Data Mining and Knowledge Discovery (2020) 34:273–308
https://doi.org/10.1007/s10618-019-00670-y

A survey and benchmarking study of multitreatment uplift modeling

Diego Olaya[1] · Kristof Coussement[2] · Wouter Verbeke[1]

Received: 17 May 2019 / Accepted: 14 December 2019 / Published online: 13 January 2020

Abstract

Uplift modeling is an instrument used to estimate the change in outcome due to a treatment at the individual entity level. Uplift models assist decision-makers in optimally allocating scarce resources. This allows the selection of the subset of entities for which the effect of a treatment will be largest and, as such, the maximization of the overall returns. The literature on uplift modeling mostly focuses on queries concerning the effect of a single treatment and rarely considers situations where more than one treatment alternative is utilized. This article surveys the current literature on multitreatment uplift modeling and proposes two novel techniques: the naive uplift approach and the multitreatment modified outcome approach. Moreover, a benchmarking experiment is performed to contrast the performances of different multitreatment uplift modeling techniques across eight data sets from various domains. We verify and, if needed, correct the imbalance among the pretreatment characteristics of the treatment groups by means of optimal propensity score matching, which ensures a correct interpretation of the estimated uplift. Conventional and recently proposed evaluation metrics are adapted to the multitreatment scenario to assess performance. None of the evaluated techniques consistently outperforms other techniques. Hence, it is concluded that performance largely depends on the context and problem characteristics. The newly proposed techniques are found to offer similar performances compared to state-of-the-art approaches.

Keywords Uplift modeling · Business analytics · Prescriptive analytics · Causality · Matching · Literature survey

Responsible editor: Johannes Fürnkranz.

This paper was supported by Innoviris, the Brussels Region Research funding agency.

Extended author information available on the last page of the article

1 Introduction

Predictive analytics supports decision-making by exploiting the patterns present in historical data to obtain insights about future states. Machine learning techniques play a crucial role, as they facilitate the estimation of the likelihood of an outcome of interest. However, a key concern in real-world applications lies in foreseeing the effects of different actions on an outcome variable. This task is performed by uplift modeling techniques and allows decision-makers to prescribe the course of action that maximizes a given objective at the individual level. Hence, uplift modeling is a type of prescriptive analytics (Bertsimas and Kallus 2019).

The identification of the most favorable action (hereafter referred to as *treatment*) for an individual corresponds to estimating the effect that a decision variable (e.g., treatment) has on an outcome variable (e.g., response). This association is known in the causal literature as the individual treatment effect (ITE) and frames uplift modeling as a causal inference task. The potential outcomes framework (Rubin 1974) defines the ITE as the difference between potential outcomes of distinct treatment alternatives. From a machine learning perspective, this consists of contrasting the predicted values of the outcome variable for each of the treatments at the individual level.

Since making causal inferences is tied to a treatment applied to an individual, uplift modeling is functional in cases where a decision-maker has control over a variable whose manipulation is expected to cause a behavioral change. For instance, marketers launch campaigns that maximize the intentions of customers to buy particular products (Gubela et al. 2017).

Uplift modeling can be implemented in different domains. However, the most common applications are found in the fields of marketing (Lo 2002; Hansotia and Rukstales 2002; Guelman et al. 2012, 2014a, b; Kane et al. 2014; Guelman et al. 2015; Gross and Tibshirani 2016; Michel et al. 2017; Gubela et al. 2017) and personalized medicine (Alemi et al. 2009; Jaskowski and Jaroszewicz 2012). Particularly, uplift modeling has helped marketers to increase the return on marketing investment by segmenting the customer base into four categories according to the recommendations of the model. Customers who respond favorably because of the campaign are categorized as *persuadables*. On the other hand, the *do-not-disturb* segment includes customers adversely affect by the campaign: they do not respond at all, while they would have responded if they were not contacted. The customers in the third and fourth categories either never respond to any offer—the *lost causes*—or always respond regardless of the offer—the *sure things*. The interest lies in targeting the *persuadables* and avoiding the other segments.

The literature on uplift modeling is primarily focused on the estimation of a single treatment effect. Studies that generalize the binary treatment effect framework to applications where the effects of different treatment alternatives are estimated are scattered and limited in number. Hence, there is at most a vague understanding of which uplift multitreatment techniques are available and limited empirical evidence regarding the uses and performances of these methods.

This study contributes to the state-of-the-art in the field of uplift modeling in three ways: (1) it provides an exhaustive survey of the literature on multitreatment uplift modeling and introduces a framework to classify multitreatment uplift modeling methods, (2) it proposes two novel multi-treatment uplift modeling methods, and (3) it presents the results of an extensive benchmarking study, which provides ample empirical evidence of the performances of thirteen multitreatment uplift modeling methods across eight multitreatment uplift data sets. The experiments are performed on data sets from diverse domains such as marketing, politics, personalized medicine and human resources. The Qini metric and the expected response are used to evaluate the performances of the models.

Additionally, uplift studies where selection bias is tested and controlled are uncommon. Therefore, we verify and, if needed, correct for the imbalance among the pretreatment characteristics of the treatment groups to ensure a correct interpretation of the estimated uplift.

The remainder of this paper is structured as follows. The first part provides a general introduction to the fundamentals of uplift modeling and an overview of current approaches to estimate uplift in a multitreatment scenario and presents two novel methods. Section 3 discusses the evaluation of multitreatment uplift models. Next, the experimental design is described in Sect. 4, and empirical results are discussed in Sect. 5. Finally, Sect. 6 concludes and provides directions for future research.

2 Uplift modeling

This section starts with a general definition of uplift modeling and a description of the single treatment and multitreatment scenarios. Next, we provide an overview of current uplift modeling techniques and propose two novel methods.

2.1 Definition

Uplift modeling is a machine learning approach that employs Rubin's causal inference framework (Rubin 1974) to estimate the ITE of (a) treatment(s) on an outcome of interest. The ITE estimation requires three elements to be present in the data: a set of variables representing the pretreatment characteristics of individuals, X, a decision-variable indicating the exposure to a treatment, T, and the corresponding outcome, Y.

In a binary treatment assignment, $Y_i(T = 1)$ and $Y_i(T = 0)$ correspond to the potential outcomes (i.e., the future state of the outcome) of an individual when she/he receives treatment and nontreatment, respectively. Then, the ITE of treatment against nontreatment on Y for the i individual, τ_i, is $Y_i(T = 1) - Y_i(T = 0)$. If the result of the subtraction is a nonzero value, it can be inferred that the treatment exerts an impact on the outcome for that particular individual. In uplift modeling, the potential outcomes are estimated by machine learning algorithms as conditional probabilities whose difference is used to determine the effect of the treatments. The multitreatment scenario is a generalization of Rubin's framework to applications where the decision variable can assume more than two values. Examples include the situations in which policy makers have to decide among various assistance programs or when marketers have to choose among different channels to reach out to customers.

Causal discovery infers causal structures from data with respect to interventions (Peters et al. 2017). The focus of uplift modeling, on the other hand, lies in customizing the treatment assignments. The aim is to target individuals on who the treatment will have the largest positive effect according to the predictions of the model. An analogous approach to uplift modeling is the estimation of heterogeneous treatment effects (Zhao and Harinen 2019). A large portion of this literature employs machine learning methods to estimate the conditional average treatment effect (CATE). The motivation behind the understanding of treatment effect heterogeneity is that the CATE can be used to select the optimal treatment rule, since it considers the treatment effectiveness to vary with the characteristics of individuals. Applications in the binary treatment case include those of Kallus (2017), Athey and Wager (2017), Kallus and Zhou (2018) and Athey and Imbens (2019).

To the best of our knowledge, the multitreatment setting has only been addressed by Imai et al. (2013) and Zhou et al. (2018). In contrast to uplift modeling, these methods serve to formulate treatment rules conditioned for individual characteristics, thus prioritizing the estimations of causal effects and statistical inference rather than their predictive power.

