test set (~ 50% or 77,771 customers). Additionally, for hyperparameter tuning, the train set was randomly partitioned into a calibration set (~ 4/7 of the train set or 44,353 customers) and a validation set (~ 3/7 of the train set, 33,264). For all customers, besides the above-discussed variables that describe the belonging to the treatment and to the control groups, the purchase information, and the generated revenue, altogether 472 metric variables with a non-zero variance that describe their past information and buying behavior were available. Table 4 gives a short description of the 472 variables.

Based on the train set, the 472 variables were preprocessed by setting means to zero, setting standard deviations to 1, and applying a Box-Cox-transformation to transform skew distributed variables into "normal shape". Moreover, since the variables were highly correlated and—according to Devriendt et al. (2018)—the "best" number of predictors for uplift models has proven to be low (say 5–15), the variables were transferred to principal components. Here, the first 88 principal components accounted for 95%, the first 55 for 90%, and the first 20 for 75%, of the variance in the transformed training data. The same preprocessing (including the transformation into principal components) was applied to the test set, using the transformation parameters and coefficients derived from the train set. It should be mentioned that most variables (407 of 472) reflect the traditional RFM (recency, frequency, monetary value) scoring aspects in direct marketing, but their diversity with respect to various discount types, item categories, and time slots—as being obvious from a practical point of view and principal component analysis demonstrates from a statistical point of view—could improve the prediction.

4.2 Applying the profit uplift modeling approaches

As described in Sect. 3, four profit uplift modeling approaches were used for training and testing a scoring system:

Table 4 472 variables of the BAUR dataset that describe past information and buying behavior

Variable category	Number of variables	Description
Recency	23	Variables that count days since last order (w.r.t. discount types, item categories, and time slots)
Frequency	193	Variables that count past orders (w.r.t. discount types, item categories, and time slots)
Monetary value	191	Variables that reflect past revenues (w.r.t. discount types, item categories, and time slots)
Shop visit	14	Variables that describe the online information behavior (w.r.t. number of visits, visit duration, basket size and value across time slots and item categories)
Sensitivity to recommendations	3	Variables that describe the number of orders and their value due to recommendations (w.r.t. time slots)
Sensitivity to discounts	26	Variables that describe the share of orders with discounts to all orders in the past (w.r.t. discount types, item categories, and time slots)
Return behavior	22	Variables that describe the number of returns and their value (w.r.t. time slots)

- Heckman: the two-stage Heckman selection model (Heckman 1979) is estimated based on the binary outcome (purchase) and—in case of a predicted purchase—on the profit uplift. For parameter estimation, first, the observed profit for all purchasers is derived from the observed revenue by multiplying with the margin (m = 30%) for the purchasers in the control group and with the margin minus discount (m–d = 10%) for the purchasers in the treatment group. Then, the profit response is transformed to "observed" profit uplift according to formula (4), and the Heckman selection model is estimated. Finally, profit uplift predictions can be directly derived for all customers using formula (3) via formulae (5) and (6). Besides this direct model approach (using the "observed" profits for estimation) also a two model approach according to formula (2) was used (using the profit responses in the treatment and control group for separate estimations). For all estimations, the R package and R function sampleSelection was applied.
- OLS and RF: as one-stage models, simple regression (OLS) and random forest (RF) (Breiman 2001) is used. We apply glm from the MASS package in R in case of OLS and the ranger implementation in R (Wright and Ziegler 2017) in case of RF to the "observed" profit uplifts as a direct modeling approach. Again, predictions for the profit uplift outcome can be derived for all customers directly according to formula (3).
- Zeroinfl: the two-stage zero-inflated Poisson regression model (Lambert 1992) and its zero-inflated negative binomial regression model alternative (Ridout et al. 2001) assume non-negative count data as input. Therefore, first, the observed profit has to be converted to Millicent (to preserve variability) and to be rounded. Also, as discussed in Sect. 3, an interaction model is useful (as an alternative to the two model approach) that includes the treatment indicator (1 for customers in the treatment group, 0 for the others) and its interactions with the other predictors. The estimated interaction model then is used for predicting the profit uplift as the difference between the predicted profit when the treatment indicator is set to 1 and the predicted profit when the treatment indicator is set to 0 according to formula (2). In our applications, we use the zero-inflated negative binomial regression model due to overdispersion in the train dataset. Hurdle models were also tested but showed no improvement compared to the zero-inflated models. The R package pscl is applied.

Before estimating the models based on the train data and comparing the results on the test data—as usual in machine learning—reflections on performance evaluation and parameter tuning are necessary. As already discussed in Sect. 3, the incremental profit uplift curve and the derived profit Qini coefficient are suitable measures for this purpose. Since only one observation per customer is available in the data (profit if treated or profit if not treated due to the belonging to the treatment or the control group), for calculating uplifts a grouping of customers and comparing average profits of treated and not treated customers in each group is needed. This grouping is based on sorting the customers according to the developed scoring system (starting with the customers where we assume the highest profit uplift) and forming quantiles (usually deciles) of the sorted customers. Basing on these groupings, now, the incremental profit uplift across the quantiles can be plotted (the incremental profit

uplift curve), and the area between this curve and a curve derived by random sorting (the profit Qini coefficient Q) can be calculated and used for selecting best scoring systems.

Figure 2 shows the profit Qini coefficients for the four discussed models (OLS, Heckman, and RF as direct models, Zeroinfl as interaction model) when estimated with varying numbers of predictors on the basis of the calibration subsample of the train data and used for predictions on the basis of the validation sample. Note that the profit Qini coefficients reflect the area between the profit Qini curve and the curve for the random model as in Fig. 1 and that larger values indicate a better sorting of the customers according to their "observed" profit uplift (calculated via groups of customers with similar uplift predictions). It can be easily seen that the profit Qini coefficients are low with small numbers of predictors as well as with high numbers of predictors. These findings are consistent with the findings of Devriendt et al. 2018), who found in their comparison of binary uplift models that 5–15 predictors typically provide the best results. Against this background, we decided to use 20 predictors in the following for training and testing our models and to compare them with revenue response and uplift as well as profit response models. It can also be seen that overall the two direct regression models regression (OLS and Heckman) performed quite similar in this tuning analysis. In the following, when we concentrate on the 20 principal components as a result of hyperparameter tuning but now analyze the stability of these results in more detail, consequently, we additionally applied the Heckman two-model approach to elaborate further differences.

Table 5 and Fig. 3 already show the results of this extended evaluation of profit uplift modeling approaches: Four profit uplift modeling approaches were applied

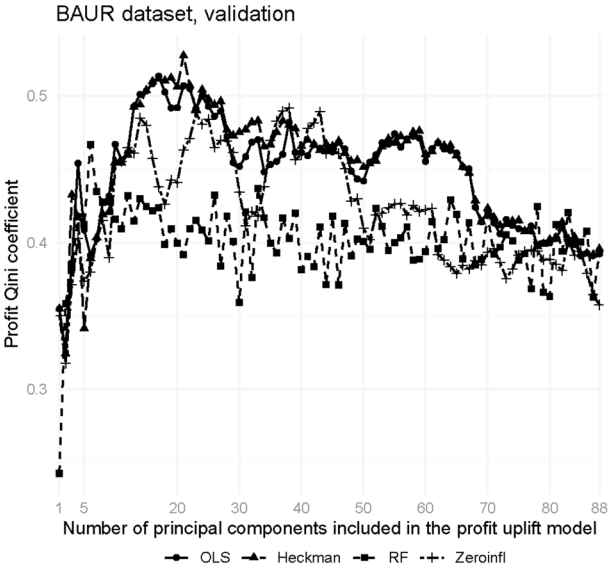

Fig. 2 Profit Qini coefficients for the validation set (3/7 of the BAUR train set) based on training the profit uplift modeling approaches on the calibration set (4/7 of the BAUR train set)

Table 5 Results of the application of profit uplift modeling approaches to the BAUR dataset (20 principal components): 50 random subsamples (6/7) of the train set (50%) were used to calibrate the models and predict the profit uplift in the test set (50%)

Modeling approach			Train set	Test set
		Estimation	Profit Qini coefficient	Profit Qini coefficient
Profit uplift	Direct	OLS	0.495 (0.022)	0.421 (0.013)
	Two model	Heckman	0.228 (0.084)	0.148 (0.097)
	Direct	RF	0.568 (0.026)	0.344 (0.009)
	Interaction	Zeroinfl	0.471 (0.024)	0.402 (0.012)

The profit Qini coefficients were averaged (standard deviations in brackets)