2.1.1 Binary model

Binary treatment uplift modeling is formally introduced by Radcliffe and Surry (1999) as a technique to predict the incremental effects of marketing activities. The difference between uplift modeling and response modeling is that the latter uses predictive models to estimate the likelihood of a favorable outcome. The former, however, predicts how much the outcome will vary when the individual is exposed to a treatment.

$$\hat{\tau}_{i,1}(x_i, T) := \hat{P}\big(Y_i = 1|x_i,\ do(T = 1)\big) - \hat{P}\big(Y_i = 1|x_i,\ do(T = 0)\big) \qquad (1)$$

Assuming a binary outcome variable $Y \in \{0, 1\}$, Eq. 1 defines the predicted individual uplift for $T = 1$ ($\hat{\tau}_{i,1}$) as a function of the individual's pretreatment characteristics X and the two treatment alternatives $T = \{0, 1\}$. This definition integrates the $do(\cdot)$ operator to indicate that the observed change in the probability of the outcome is due to the treatment itself and not to the presence of confounders (Pearl 2009). A fundamental assumption is that individuals are somehow sensitive to the given treatment (Guelman et al. 2014b). The contrast between the two groups allows the identification of the individuals who are most likely to have a favorable outcome when treated. This makes uplift modeling an appropriate tool for customizing treatment assignment and prescribing the course of action that maximizes a given objective.

2.1.2 Multitreatment model

A binary uplift model can be extended to applications where the interest lies in evaluating the ITE of a diverse set of treatments. This corresponds to real-world scenarios where decision-makers must choose between multiple treatment alternatives in order to optimize the performance of treatments and to personalize the experience of users. Examples of such decisions are identifying the product design, the communication channel or promotion that is the most appealing to a customer, the most favorable medical treatment option for a patient, or selecting the assistance program with the largest benefit for a vulnerable individual.

Multitreatment uplift modeling (MTUM) requires a set $T = \{0, 1, \ldots, k\}$ of mutually exclusive treatments, a collection of observed pretreatment characteristics X, and a binary outcome variable $Y \in \{0, 1\}$. Similarly to binary treatment uplift models, the aim of MTUM is to find the treatment whose effect on the outcome is the most favorable from a larger set of treatment alternatives. The machine learning task consists in estimating the conditional probabilities of a positive outcome for each individual, given the pretreatment characteristics and the exposure to the treatments. Later, these estimates are contrasted to identify the treatment whose ITE is the largest.

MTUM takes into account two different contrasts: multiple treatment groups without a control group and multiple treatment groups with a control group. The former consists of $\binom{k}{2}$ simultaneous pairwise comparisons and seeks to identify the best rank order for each individual. The latter compares each treatment alternative against a control group and aims to determine the optimal action for each individual (Zhao and Harinen 2019). To maintain similarity with the current MTUM literature, this study applies to scenarios with multiple treatment groups, including a control group. For example, a government agency wants to send personalized letters to motivate individuals to vote by: (1) sending a letter with the message “Do your civic duty” (treatment 1), (2) sending a letter with the message “You are being studied” (treatment 2), or not sending a letter at all (control group). The goal of MTUM is then to identify whether a letter should be sent for each individual and, if so, which type of message it should contain.

Formally, the optimal treatment ($\pi^*_{i,k}$) for individual i is the treatment for which the uplift $\hat{\tau}_{i,k}$ is the largest,

$$\pi^*_{i,k} = argmax(\hat{\tau}_{i,1}, \ldots, \hat{\tau}_{i,k}). \tag{2}$$

This is obtained after estimating the differences in the probabilities of a positive outcome between the treatments under evaluation and the control group ($T = 0$) at the individual level, as shown in Eq. 1.

2.2 Survey of multitreatment uplift modeling approaches

The MTUM literature is still limited. This study categorizes the different MTUM approaches according to the classification proposed by Devriendt et al. (2018) for binary uplift models. The authors distinguish two main methods to obtain uplift estimates: the data preprocessing approach and the data processing approach. The former learns an uplift model by means of conventional machine learning algorithms by redefining the original outcome variable or by modifying the input space before training. The data processing approach comprises methods wherein standard machine learning algorithms are trained separately, or their internal structures are adapted to the multitreatment case. Table 1 provides an overview of the modeling strategies that are surveyed in this study. In particular, the naive uplift approach and the multitreatment modified outcome approach are our contributions to the current uplift literature. These methods are introduced in Sect. 2.3.

The dummy and interactions approach (DIA) is the only data preprocessing approach that has been proposed within the multitreatment uplift literature. This method extends the input space by adding treatment indicators encoded as dummies $D = \{0, \ldots, k\}$ and interaction terms. The latter capture the interplay between the dummies and the pretreatment characteristics. Uplift is then modeled by means of any machine learning algorithm that receives as input the pretreatment characteristics X, the dummy variables D, and the interaction terms $D \times X$, so that $P(Y = 1|X, do(T)) = f(X, D, D \times X)$.

Lo (2002) and Tian et al. (2014) implement the DIA for binary treatment uplift models and Chen et al. (2015) for the MTUM case. The Personalized Revenue Maximization (PRM) algorithm proposed by the latter authors is particularly discussed in the context of customized pricing and personalized assortment optimization. The inputs the algorithm uses are the vector of individual characteristics, the assigned treatment (e.g., price offered), the interaction terms and the outcomes. The optimization problem lies in minimizing the gap between the predicted expected revenue according to the optimal treatment assignment and the expected revenue obtained with complete knowledge of the parameters that specified customer behavior. The results of the customized pricing for airline priority seating show that the SMA using a random forest algorithm slightly outperforms the PRM method for all data sizes.

The DIA is a simple approach, since conventional algorithms do not need to be modified and the outcome variable does not necessarily have to be binary. However, the enlargement of the input space can cause overfitting and multicollinearity problems when the amount of interactions is considerably large (Kane et al. 2014).

Most studies addressing the MTUM case can be categorized within the data processing approach. This implies that the uplift is modeled in either an indirect or a direct way. Modeling uplift indirectly corresponds to a strategy in which training cases are grouped according to the treatment that they received. Later, a model is trained for each group. By contrast, a direct uplift estimation trains a single model by employing multitreatment uplift algorithms.

Estimating uplift indirectly is also known as the separate model approach (SMA). This is the baseline technique and was initially proposed to train binary uplift models. Later, it was extended to multitreatment applications due to its simplicity. It employs standard machine learning algorithms to train separate predictive models for each treatment group. Afterwards, the models are used to compute the $\hat{P}(Y = 1|X, do(T = k))$ for each test case, so that the optimal treatment is the one for which the largest difference is obtained (see Eq. 2).

Lo and Pachamanova (2015) demonstrate the estimation of the ITE in a multitreatment scenario by applying the SMA, due to its simplicity and general acceptance as a baseline method. The authors present a framework that formulates the MTUM task as an optimization problem and considers the level of risk aversion of the modeler. An application is presented in which separate logistic regressions are trained to estimate the $\hat{\tau}_{i,k}$. Later, these estimates are used as input variables to determine the cluster level uplift of each treatment. Treatments are then allocated by considering the estimated uplift scores and the variability among estimates.

Table 1 A summary of MTUM approaches

Current methods		
Previous studies	Machine learning technique	Data sets
Data preprocessing approach		
Direct estimation: dummy and interactions approach (DIA)		
A single predictive model with a modified input space is trained. In addition to the pretreatment characteristics, dummies indicating the exposure to treatments and interaction terms are added.		
Chen et al. (2015)	Logistic regression	Airline priority seating (private)
Data processing approach		
Indirect estimation: separate model approach (SMA)		
A predictive model is trained for each treatment group using the pretreatment characteristics as predictors and the outcome variable as target. Then, each model is used to predict the conditional probabilities $\hat{P}(Y = 1 \| X, do(T = k))$ for each test individual, so that the $\hat{\tau}_{i,k}$ can be estimated to identify the optimal treatment $\pi^*_{i,k}$.		
Lo and Pachamanova (2015)	Logistic regression	MineThatData (public)
Direct estimation: adapted algorithms		
An uplift model is trained with a machine learning technique that is specially adapted to the multitreatment setting.		
Rzepakowski and Jaroszewicz (2012)	Decision tree	`splice` in UCI repository (public)
Guelman (2015)	K-nearest-neighbor (CKNN)	–
Zhao et al. (2017b)	Random forest (CTS)	Synthetic data (private) & Seat reservation data (private)
Zhao et al. (2017a)	Random forest (UCTS)	Synthetic data (private)
Li et al. (2018)	Reinforcement learning (Rlift)	Synthetic data (private) & Marketing campaign (private)
Sawant et al. (2018)	Reinforcement learning	Amazon fashion marketing (private)
Zhao and Harinen (2019)	Meta-learners *(X-Learner and R-Learner)*	Synthetic data (public) & Promotion campaign (private)

Table 1 continued

Proposed methods
Data preprocessing approach: multitreatment modified outcome approach (MMOA)
The modified outcome variable approach (MOVA) proposed by Kane et al. (2014) and Lai (2006) for binary uplift models is generalized to MTUM.
Data processing approach: naive uplift approach (NUA)
Separate binary uplift models directly estimate the uplift between each treatment group and the control group.