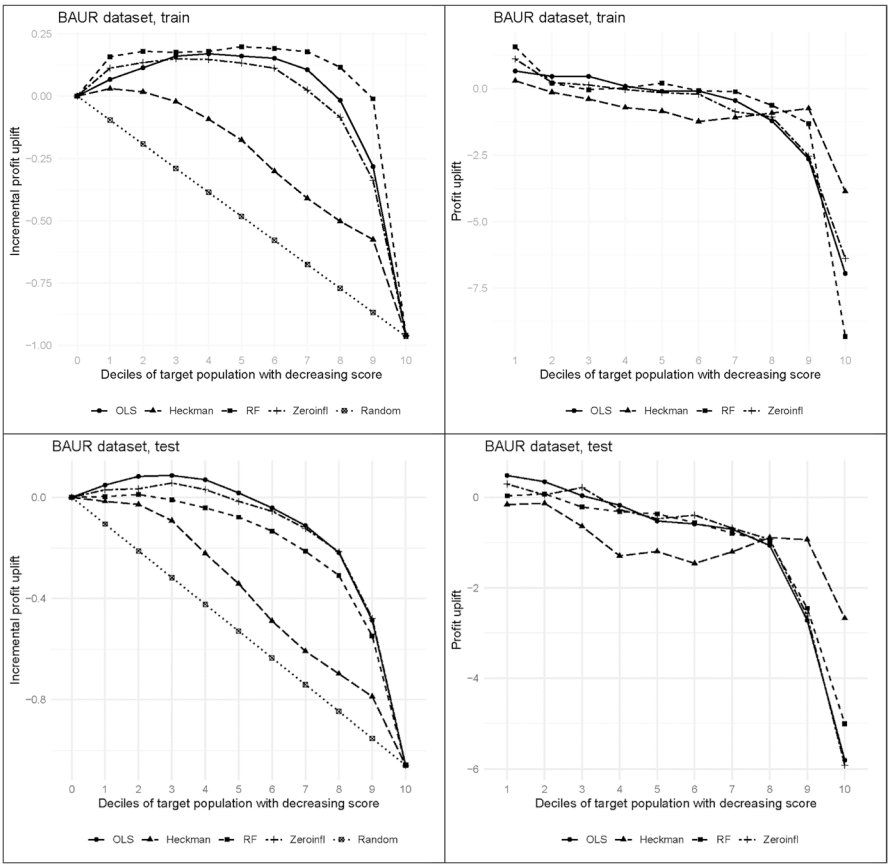

Fig. 3 Application of profit uplift modeling approaches to the BAUR dataset (20 principal components): 50 random subsamples (6/7) of the train set (50%) were used to calibrate the models and predict the profit uplift in the test set (50%). The resulting profit Qini curves (left) and mean profit uplifts (right) were averaged

to 50 randomly drawn subsamples (6/7) of the train set (50% of the BAUR dataset). The profit Qini coefficients were calculated and averaged (see mean values and standard deviations in Table 5 and the mean Qini curve in Fig. 3). Then, the estimated profit uplift models were applied to the test data, and—again—the profit Qini coefficients were calculated and averaged (mean values and standard deviations in Table 5, mean Qini curve in Fig. 3).

The results reflect the results of parameter tuning (Please note that since the OLS and Heckman direct model performed similar, therefore in Table 5 and Fig. 3 the results of the Heckman two model approach are given instead): the direct model with OLS parameter estimation performs best with respect to the holdout test set, followed by the Zeroinfl interaction model, and the RF direct model. The Heckman two-model approach is inferior to these three approaches. However, as Fig. 3 demonstrates, OLS, RF, and Zeroinfl provide quite similar results, which is—to some extent – surprising since the modeling assumptions ("normal shape" vs. count data, direct model vs. difference of two predictions based on the interaction model) and the estimation algorithms (one-step vs. two-step estimation) are very different.

It should be mentioned that the profit Qini coefficients in Table 5 and the Qini curves in Fig. 3 are used to select a "best" predictive model and not for deciding on "best" customers for the direct marketing campaign. All customers of the train set and the test set were allocated to the treatment or the control group, and the Qini coefficients and the Qini curves just reflect whether a derived model from the train set is able to correctly predict the uplifts in the test set. When the decision with respect to a best model is made, this best model then is applied to all customers and customers with a predicited positive profit uplift should be included into the direct marketing campaign. However, for these final step, no Qini coefficients or Qini curves can be derived since the necessary balanced distribution of respondents in the tran and test group is not given. The application at least shows that it seems to be possible that—besides already existing binary uplift and revenue uplift models—it is possible to estimate profit uplift models which show clear practical advantages. In the following, we analyze this theoretical superiority using the BAUR dataset by comparing the new approaches with traditional ones.

4.3 Comparison of revenue and profit response and uplift modeling approaches

A detailed comparison of the proposed profit uplift modeling approaches to already known revenue response and uplift modeling approaches, but also profit response modeling approaches is used to clarify differences between these approaches. Again, the train set (50%) and the test set (50%) of the BAUR dataset is used for comparisons, 50 subsamples (6/7) of the train set were drawn and used to calibrate the various models under study. For each model and for each subsample of the datasets, scores are predicted for the customers in the data used for training and for the customers in the test set. Please note that in the case of response models, these scores reflect a revenue or profit response (depending on the model), and in the case of uplift models, these scores reflect a revenue or profit uplift. Based on the sorting of the customers according to these predicted (revenue or profit, response or uplift)

scores, then, revenue and profit Qini curves, as well as revenue and profit Qini coefficients, can be calculated. These coefficients were averaged across the 50 random subsamples similar as in the previous subsection (Indeed, the profit Qini values for the profit uplift modeling approaches are the same in Tables 5 and 6).

Table 6 as well as Figs. 4, 5, and 6 reflect the results of these modeling and prediction endeavors (Please note that Figs. 4 and 5 show revenue uplift curves whereas Figs. 3 and 6 show profit uplift curves): first, one can see, that altogether 16 modeling approaches were used in this comparison, each applied to 50 subsamples of the train set. The response models were calibrated based on the treated customers in the train set. The aim was to predict the individual revenue or profit of treated customers without taking into account whether the customer would have also bought without being treated. As Fig. 4 (for revenue response models) and Fig. 6 (for profit response models) demonstrate, these response modeling approaches also convince when the customers of the train and test set should be sorted according to their estimated revenue uplift, but—according to Table 6—not when a profit uplift sorting is needed. However, as Table 6 clearly demonstrates: the response models are inferior to the uplift models in all cases, i.e. when predicting uplifts is needed and measured via the revenue and the profit Qini coefficients.