There are two main disadvantages in applying the SMA. First, training several models increases computational costs. Second, the modeling objective of the different predictive models does not correspond to estimating the uplift. Each model learns the likelihood of a positive outcome, rather than the what-if difference in behavior (Radcliffe and Surry 2011). Nonetheless, Rudaś and Jaroszewicz (2018) demonstrate that the SMA performs competitively for uplift regression when the sample size is sufficiently large and highly correlated variables are removed.

Modified machine learning algorithms are proposed in the MTUM literature to improve the accuracy of the uplift estimate and offset the main drawbacks of the methods mentioned above. In this regard, Alemi et al. (2009) and Guelman (2015) proposed to adapt the K-nearest neighbor classifier (Cover and Hart 1967) to infer the optimal treatment based on the treatment that has worked the best for individuals who are similar to the test case. In personalized medicine, the Sequential K-Nearest Neighbor Analysis (SKNN) (Alemi et al. 2009) sequentially examines the K most similar individuals until the success or failure of the treatment is determined to be statistically significant. Likewise, the Causal K-Nearest-Neighbor (CKNN) approach (Guelman 2015) predicts the optimal treatment for a given individual by weighting the evidence of similar individuals more strongly. This approach is computationally expensive, since all of the training data must be stored to score test cases.

The splitting criterion and pruning method of the most common decision tree classifiers, such as the classification and regression trees (CART) (Leo et al. 1984), chi-square automatic interaction detection (CHAID) (Kass 1980), and C4.5 (Quinlan 1993), can be adjusted for MTUM. Rzepakowski and Jaroszewicz (2012) propose a splitting criterion that compares the probability distributions of treatment groups by using divergence measures from the information theory literature: the Kullback–Leibler (KL), the squared Euclidean distance (ED) and the chi-squared divergence. Pruning is based on the maximum class probability approach. The measure of divergence for multiple distributions allows the modeler to determine the relative importance assigned to the dissimilarity between all of the treatments and the control, and between the treatments themselves. The relative importance of the treatments is also considered.

Adjustments to the splitting criterion and termination rules of the random forest algorithm (Breiman 2001) are suggested to counteract the instability of a single decision tree. The Contextual Treatment Selection (CTS) algorithm (Zhao et al. 2017b) is a forest of randomized trees whose splitting criterion directly maximizes a measure of performance: the expected response. This ensures that the split with the largest increase in expectation is performed at each step. The Unbiased Contextual Treatment Selection (UCTS) algorithm (Zhao et al. 2017a) eliminates the estimation bias present in the CTS by randomly splitting the training set into an approximation set that generates the tree structure and an estimation set that estimates the leaf response. According to the authors' findings, the UCTS proves to be more competitive in terms of performance for some data sets compared to the CTS.

Li et al. (2018) propose a reinforcement learning application that relates MTUM with an offline contextual bandit problem. Since the objective of offline contextual bandits is to maximize the expected response to an action instead of maximizing the expected uplift, the authors formulate the uplift modeling task as a Markov Decision

Process (MDP). This is solved by using Sutton et al. (2000)'s neuralized policy gradient method. In addition, Sawant et al. (2018) use counterfactual matching as part of the data collection and incorporate contextual Bayesian multiarmed bandits to optimize causal treatment effects.

Last, the cost difference of applying treatments in MTUM is incorporated by Zhao and Harinen (2019). The authors adapt the X-Learner (Künzel et al. 2019) and the R-Learner (Nie and Wager 2017) meta-learners to the multitreatment uplift setting and propose a net-value optimization framework to consider the cost of each treatment.

2.3 Proposed methods

This section presents the two methods proposed in this article to estimate uplift in multitreatment applications. First, the MTUM task is transformed into a multiclass prediction problem that can be solved by conventional machine learning algorithms. This is a generalization of the Modified Outcome Variable Approach (MOVA), a conventional method in the binary uplift modeling setting. It considers the information in the data set about the treatment allocated to individuals and their corresponding observed outcome in order to create a new outcome variable. The second method builds separate uplift models employing modified binary uplift algorithms. Each model contrasts the $T = k$ treatment group against the control group ($T = 0$).

2.3.1 Multitreatment modified outcome approach (MMOA)

The MOVA is proposed by Kane et al. (2014) and Lai (2006) for the binary treatment case. The aim is to use any standard multiclass classification algorithm to obtain the required predictions to compute the ITE from a single model. Since a data set suitable for uplift modeling contains information regarding the treatments received by individuals and their observed outcomes, we can segment cases into different categories. These will be the labels of the new outcome variable. For example, the new outcome variable consists of four segments of individuals in a binary treatment case: treated responders ($R_{T=1}$), control nonresponders ($NR_{T=0}$), treated nonresponders ($NR_{T=1}$) and control responders ($R_{T=0}$). The multiclass algorithm outputs the likelihood of each test case belonging to each of these categories. The intuition behind this approach is that the ITE ($\hat{\tau}_{i,1}$) can be computed as follows:

$$\hat{\tau}_{i,1} = \left(\frac{\hat{P}(R_{T=1}|x_i)}{P_{T=1}} + \frac{\hat{P}(NR_{T=0}|x_i)}{P_{T=0}} \right) - \left(\frac{\hat{P}(NR_{T=1}|x_i)}{P_{T=1}} + \frac{\hat{P}(R_{T=0}|x_i)}{P_{T=0}} \right). \quad (3)$$

Equation 3 is analogous to Eq. 1, since the left side indicates the individual's likelihood to have a favorable outcome due to the treatment. Depending on its magnitude, it determines whether an individual should be targeted. Additionally, the prior probabilities of the treatments ($P_{T=k}$) are incorporated as weights to counteract the imbalance of treatment groups.

The extension to the multitreatment case is straightforward, since Eq. 3 can be generalized to calculate the $\hat{\tau}_{i,k}$ for any number of treatments. In the case of two

Table 2 Modified outcome variable for three treatment groups

Treatment group (T)	Observed outcome (Y)	Modified outcome
$T = 0$	1	$R_{T=0}$
$T = 0$	0	$NR_{T=0}$
$T = 1$	1	$R_{T=1}$
$T = 1$	0	$NR_{T=1}$
$T = 2$	1	$R_{T=2}$
$T = 2$	0	$NR_{T=2}$

treatment groups and one control group $T = \{0, 1, 2\}$, the new labels of the outcome variable are shown in the third column of Table 2.

A multiclass probabilistic model is trained to later predict for each individual the probabilities of responding positively ($Y = 1$) and negatively ($Y = 0$) to every treatment alternative. The predicted optimal treatment for the i individual is $\pi^*_{i,k} = argmax(\hat{\tau}_{i,1}, \hat{\tau}_{i,2})$. The $\hat{\tau}_{i,1}$ and $\hat{\tau}_{i,2}$ are calculated as follows:

$$\hat{\tau}_{i,1} = \left(\frac{\hat{P}(R_{T=1}|x_i)}{P_{T=1}} + \frac{\hat{P}(NR_{T=0}|x_i)}{P_{T=0}}\right) - \left(\frac{\hat{P}(NR_{T=1}|x_i)}{P_{T=1}} + \frac{\hat{P}(R_{T=0}|x_i)}{P_{T=0}}\right),$$
$$\hat{\tau}_{i,2} = \left(\frac{\hat{P}(R_{T=2}|x_i)}{P_{T=2}} + \frac{\hat{P}(NR_{T=0}|x_i)}{P_{T=0}}\right) - \left(\frac{\hat{P}(NR_{T=2}|x_i)}{P_{T=2}} + \frac{\hat{P}(R_{T=0}|x_i)}{P_{T=0}}\right).$$

The advantage of the MMOA is that the uplift estimation is reduced to a multiclass classification problem, where a wide variety of classifiers can be used. Additionally, this setting allows the implementation of models that are easier to interpret. For instance, favoring simple models facilitates the observation of the influence that the pretreatment characteristics exert on the uplift estimation. Nevertheless, the MMOA can become inefficient when the amount of treatments rises exponentially.

2.3.2 Naive uplift approach (NUA)

The binary treatment uplift models presented in the survey by Devriendt et al. (2018) can be extended to indirectly predict the optimal treatments in the MTUM scenario. The NUA is a data processing method in which uplift is estimated indirectly. It trains different binary treatment uplift models separately. Each binary treatment model contrasts a treatment group against the control group and outputs the probabilities that are needed to predict the best treatment for the i individual ($\pi^*_{i,k}$).

In the example with two treatment groups and a control group, we build two separate binary uplift models. One model directly estimates the individual-level probabilities of a positive outcome by contrasting $T = 1$ (treatment 1 group) and $T = 0$ (control group), whereas a second model does the same by comparing $T = 2$ (treatment 2 group) and $T = 0$ (control group). Then, test cases are scored using the two models, and the best treatment is predicted as specified in Eq. 2.

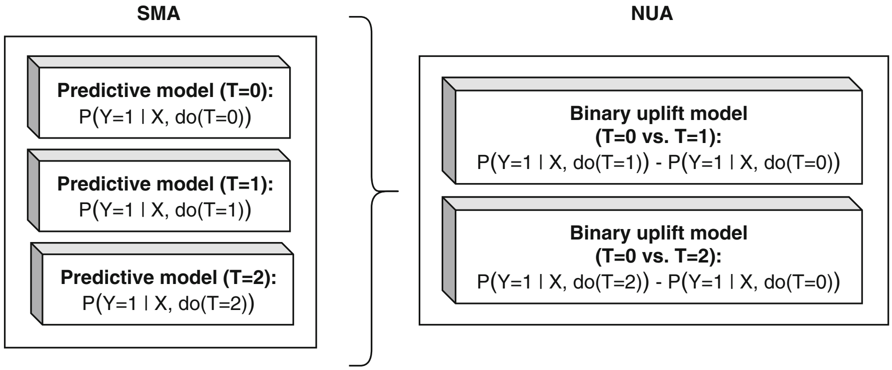

Fig. 1 Comparison of the training schemes of the SMA and the NUA when three treatment groups are considered. Whereas three separate conventional classifiers are trained under the SMA, the NUA estimates the uplift by employing two binary uplift models

The difference between the NUA and the SMA lies in the number of models to train and the algorithms that can be used. The individual-level uplift is calculated by SMA based on the predictions of the models built on each treatment group, a task that can be performed by any standard classifier. In contrast, the NUA takes advantage of existing binary uplift modeling machine learning algorithms to train $k-1$ models, which directly compare the treatments with the control group.

Figure 1 illustrates the difference between the two methodologies in the case of three treatment groups and a binary outcome variable, where $Y = 1$ represents a positive outcome.

3 Evaluation metrics

In predictive analytics, the model with the lowest prediction error (e.g., error rate or loss function) is typically considered to be the best performing model. In this regard, error refers to the lack of fit between the predicted outcome value and the true outcome value for an individual in the holdout set. However, in the uplift modeling case, such an approach is infeasible because the true effect of the treatments is not observed, as a consequence of the fundamental problem of causal inference (Holland 1986). This makes direct test set evaluation impossible: hence, an error cannot be computed. Suggestions to tackle this problem are proposed in the uplift literature, but none have proven to be optimal. One such suggested approach creates groups of test set individuals similarly ranked by the model and extracts the uplift estimate from their respective true outcomes and observed treatments. A second method computes the expected response given the optimal treatment suggested by the uplift model.

Lo and Pachamanova (2015) and Chen et al. (2015) evaluate uplift models in accordance with the optimization objective of their study. The first authors present a framework that formulates the MTUM task as an optimization problem and considers the level of risk aversion of the modeler. The $\hat{\tau}_{i,k}$ estimates are used as input variables for cluster analysis to determine the cluster level uplift for each treatment.

The risk/return trade-off is summarized using the $efficient\ frontier$ graph. As such, the modeler selects the treatment assignment according to her/his risk aversion profile. The second study determines model performance based on the expected revenue that can be achieved by targeting individuals with the suggested optimal treatment. Li et al. (2018) propose the Uplift Modeling General Metric (UMG) and the Self-Normalized Uplift Modeling General Metric (SN-UMG). Their objective is to find a treatment rule that maximizes the expected uplift response under a specific treatment policy by comparing the expected treatment responses with the expected natural responses. The difference between the UMG and the SN-UMG is that the latter reduces the variance by adding standardized weights to the UMG.

In the remainder of this section, we further discuss the uplift evaluation techniques used to compare the results of our experiments.

3.1 Conventional uplift metrics

A conventional uplift methodology assumes that test cases which are similarly scored by a model behave in a similar manner. The performance of a model is then assessed at the level of groups of individuals. First, the estimated uplift score $\hat{\tau}_{i,k}$ of the optimal treatment $\pi^*_{i,k}$ is used to rank each individual in the test set in descending order. Later, groups of test cases are formed from the resulting split of the test set in various bins (e.g., deciles). Given that we observe the assigned treatment and the corresponding outcome for each individual, the response rate for each treatment can be calculated within the group. The uplift is then estimated within each group as the difference in response rates. The intuition behind this approach is that a model with an outstanding performance places potential responders at the top of the ranking. Therefore, larger uplifts are expected in the first groups than in the bottom groups (Hansotia and Rukstales 2002). The advantage of this method is that it provides a comprehensive view of model performance and facilitates decision-making (Moro et al. 2011). However, this technique can be misleading when there are large differences in the pretreatment characteristics of test individuals or large imbalances in the size of the treatment groups. In Sect. 4, propensity score matching is proposed to offset these concerns.

Although the evaluation of MTUM following this approach poses some challenges, examples of its implementation are found in Sawant et al. (2018) and Zhao and Harinen (2019). The main difficulty lies in the fact that individuals in the test set are exposed to treatments at random and their predicted optimal treatment does not necessarily match the treatment that is observed. Imai et al. (2013) propose as an alternative the assignment of a pay-off to the test cases whose responses are favorable to the treatments recommended by the model. However, such practice can generate biases in the uplift estimation. For this reason, we adopt the solution suggested by Chen et al. (2015) and Rzepakowski and Jaroszewicz (2012), in which the mismatched test cases are not considered for the evaluation. This naturally leads to a considerable loss of data but assures an unbiased assessment.

The performance of an uplift model can be visualized by means of an uplift curve (Rzepakowski and Jaroszewicz 2012). Given the ranking of individuals, this curve illustrates the cumulative difference in response rate by applying the optimal treatment to p percent of test cases relative to the control group. The x-axis displays the percentage of targeted individuals, whereas the y-axis shows the cumulative difference between the response rates of the predicted optimal treatments and the response rates of the control group. The overall effect of the treatments when all individuals are targeted (i.e., $p = 100$ percent) is implicitly observed on the plot. A highly right-skewed uplift curve is desirable, since it indicates that the likely responders are primarily grouped in the top segments. An uplift curve is comparable to the lift curves in standard classification models, since it results from subtracting the estimated lift curve of the group with the optimal treatments from the estimated lift curve of the standard treatment group. In addition, another straight line is drawn within the two extremes of the uplift curve to represent the net incremental gains of randomly intervening individuals. This line serves as a baseline to graphically observe how well a model outperforms the action of targeting subjects at random.