The same holds when revenue response and uplift are used to predict profit uplifts: Figs. 4 and 5, as well as Table 6, demonstrate that these models are quite good in predicting revenue responses and uplifts. However, the profit Qini curves (not shown in the Figures) evaluated via the profit Qini coefficients in Table 6 show negative values, which means that these models are inferior even to a random model. Finally, Fig. 6 also shows that for profit uplift prediction, the application of a profit response model is not enough.

To summarize: the extensive comparison of various response and revenue uplift models applied to the BAUR dataset reflects promising results for the usefulness of the new profit uplift modeling approaches. Even an application with a simple (one-stage) regression or a random forest estimation algorithm applied to transformed profit data outperforms these traditional models. In the following, we investigate whether this superiority can also be found when a well-known and publicly available dataset is used, which has often been the basis for introducing new uplift modeling approaches.

5 Application to the Hillstrom dataset

In order to demonstrate that the new profit uplift modeling approaches are applicable and superior to traditional response and uplift modeling approaches, also a standard dataset from the uplift modeling literature is analyzed, the Hillstrom dataset (Radcliffe 2008). This dataset was made available by Kevin Hillstrom through his MineThatData blog and contains a sample of 64,000 customers which had been divided up into three nearly equally sized subsamples, two of them contacted via two direct marketing campaigns and one not contacted, serving as a control group (see the similar usage of this dataset, e.g., in Rudaś and Jaroszewicz 2018). Table 7 summarizes the descriptive uplift statistics of this dataset, where the two treated

Table 6 Results of the application of revenue and profit response and uplift modeling approaches to the BAUR dataset (20 principal components): 50 random subsamples (6/7) of the train set (50%) were used to calibrate the models and predict the revenue and profit uplift in the test set (50%)

Modeling approach			Train set		Test set	
		Estimation	Revenue Qini coefficient	Profit Qini coefficient	Revenue Qini coefficient	Profit Qini coefficient
Revenue response	Direct	OLS	2.612 (0.099)	− 0.329 (0.019)	2.370 (0.026)	− 0.318 (0.006)
		Heckman	2.424 (0.096)	− 0.302 (0.018)	2.133 (0.000)	− 0.295 (0.000)
		RF	2.920 (0.102)	− 0.296 (0.018)	2.355 (0.018)	− 0.317 (0.004)
		Zeroinfl	2.575 (0.103)	− 0.300 (0.020)	2.400 (0.017)	− 0.278 (0.004)
Revenue uplift	Direct	OLS	2.744 (0.088)	− 0.178 (0.022)	2.531 (0.045)	− 0.162 (0.020)
	Two model	Heckman	2.648 (0.135)	− 0.106 (0.058)	2.227 (0.102)	− 0.139 (0.055)
	Direct	RF	3.785 (0.126)	0.026 (0.032)	2.438 (0.036)	− 0.234 (0.012)
	Interaction	Zeroinfl	2.526 (0.109)	− 0.127 (0.028)	2.370 (0.079)	− 0.118 (0.019)
Profit response	Direct	OLS	2.612 (0.099)	− 0.329 (0.019)	2.370 (0.026)	− 0.318 (0.006)
		Heckman	2.504 (0.119)	− 0.296 (0.020)	2.210 (0.090)	− 0.286 (0.006)
		RF	2.920 (0.102)	− 0.296 (0.018)	2.355 (0.019)	− 0.317 (0.004)
		Zeroinfl	2.575 (0.103)	− 0.300 (0.020)	2.400 (0.017)	− 0.278 (0.004)
Profit uplift	Direct	OLS	− 0.118 (0.150)	0.495 (0.022)	− 0.427 (0.101)	0.421 (0.013)
	Two model	Heckman	1.118 (0.391)	0.228 (0.084)	0.655 (0.281)	0.148 (0.097)
	Direct	RF	− 0.778 (0.202)	0.568 (0.026)	− 1.990 (0.069)	0.344 (0.009)
	Interaction	Zeroinfl	0.306 (0.164)	0.471 (0.024)	− 0.025 (0.109)	0.402 (0.012)

The revenue and profit Qini coefficients were averaged (standard deviations in brackets)