Since the uplift curve is a subtraction of lift curves, this facilitates the estimation of a modified metric which is conceptually similar to the Gini coefficient (Kuusisto et al. 2014). The Qini metric (Radcliffe 2007), also known as the Area Under the Uplift Curve (AUUC) (Rzepakowski and Jaroszewicz 2010), is a standard tool to compare the performance among uplift models. This is calculated as the area between the uplift curve and the random model line. The greater this metric, the larger the incremental effects of the predicted optimal treatments.

3.2 Expected response

The expected response is proposed by Zhao et al. (2017b) as an alternative to evaluate the performance of uplift models. This method is generalized for applications where multiple treatments are considered, as well as for different types of outcome variables. In addition, it addresses potential biases that may result when the sizes of the treatment groups are highly imbalanced.

The expected response method calculates a new variable Z that depends on the observed treatment in the test set, the predicted optimal treatment by the uplift model, the prior probabilities of the treatments $P_{T=k}$, and the observed outcome Y. The computation also considers the Iverson bracket $\mathbb{I}(\cdot)$, which is equal to one if the predicted optimal treatment matches the observed treatment, and zero otherwise. Formally, the individual expected response is as follows:

$$z_i = \sum_{k=1}^{N} \frac{y_i}{P_{T=k}} \mathbb{I}\{h(x_i) = k\}\mathbb{I}\{T = k\}.$$

When the predicted optimal treatment equals the observed treatment, z_i represents the observed outcome scaled by the prior probability of being exposed to the treatment. The expected response of a multitreatment uplift model is then calculated as follows:

$$\bar{z} = \frac{1}{N} \sum_{i=1}^{N} z_i. \tag{4}$$

The modified uplift curve illustrates the performance of a multitreatment uplift model in terms of the expected response (Zhao et al. 2017b). This curve is a plot of the cumulative expected response as a function of the percentage of test set cases that are targeted according to model suggestions. Similarly to the conventional uplift evaluation, test individuals are ranked in descending order according to their predicted uplift scores, and $\bar{z}$ is calculated for a given p percent.

4 Experimental setup

The experimental evaluation contrasts the performances of a subset of the above presented MTUM approaches with respect to eight data sets. First, we provide an overview of the main characteristics of the data sets. Later, we describe the data preprocessing and partitioning strategy, along with the MTUM techniques considered for the experiments. At the end, the statistical tests and their implementation are discussed.

4.1 Data sets

Customizing treatment allocation is a main concern among decision-makers in different domains. This study evaluates uplift models with respect to eight multidisciplinary data sets. Table 3 summarizes the most relevant information in relation to the data sets. Because some data sets are not specifically designed to estimate individual treatment effects, the treatment groups are formed according to the observed values of a specific decision variable (see Rzepakowski and Jaroszewicz (2012)). To assure that the uplift estimate is unbiased, we assess the balance of pretreatment characteristics among treatment groups before training. When imbalance is detected, we implement propensity score matching as proposed by Guelman (2015). This technique is further discussed in the next subsection. Overall, aside from profile, sociodemographic or transactional information, each data set also contains a treatment indicator encoded as a categorical variable with K possible treatments, along with a binary outcome variable. The following data sets are included in the experiments:

- The *Hillstrom* direct marketing campaign data set (Hillstrom 2018) comprises a sample of 64.000 individuals. Three treatment groups are identified. Some customers receive an e-mail with men's merchandise, a second group is targeted with an e-mail corresponding to women's merchandise, and a last segment is not contacted. Success is considered when a customer visits the website within two weeks after receiving the e-mail.
- The *Gerber* data set (Gerber et al. 2008) relates to the study of the political behavior of voters. The aim is to analyze whether social pressure increases turnout from a sample of 180.002 households. Direct mailings were randomly sent 11 days before the August 2006 primary election. The households that received either the "Self" message or the "Neighbors" message are the treated groups to evaluate, whereas those who were targeted with the "Civic duty" message represent the control group. The outcome variable is positive if a vote was given in the elections.

- The *Bladder* data set (Therneau 2015) contains information regarding recurrence of bladder cancer for three treatment groups: 1) pyridoxine, 2) thiotepa, and 3) placebo. As in Sołtys et al. (2015), patients who had remaining cancer, or at least one recurrence, are classified as negative cases.
- The colon data set (Therneau 2015) includes data of chemotherapy trials against colon cancer. A low-toxicity medication, Levamisole, was administered to some patients, whereas a combination of Levamisole with the moderately toxic 5-FU chemotherapy agent was received by another subsample. The control treatment group corresponds to the nontreated patients. Following the setup proposed by Sołtys et al. (2015), two outcome variables can be extracted: 1) recurrence or death (Colon1) and 2) death (Colon2). The two data sets slightly differ in the way that the predictor variable *time* is processed. For the Colon1 data set, this variable is split into two factors: 1) the number of days until the recurrence event and 2) the number of days until the death event. In the Colon2 data set, *time* refers only to the number of days until death, since there is no recurrence.
- The *AOD* data set corresponds to alcohol and drug usage (McCaffrey et al. 2013). In this subset of 600 observations, three treatment groups are identified: "community," "metcbt5" and "scy." We assigned individuals within the former category to the control group. Given that the outcome variable is continuous, we apply binary encoding by assuming that a positive case is an individual whose substance use frequency declines by the 12th month after the treatment is applied. An important observation is that only 5 out of the 23 original pretreatment variables are available in this subset. Therefore, information on demography, substance use, criminal activities, mental health function and environmental risk is mostly absent.
- The *Bank Marketing* data set (Moro et al. 2014) is publicly available in the UCI repository. This set contains information regarding a direct marketing campaign conducted by a commercial bank. To obtain a multitreatment set, the categorical variable "contact" is chosen as the decision variable to determine the different treatment groups. Depending on the type of contact communication, individuals are assigned to either the "cellular" group or "telephone" group. The "unknowns" are the control group. The outcome variable is positive if a customer decides to open a term deposit with the institution.
- The *Turnover* data set provided by a private Belgian organization comprises information regarding retention strategies aiming to reduce voluntary turnover. A subset of the 1.951 white collar employees is targeted with two retention campaigns: "recognition" and "flexibility." The remaining group is not treated, and hence is classified as control. A positive case is represented by an employee who does not voluntarily leave the company the year after the strategies are deployed.

4.2 Data preprocessing and partitioning

Estimating the ITE of multiple treatments conveys some degree of uncertainty because an individual can only be assigned to one treatment group. Hence, the outcomes under the remaining alternatives are never observed in reality. If K represents the amount

Table 3 Multitreatment data sets

Dataset	Source	Domain	Channel	Response	No. of variables	Groups		
						Treatment 1	Treatment 2	Standard treatment
Hillstrom	Hillstrom (2018)	Marketing	E-mail	Visit	18	WomensEmail (21.387) 4.52%*	MensEmail (21.307) 7.66%*	Control (21.306)
Gerber	Gerber et al. (2008)	Political behavior	Mail	Vote	11	Self (38.218) 6.34%*	Neighbors (38.201) 3.06%*	Civic duty (38.218)
Bladder	Therneau (2015)	Clinical trial	Medication	No recurrence	8	Pyridoxine (85) −5.16%*	Thiotepa (81) −9.86%*	Placebo (128)
Colon 1	Therneau (2015)	Clinical trial	Medication	Recurrence or death	13	Levamisole (310) −0.08%*	Levamisole & 5FU (304) −17.08%*	Observation (315)
Colon 2	Therneau (2015)	Clinical trial	Medication	Death	12	Levamisole (310) −0.08%*	Levamisole & 5FU (304) −17.08%*	Observation (315)
Bank	Moro et al. (2014)	Marketing	Call	Subscribe	16	Cellular (29.285) 10.85%*	Telephone (2.906) 9.35%*	Unknown (13.020)
Turnover	Private organization	Human resource	Retention	No turner	24	Recognition (363) 1.84%*	Flexibility (491) −1.52%*	Control (690)
AOD	McCaffrey et al. (2013)	Public policy	Program	Reduce use	5	Metcbt5 (200) −6%*	Scy (200) −4%*	Community (200)

Numbers within brackets refer to the size of the respective treatment group, while asterisks denote the observed uplift in the data set of total treatment states, there are $K - 1$ unknown outcomes that correspond to the different counterfactual scenarios.

In a randomized control trial, the counterfactuals can be imputed from the observed outcomes of “similar” individuals who were exposed to the alternative treatments (Rosenbaum and Rubin 1983). In this context, similarity indicates that there are no considerable differences in the pretreatment characteristics among the treatment groups. This assures that the only cause of the behavioral change of an individual is the exposure to a particular treatment, all other factors being equal. However, in situations where the selection rule to allocate treatments is unknown or is not random (i.e., observational study), the estimation of treatment effects can be biased due to the heterogeneity of the treatment groups.

The majority of uplift models implicitly assume that treatments are allocated to individuals at random: as a result, the balance of pretreatment characteristics is often not validated. We believe that an unbiased estimation of treatment effects at the individual level demands verification and, if needed, corrective action. As such, the influence of the pretreatment characteristics in the assignment of treatments is removed. In this study, we implement the propensity score matching (PSM) (Rosenbaum and Rubin 1983) method as an attempt to form a quasi-randomized experiment to control for any selection bias that may affect the uplift estimate (Lopez et al. 2017).

Matching is convenient for estimation of treatment effects, since, in principle, it guarantees a homogeneous sample of individuals in terms of their observed pretreatment characteristics. Diverse matching strategies are proposed in the literature (see Morgan and Winship (2015) for an overview of matching techniques). The main differences among the techniques lie in how the sets of “similar” individuals are formed. For example, exact matching consists in grouping individuals whose only difference is the allocated treatment, whereas PSM joins the information of all pretreatment variables into a “score” that is later used to perform the matching.

PSM consists of estimating the probability for each individual of being treated as a function of the pretreatment characteristics, which is known as the propensity score (PS), $PS_{i,k} = \hat{P}(T = k|x_i)$. Later, individuals with similar PS values are matched, and an estimate of the treatment effect is computed based on the differences in their observed outcomes. In sum, this technique aims to balance the observed pretreatment variables X among the treated groups T to obtain an unbiased estimate of the causal effect of T on Y. This assures that the remaining differences in the observed outcomes among treatment groups can be attributed solely to the effects of the treatments (Morgan and Winship 2015).

The PSM approach creates sets of individuals who are “similar” to some degree with respect to the observable pretreatment variables. As such, it provides transparency with respect to the mechanism of treatment assignment. Most common techniques for PSM employ nearest neighbor algorithms, kernel matching or one-to-one matching. They differ in the selection of distance measure, as well as in the amount of cases to group. Diamond and Sekhon (2013) proposed a genetic search algorithm to assure that optimal balance is achieved. Optimal matching has proven to be efficient, since it minimizes the distance between matched individuals and works well when the control group is smaller than the other treatment groups.

One advantage of PSM is that it achieves good performance for small data sets and counteracts the difficulties of applying matching to high-dimensional data sets. One limitation, however, is that unmatched cases are excluded from the analysis, leading to an important loss of information that can hamper the generalization of the findings. In addition, given the differences in pretreatment characteristics among treated groups, biases may not be completely solved, since the estimation of the PS is highly dependent on the correct specification of the set of observable characteristics that account for the systematic differences between these groups.

Regarding the partitioning of the data sets, a cross-validation strategy is used for evaluating model performance. This decreases the risk of overfitting and assures the generalization of model estimates. We partition each data set into five folds of approximately equal size, without overlap. In addition, stratification is applied with respect to the treatment groups to preserve the observed treatment effects to remain as similar as possible among the folds. Models are then fitted by performing multiple rounds, in which one fold is left out for testing and the remaining folds are considered as the training set. Later, the final models are applied to individuals in the test set, and performance is evaluated. The results of each round are averaged to obtain the overall performance.

4.3 Uplift modeling techniques

Table 4 provides an overview of the MTUM techniques whose performances are evaluated in this study. These methods are a selection of data preprocessing and data processing approaches.

Among the techniques are well-established standard algorithms, such as the Logistic regression and the Random forest. Moreover, four modified algorithms that estimate the uplift directly in multitreatment applications are considered: Causal K-nearest neighbor (Guelman 2015), CTS random forest (Zhao et al. 2017b), ED random forest (Rzepakowski and Jaroszewicz 2012), X-Learner random forest and R-Learner random forest (Zhao and Harinen 2019). For the NUA and the MMOA, we use the binary Uplift random forest developed by (Guelman 2014) and the Multinomial logistic regression, respectively. These latter approaches complement the existing set of methods in the MTUM literature.

When a variable selection procedure is not embedded within an algorithm, the generalized linear model with a stepwise variable selection procedure is deployed. This wrapper method removes some of the pretreatment variables until it finds the optimal combination that maximizes the performance of the model as guided by the Akaike Information Criterion (AIC). At the end of the iterations, a vector with the final variables is returned. Later, these variables are used to fit the models. An optimal sample of pretreatment characteristics decreases not only the computational time but also the complexity of the models. A more parsimonious and interpretable model can then be achieved, with potential gains in model performance and stability (Kuhn and Johnson 2013).

Table 4 MTUM techniques considered in the benchmarking study

Method	Approach	Modeling technique	Implementation
Current	Data preprocessing	DIA	
		Logistic regression (*DIALR*)	`train`, `method=“glmStepAIC”`
		Random forest (*DIARF*)	`train`, `method=“rf”`
	Data processing	Indirect estimation	
		SMA	
		Logistic regression (*SMALR*)	`train`, `method=“glmStepAIC”`
		Random forest (*SMARF*)	`train`, `method=“rf”`
		Direct estimation	
		Adapted algorithms	
		Causal K-nearest neighbor (*CKNN*)	`uplift`, `upliftKNN`
		CTS random forest (*CTS*)	`causalML`, `evaluationFunction=“CTS”`
		ED random forest (*ED*)	`causalml`, `evaluationFunction=“ED”`
		XLearner random forest (*XLearner*)	`causalML`, `evaluationFunction=“XLearner”`
		RLearner random forest (*RLearner*)	`causalML`, `evaluationFunction=“RLearner”`
Proposed	Data preprocessing	MMOA	
		Multinomial log-linear (*MMOALR*)	`nnet`, `multinom`
		Random forest (*MMOARF*)	`randomForest`
	Data processing	Indirect estimation	
		NUA	
		Uplift random forest (*NUARF*)	`uplift`, `upliftRF`
		Uplift causal conditional inference forest (*NUACCIF*)	`uplift`, `ccif`

4.4 Statistical test

The results of the benchmarking experiments are contrasted using a statistical test in order to detect whether the observed differences in performance are significantly different. We adopt the procedure documented by Demšar (2006), which performs a nonparametric Friedman test (Friedman 1940) with a corresponding post hoc test. First, we calculate the ranking for a model $j \in J = \{j_1, \ldots, j_9\}$ within each data set. In the case that some results are identical, the final ranking is an average of the ranks that were initially assigned. Second, we calculate the average ranking $\bar{r}_j$ for the j model over the n data sets and estimate the test statistic as follows:

$$\chi_F^2 = \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left(\bar{r}_j - \frac{k+1}{2} \right)^2 .$$

At a level of $\alpha = 0.05$, we are interested in rejecting the null hypothesis stating that there are no significant differences in performance across the data sets. The probability distribution of χ_F^2 is accurately approximated by that of a chi-squared solely when both n and k are sufficiently large, which is fulfilled in this study (i.e., $n = 8$ and $k = 13$). If the p-value $P(chi_{k-1}^2 \geq \chi_F^2)$ indicates that there are statistically significant differences, a post hoc Nemenyi (Nemenyi 1963) test is suggested to compare all of the models to each other.