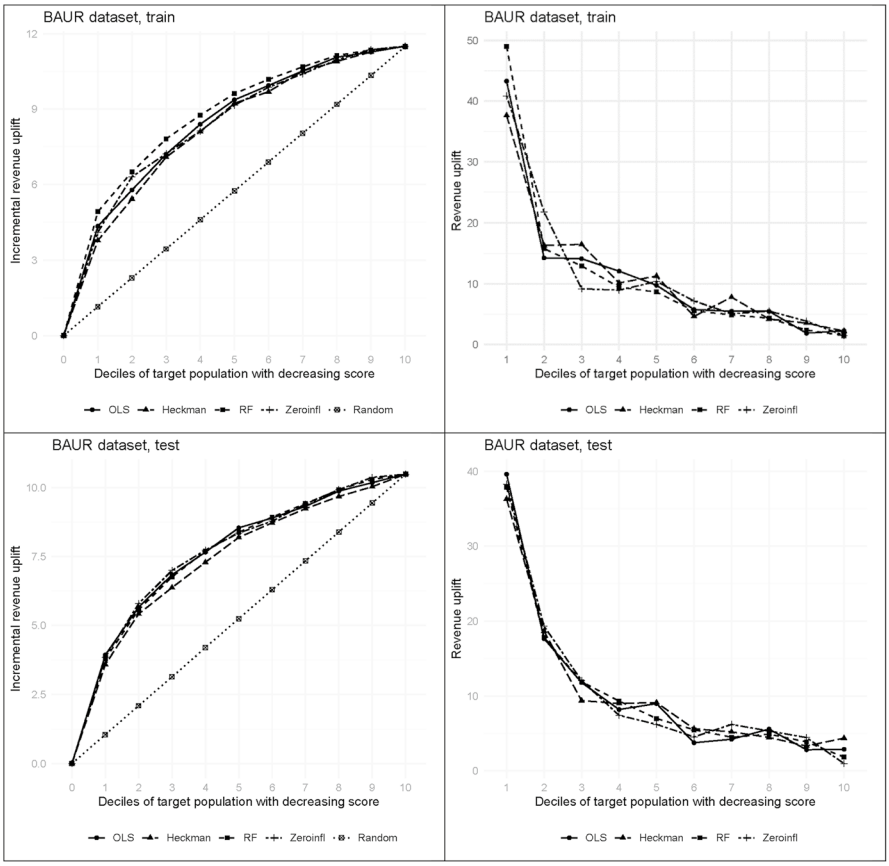

Fig. 4 Application of revenue response modeling approaches to the BAUR dataset (20 principal components): 50 random subsamples (6/7) of the train set (50%) were used to calibrate the models and predict the revenue uplift in the test set (50%). The resulting revenue Qini curves (left) and mean revenue uplifts (right) were averaged

subsamples are merged. As can easily be seen, the conversion rate is much lower as in the BAUR dataset (on average 1.07% in the treatment group) but nevertheless shows a conversion rate uplift compared to the control group (on average, an uplift of 0.50%). The revenue uplift per customer is 0.60$, but this uplift seems to be arising solely from the conversion rate uplift since the average revenue spend by a purchaser in the treatment group (117.00$) is only slightly higher than in the control group (114.00$). Again, as in the BAUR dataset, we assume that the campaign offers a 20% discount and that the margin for the retailer is 30%. With these assumptions (not part of the original communication of the dataset, just an assumption to be able to analyze the dataset with our profit uplift modeling approaches), the overall profit uplift per customer is negative (− 0.07$). So, again we have to develop a scoring system that helps to restrict the direct marketing campaign to customers with a positive profit uplift prediction.

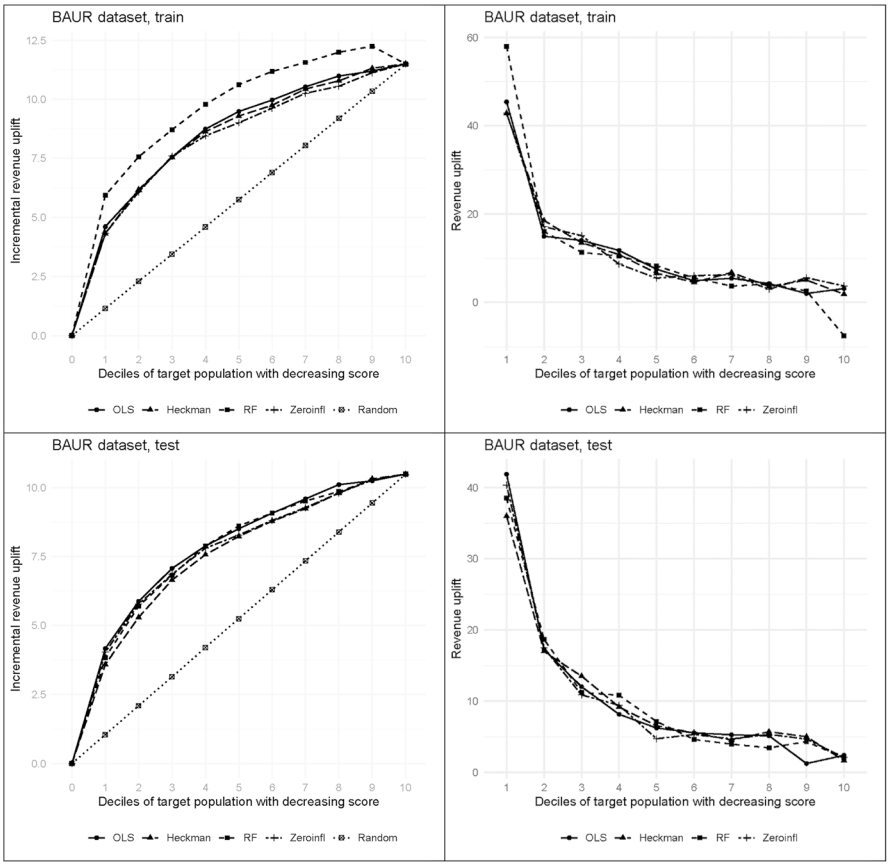

Fig. 5 Application of revenue uplift modeling approaches to the BAUR dataset (20 principal components): 50 random subsamples (6/7) of the train set (50%) were used to calibrate the models and predict the revenue uplift in the test set (50%). The resulting revenue Qini curves (left) and mean revenue uplifts (right) were averaged

The original dataset also contains potential predictors for this scoring system, as given in Table 8. The original eight potential predictors (in Table 8 described as variable categories) were scaled nominally (e.g., history_segment with 7 values or channel with three values) or metrically (e.g., recency or history). For our further analysis with the three models, we dummy-coded the nominally scaled potential predictors and so received in total 25 metrically scaled variables (see Table 8).