4.5 Implementation

The prescribed experiments are implemented in R (R Core Team 2017) and Python (Van Rossum and Drake Jr 1995). For the analysis of selection bias, the `RItools` package is used to check the imbalance in pretreatment characteristics among treatment groups. In the case of detecting any imbalance, the `MatchIt` package applies optimal matching based on the propensity scores.

In R, the `caret` package includes the standard Logistic regression and the Random forest algorithms. Furthermore, in this programming language, the `uplift` package (Guelman 2014) incorporates the CKNN (`upliftKNN`), the Uplift random forest (`upliftRF`) and the Uplift causal conditional inference forest (`ccif`). For the setup of the modified outcome techniques, the `randomForest` algorithm (Liaw and Wiener 2002) and the Multinomial log-linear model algorithm (`multinom`) (Ripley and Venables 2011) are chosen. Recent implementations of the CTS, ED, X-Learner and R-Learner algorithms are available in the `causalML` Python package.

Upon publication of this article, we will make the implementation of our experiments publicly available via Github. Our intention is to make the presented results reproducible and verifiable, as well as to stimulate and facilitate further MTUM research.

5 Empirical results

This section presents the assessment of balance of the pretreatment characteristics among treatment groups, and the respective correction by means of PSM. Later, the results of the benchmarking experiments are reported and discussed. The Qini metric and the expected response are used to evaluate model performance. The Friedman test is applied to determine whether the observed differences in performance are significantly different. In the end, the models' average rankings are calculated and visualized.

5.1 Identifying and correcting selection bias

We perform a PSM preprocessing step in the case of detecting any imbalance in the pretreatment characteristics among the treatment groups. The purpose of this corrective action is to decrease the possibility of obtaining a biased uplift estimate. Considering that all data sets in this study consist of two treatments and a control group, balance is assessed in a pairwise fashion, as shown in Table 5. To verify whether there is at least one pretreatment variable for which the two groups are different, we compute a $\chi^2 - test$ that performs the omnibus test proposed by Hansen and Bowers (2009).

Table 5 illustrates the results of the balance assessments and indicates whether matching is performed. The resulted p-values of the initial chi-square tests do not provide evidence of imbalance among the pairs of treatment groups that are part of the Hillstrom, Gerber and AOD data sets. Therefore, matching is not required. However, the test suggests that at least one of the pretreatment variables in the Bladder, Colon1, Colon2 and Turnover data sets is creating an imbalance between the treatment pairs. Given this result, we apply matching. The p-values of the postmatching chi-square test (i.e., final p-value) indicate that the imbalance is considerably reduced. An important relevant remark is that the prior imbalances among the groups of the bank data set are not successfully corrected by the chosen matching strategy. Therefore, in this specific case, the uplift estimates can be biased, given the differences in the pretreatment characteristics of the individuals.

5.2 Assessing model performance: the Qini metric

Table 6 reports the results of the benchmarking study for the Qini metric. We consider two scenarios in which the MTUM technique is used to target the full sample of test cases (Panel A) and the top 10 percent of individuals most likely to respond favorably to the treatments (Panel B). The CKNN algorithm is not implemented in the Hillstrom, Gerber and Bank data sets, given its operational inefficiency for large data sets. The Qini metrics of the best performing models are in bold, and the corresponding standard deviations are within brackets. Overall, in panel A, it is observed that none of the MTUM approaches evaluated in this study outperform the others. Most of the techniques perform well for some data sets, but poorly for others. Nonetheless, in five out of the eight data sets, our proposed approaches perform better than current methods. Among the recent algorithms such as CTS, XLearner and RLearner, only the former slightly excels beyond the performance of our proposed approaches with respect to

Table 5 Balance assessment and indication of matching

Data set	Treatment groups	Imbalance	p-value	Matching	Balance	Final p-value
Hillstrom	WomensEmail vs. control	×	0.73	×	–	–
	MensEmail vs. control	×	0.58	×	–	–
Gerber	Self vs. civic duty	×	0.31	×	–	–
	Neighbors vs. civic duty	×	0.22	×	–	–
Bladder	Pyridoxine vs. placebo	×	0.73	✓	✓	0.99
	Thiotepa vs. placebo	✓	9.12e−5	✓	✓	0.66
Colon 1	Levamisole vs. observation	×	0.91	✓	✓	0.97
	Levamisole & 5FU vs. observation	✓	0.005	✓	✓	0.06
Colon 2	Levamisole vs. observation	×	0.87	✓	✓	0.96
	Levamisole & 5FU vs. observation	✓	0.015	✓	✓	0.09
Bank	Cellular vs. unknown	✓	3.55e−65	×	×	–
	Telephone vs. unknown	✓	1.47e−302	×	×	–
Turnover	Recognition vs. control	✓	3.87e−11	✓	✓	0.9
	Flexibility vs. Control	✓	1.52e−16	✓	✓	0.29
AOD	Metcbt5 vs. community	×	0.60	×	–	–
	Scy vs. community	×	0.76	×	–	–

the Colon2 data set. Moreover, the proposed approaches generally exhibit reduced variability among folds. Their predictions are more stable, and therefore more reliable. The Friedman test is applied to the results of Panel A to corroborate whether there are statistically significant differences among the performances of the different models. The estimated *p*-value for this test is 0.24, which allows us to conclude that there is no proof of a statistically significant difference in performance among techniques.

The Qini metric for the top 10 percent of targeted test cases indicates how well an MTUM technique prioritizes treatment allocation. In practical settings, campaigns have budgetary constraints that limit their scope. Therefore, model performance is assessed within a smaller proportion of test cases. Panel B of Table 6 shows that the Qini metric varies when the targeted population is reduced to the top 10 percent of responders. Under this restriction, MTUM techniques with outstanding performance when targeting the whole population are no longer suitable. For instance, the same MTUM approach can be employed in only three out of the eight data sets when targeting 100 percent and 10 percent of test cases. Generally, there are no significant differences in performance between MTUM techniques (Friedman test *p*-value of 0.37). Every model, without distinction, performs well for some data sets and poorly for others. Furthermore, their predictions become more unstable, as shown by their standard deviations. The bias-variance trade-off is more evident, since performance improvements are made at the expense of decreasing reliability. This is especially observed for small data sets such as Bladder, Colon1, Colon2 and AOD, whose Qini metrics and standard deviations are larger for 10 percent targeting than for 100 percent targeting.

On the other hand, in data sets where the observed overall effect of treatments is negative (e.g., Bladder), MTUM techniques prove to be valuable instruments to improve treatment effectiveness. For example, the treatments considered in the Bladder, Colon1, Colon2 and AOD data sets would exhibit unfavorable effects if target assignment was not customized according to the predictions of the MTUM techniques.

A Qini curve is a useful visualization tool to assess model performance. This curve graphically displays the performance of a model compared to random targeting. Figure 2 shows the Qini curves of the MTUM techniques evaluated with respect to the Hillstrom data set. The results for this data set are exemplary of those of most data sets. The diagonal line represents a random assignment of treatments, whereas the lines in different colors correspond to the different MTUM techniques. The Qini curve of a model with outstanding performance is as far away as possible from the random line curve (in black). Overall, any of the MTUM techniques boosts the effect of the treatments for a particular proportion of targeted individuals. Nonetheless, DIALR, SMALR and ED appear to be more suitable to achieve superior treatment effects when targeting small samples: for instance, when launching campaigns with a high constraint on the number of participants. On the other hand, MMOALR or RLearner can be more appropriate when considering larger exposure groups.