As in Sect. 4, the customers were randomly partitioned into a train set (~ 70% or 44,800 customers) and a holdout test set (~ 30% or 19,200 customers), and the train set was preprocessed by setting means to zero, setting standard deviations to 1, and applying a Box–Cox-transformation to transform skew distributed variables into "normal shape". The same preprocessing was applied to the test set, using the transformation parameters derived from the train set. Then, five models, similar as in Sect. 4, were applied: 50 subsamples (6/7) of the train set were

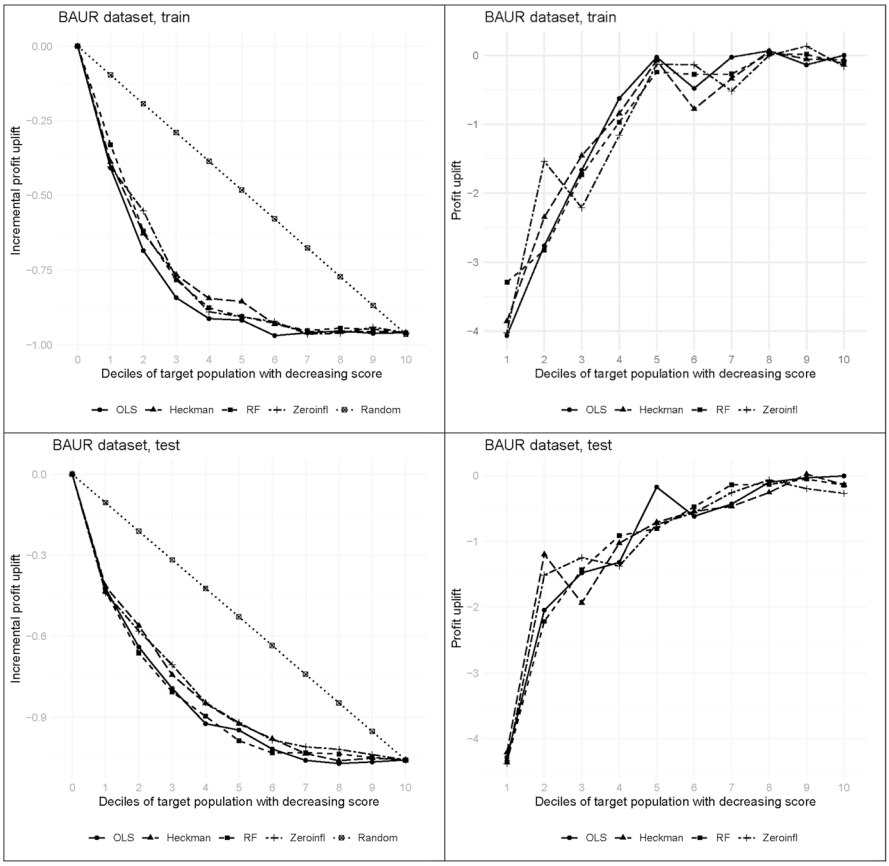

Fig. 6 Application of profit response modeling approaches to the BAUR dataset (20 principal components): 50 random subsamples (6/7) of the train set (50%) were used to calibrate the models and predict the profit uplift in the test set (50%). The resulting profit Qini curves (left) and mean profit uplifts (right) were averaged

drawn randomly, the models were calibrated, and the Qini curves and Qini coefficients were averaged. Figure 7 and Table 9 reflect the results of this modeling and prediction task. It should be mentioned that we only use a subsample of the models applied in Sect. 4 but the selection contains the "best" models from this comparison (e.g. especially the three "winners" OLS and RF direct models and Zeroinfl interaction model).

One can easily see that the three "best" profit uplift modeling approaches (one-stage OLS and RF as well as two-stage Zeroinfl), again, show similar results with random forest providing the best performance. But it should be mentioned that—maybe due to the few purchasers in the dataset with a high concentration of revenues and profits from few purchasers—the modeling leads to a worse performance compared to the application of the BAUR dataset. This problem of the Hillstrom dataset when it comes to modeling continuous outcomes has also been

Table 7 Descriptive uplift statistics of the Hillstrom dataset (with margin m = 30% and discount d = 20%)

Group	Share (%)	Customers	Purchasers	Conv rate (%)	Conv. upl. (%)	Revenue/conv. ($)	Revenue/cust. ($)	Revenue/cust. upl.($)	Profit/ conv ($)	Profit/ cust ($)	Profit uplift/ cust. ($)
Treatment	66.71%	42,694	456	1.07%	0.50%	117.00	1.25	0.60	11.70	0.12	− 0.07
Control	33.30%	21,306	122	0.57%		114.00	0.65		34.20	0.20	
Total		64,000	578	0.90%		116.36	1.05		16.45	0.15	

Table 8 Variables of the Hillstrom dataset that describe past buying behavior

Variable category	Number of variables	Description
Recency	12 (− 1)	Indicators for months since last purchase (1,…,12)
History_ segment	7 (− 1)	Indicators for revenue categories last year ([0,100$), [100$,200$), [200$,350$), [350$,500$), [500,750$), [750$,1000$), [1000$,)
History	1	Revenue generated last year (in $)
Mens	1	Indicator whether customer bought men's merchandise last year
Womens	1	Indicator whether customer bought women's merchandise last year
Zipcode	3 (− 1)	Indicator whether the customer's zip code is rural, suburban, urban
Newbie	1	Indicator whether customer bought last year for the first time
Channel	3 (− 1)	Indicator whether customer bought last year via phone, web, both

(− 1 indicates that one indicator is dependent on the others and therefore is omitted for estimation)

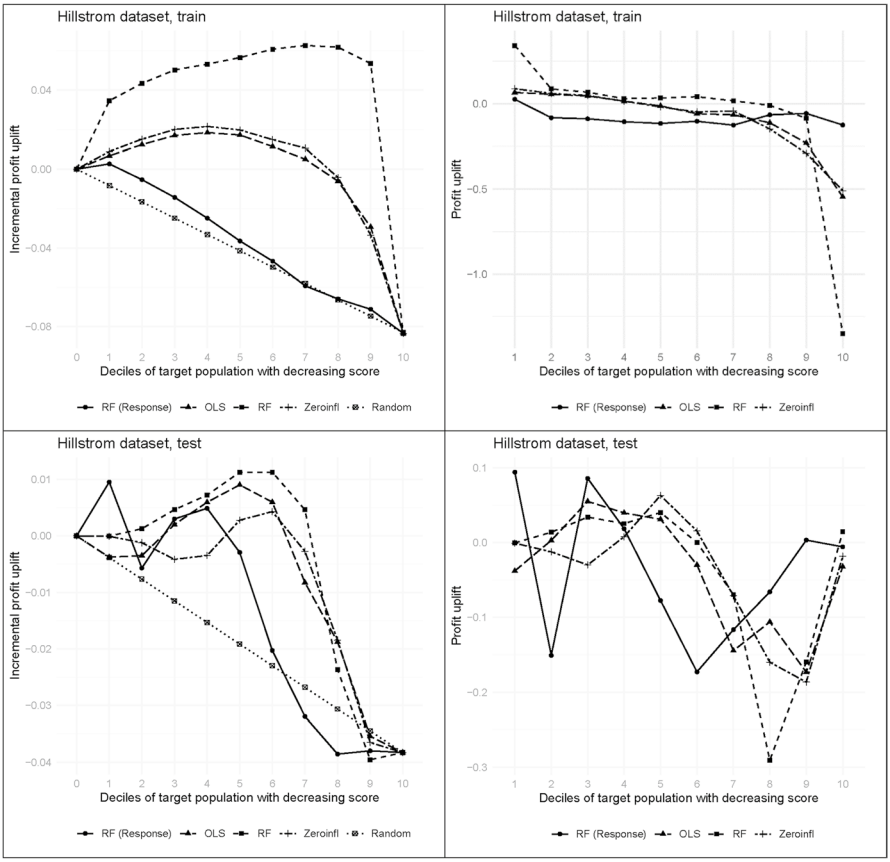

Fig. 7 Results of the application of revenue and profit response and uplift modeling approaches to the Hillstrom dataset: 50 random subsamples (60% of the data) of the train set (70%) were used to calibrate the models and predict the profit uplift in the test set (30%). The resulting Qini curves (left) and mean uplifts (right) were averaged

Table 9 Results of the application of revenue and profit response and uplift modeling approaches to the Hillstrom dataset (25 variables): 50 random subsamples (6/7) of the train set (70%) were used to calibrate the models and predict the profit uplift in the test set (30%)

Modeling approach		Estimation	Profit Qini coefficient for the train set	Profit Qini coefficient for the test set
Revenue response	Direct	RF	0.005 (0.005)	0.005 (0.003)
Profit response	Direct	RF	0.005 (0.005)	0.005 (0.003)
Profit uplift	Direct	OLS	0.043 (0.005)	0.013 (0.004)
	Direct	RF	0.085 (0.004)	0.015 (0.005)
	Interaction	Zeroinfl	0.045 (0.005)	0.011 (0.005)

The profit Qini coefficients were averaged (standard deviations in brackets)

discussed by other authors in context of revenue uplift modeling; here we refer to the analysis in the paper by Rudaś and Jaroszewicz 2018).

6 Conclusions and outlook

In this paper, we demonstrated the usefulness of new profit uplift modeling approaches for direct marketing campaigns: the main idea is to contact a sample of customers testwise. Then, profit uplifts are modeled and the model is used to make predictions across all customers. Finally, these predictions are used to decide whether a customer should be contacted.

In contrast to former approaches, the proposed new approaches model profit uplift at the individual level and do not need an unrelated second step to transform modeled binary outcomes or revenues to profits. Various algorithms can be applied to estimate the model parameters. On the one side, two-stage algorithms especially tackle the problem of low rates of purchasers (in many cases: zero revenues and profits). So, the Heckman sample selection model separately models the observed binary outcome (purchase or not) and the related observed continuous outcome (profits for the treatment group, profits with negative sign for the control group in the direct model case, profits for both groups in the two model approach). Zero-inflated negative binomial or hurdle models—as an alternative—assume count data and therefore need an interaction model for estimation and prediction. On the other side, one-stage algorithms like OLS and RF performed surprisingly well, especially when applied to transformed profits.

The proposed profit uplift modeling approaches are based on very different assumptions but nevertheless provide quite similar predictions with a clear ordering of the customers according to their predicted profit uplift. The results support the meaningfulness of the approaches via cross-validation as the main contribution of this paper. Also, an extensive comparison with response models and revenue uplift models using two large datasets supports this superiority of the new approaches.

Of course, further research is needed. So, e.g., the presented profit uplift modeling approaches have to demonstrate their usefulness also with other datasets.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was conducted within the research project "Wissenschaftscampus E-Commerce" funded by Bayerisches Staatsministeriums für Wirtschaft, Landesentwicklung und Energie, Munich, Germany, and BAUR Group, Burgkunstadt, Germany.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Baier D, Rese A, Nonenmacher N, Treybig S, Bressem B (2019) Digital technologies for ordering and delivering fashion: how Baur integrates the customer's point of view. In: Urbach N, Röglinger M (eds) Digitalization cases. Springer, Cham, pp 59–77

Blattberg RC, Kim B-D, Neslin SA (2008) Database marketing—analyzing and managing customers. Springer, New York, NY

Breiman L (2001) Random forests. Mach Learn 45:5–32

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 6:13–41. https://doi.org/10.1089/big.2017.0104

Gubela RM, Lessmann S, Jaroszewicz S (2020) Response transformation and profit decomposition for revenue uplift modeling. Eur J Oper Res 283:647–661. https://doi.org/10.1016/j.ejor.2019.11.030

Guelman L, Guillén M, Pérez-Marín AM (2015) Uplift random forests. Cybern Syst 46:230–248. https://doi.org/10.1080/01969722.2015.1012892

Hansotia B, Rukstales B (2002) Incremental value modeling. J Interact Mark 16:35–46. https://doi.org/10.1002/dir.10035

Heckman J (1979) Sample selection bias as a specification error. Econometrica 47:153–161

Kane K, Lo VS, Zheng J (2014) Mining for the truly responsive customers and prospects using true-lift modeling: comparison of new and existing methods. J Market Anal 2:218–238. https://doi.org/10.1057/jma.2014.18

Lai LY-T (2006) Influential marketing: a new direct marketing strategy addressing the existence of voluntary buyers. Master Thesis, School of Computing Science—Simon Fraser University

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

Lo VSY (2002) The true lift model. ACM SIGKDD Explorations Newsl 4:78–86

Mullahy J (1986) Specification and testing of some modified count data models. J Econom 33:341–365. https://doi.org/10.1016/0304-4076(86)90002-3

Radcliffe NJ (2007) Using control groups to target on predicted lift: building and assessing uplift models. Direct Mark Anal J 1:14–21

Radcliffe NJ (2008) Hillstrom's MineThatData email analytics challenge: an approach using uplift modelling. Stochastic Solutions Ltd., Edinburgh. https://www.stochasticsolutions.com/pdf/HillstromChallenge.pdf

Radcliffe NJ, Surry PD (1999) Differential response analysis: Modeling true response by isolating the effect of a single action. In: Proceedings of Credit Scoring and Credit Control VI. Credit Research Center, University of Edinburgh Management School

Radcliffe NJ, Surry PD (2011) Real-world uplift modelling with significance-based uplift trees. Technical Report TR-2011-1, Stochastic Solutions Ltd., Edinburgh. https://www.stochasticsolutions.com/pdf/sig-based-up-trees.pdf

Ridout M, Hinde J, Demétrio CG (2001) A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57:219–223. https://doi.org/10.1111/j.0006-341X.2001.00219.x

Rudaś K, Jaroszewicz S (2018) Linear regression for uplift modeling. Data Min Knowl Disc 32:1275–1305. https://doi.org/10.1007/s10618-018-0576-8

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

Sołtys M, Jaroszewicz S, Rzepakowski P (2015) Ensemble methods for uplift modeling. Data Min Knowl Disc 29:1531–1559. https://doi.org/10.1007/s10618-014-0383-9

Toomet O, Henningsen A (2008) Sample selection models in R: package sample selection. J Stat Soft. https://doi.org/10.18637/jss.v027.i07

Wright MN, Ziegler A (2017) Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Soft 77:1–17. https://doi.org/10.18637/jss.v077.i01

Zaniewicz L, Jaroszewicz S (2013) Support vector machines for uplift modeling. In: 2013 IEEE 13th International Conference on Data Mining Workshops. IEEE

Zaniewicz Ł, Jaroszewicz S (2017) L_p-support vector machines for uplift modeling. Knowl Inf Syst 53:269–296. https://doi.org/10.1007/s10115-017-1040-6