One important remark is that there are slight differences between the Qini plots of binary uplift models and the Qini plot in MTUM. In the latter case, the Qini curves of the different models, including the random targeting line, do not converge at the end (when targeting 100 percent of test cases). We explained in Sect. 3.1 that the treatment given to an individual in the test set does not necessarily match her/his

Table 6 Qini metric

Data set	Hillstrom	Gerber	Bank	Bladder	Colon1	Colon2	Turnover	AOD
A. at 100%								
SMALR	**1.03** (0.25)	1.17 (0.17)	3.08 (0.63)	−10.47(5.02)	−0.56(4.19)	1.27(3.63)	0.88(1.4)	−3.03(4.32)
SMARF	0.49 (0.12)	1.21 (0.27)	4.85(0.24)	−0.24(5.86)	2.84(3.64)	1.95(4.22)	1.1(1.06)	−0.48(8.69)
DIALR	1 (0.24)	1.2 (0.17)	3(0.58)	−10.44(5.07)	−0.56(4.19)	0.97(3.58)	1.36(1.02)	−1.49(4.58)
DIARF	0.54 (0.39)	1.12 (0.42)	4.14(0.19)	1.78(1.11)	**7.52**(3.95)	1.24(3.74)	−0.01(0.95)	−1.51(5.21)
CKNN	–	–	–	0.18(6.96)	−1.36(7.41)	−1.47(4.16)	1.44(1.01)	−0.93(4.82)
NUARF	0.66 (0.34)	1.11 (0.35)	2.95 (0.95)	−12.23(23.65)	1.03(6.71)	3.41(1.71)	0.42(1.25)	−0.9(6.94)
NUACCIF	0.92 (0.15)	1.18 (0.21)	1.49 (0.42)	−38.02(2.63)	−2.02(5.82)	5.16(1.13)	**1.46**(0.51)	−20.92(1.94)
MMOALR	0.95 (0.22)	**1.27** (0.11)	3.66 (0.61)	1.77(1.22)	−1.07(2.6)	5.02(2.63)	−0.52(0.73)	−1.41(6.87)
MMOARF	0.21 (0.14)	1.14 (0.27)	**7.89** (0.36)	**7.23**(1.16)	0.99(1.21)	1.27(3.74)	−1.09(0.59)	**4.14**(4.11)
CTS	0.93 (0.15)	1.16 (0.19)	4.28 (0.52)	−5.01(17.62)	0.79(4.4)	**5.37**(1.74)	−1.33(1.36)	0.39(4.53)
ED	1 (0.18)	1.22 (0.15)	4.26 (0.51)	−21.05(22.38)	−0.26(4.29)	5.36(2.48)	−1.12(1.34)	1.36(3.92)
XLearner	0.87 (0.22)	1.24 (0.35)	3.5 (0.12)	−28.84(3.14)	−9.61(9.64)	0.49(2.6)	−2.12(1.02)	−15.07(4.19)
RLearner	0.97 (0.15)	1.22 (0.25)	2.58 (0.31)	−25.93(1.91)	7.07(4.25)	1.85(1.88)	−0.84(1.12)	−11.33(12.48)
B. at 10%								
SMALR	3.44 (1.1)	1.47 (1.27)	6.97(2.51)	−19.36(13.16)	−21.01(17.56)	1.55(5.64)	0.8(4.33)	−6.21(8.8)
SMARF	2.36 (1.19)	**3.19** (1.13)	16.35(1.38)	−12.28(18.16)	7.9(25.23)	7.11(12.48)	−1.65(8.74)	0.35(8.92)
DIALR	3.59 (0.79)	1.66 (1.21)	6.53(1.95)	−16.65(14.91)	−21.01(17.56)	−0.83(6.96)	2.33(4.72)	**4.84**(10.34)
DIARF	1.86 (0.95)	2.62 (0.84)	14.25(1.07)	−33.85(1.58)	**17.02**(18.08)	4.4(19.19)	0.52(3.17)	−11.8(17.14)
CKNN	–	–	–	1.17(34.07)	3.27(17.71)	−6.73(10.51)	−5.81(8.26)	−10.63(16.47)
NUARF	0.39 (0.87)	2.31 (0.97)	15.69 (1.15)	−9.12(29.22)	−7.5(22.98)	−3.34(17.04)	1.99(3.52)	−21.58(28.12)
NUACCIF	2.1 (0.79)	2.48 (1.61)	9.76 (2.51)	−36.79(0.95)	−4.53(16.23)	**7.81**(22.58)	2.98(2.87)	−10.67(11.31)
MMOALR	3.52 (0.92)	3.09 (0.43)	11.11(3.21)	−10.67(9.66)	−6.04(9.83)	−4.79(9.15)	−2.56(2.58)	−13.52(21.41)

Table 6 continued

Data set	Hillstrom	Gerber	Bank	Bladder	Colon1	Colon2	Turnover	AOD
MMOARF	0 (0.74)	2.36 (1.49)	**17.05** (3.28)	**27.19**(16.01)	5.49(11.67)	−6.54(16.25)	**17.81**(22.48)	2.95(17.98)
CTS	2.6 (1.93)	2.47 (0.83)	19.34 (5.17)	−17.82(24.18)	9.02(10.94)	1.03(22.09)	−5.56(9.03)	−6.77(18.43)
ED	**3.62** (0.99)	2.72 (0.69)	20.48 (5.32)	−29.75(10.69)	4.14(21.69)	−0.21(18.88)	−0.69(3.27)	−9.47(21.69)
XLearner	2.83 (1.69)	1.9 (1.91)	11.85 (2.34)	−6.34(12.68)	−7.62(12.16)	−0.36(20.89)	−27.4(27.38)	1.49(1.51)
RLearner	2.69 (1.35)	1.92 (0.76)	13.91 (1.93)	3.92(0)	7.5(7.88)	−11.35(7.71)	0.64(4.54)	2.55(14.76)

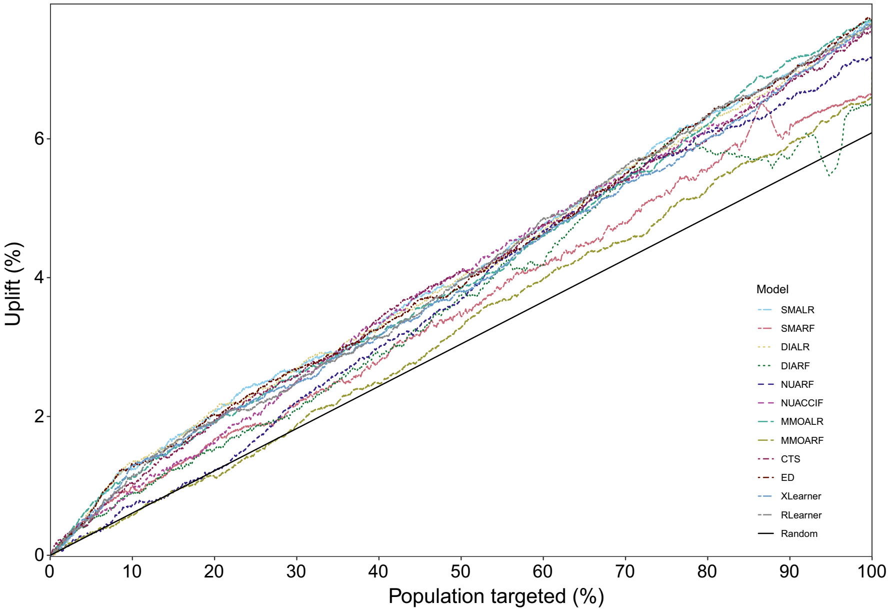

Fig. 2 Qini curves as a function of the targeted population for the Hillstrom data set. The curves correspond to the 12 different experimentally evaluated MTUM approaches, and the straight line is the baseline indicating random targeting

predicted optimal treatment, due to the random assignment of treatments. For this reason, the mismatched test cases are not considered in the evaluation. Therefore, MTUM techniques achieve distinct uplift levels when the full population is targeted.

5.3 Assessing model performance: the expected response

Alternatively to the Qini metric, the expected responses of optimal targeting as predicted by the MTUM techniques are reported in Table 7. Panel A and Panel B show the expected responses of targeting the full test sample and only the top 10 percent of test cases, respectively. The largest expected responses are in bold. In large data sets such as Hillstrom, Gerber and Bank, the expected responses for both exposure segments do not differ significantly across models. However, there are slight differences in the expected responses among the MTUM approaches for small data sets. The CTS, ED, XLearner and RLearner methods are as competitive as the approaches we propose in this study. Conventional techniques, on the other hand, are clearly suboptimal. The p-values of the Friedman test, 0.19 and 0.13, indicate that none of the applied approaches differ significantly in terms of performance when targeting either 100 percent or 10 percent of the test cases, respectively. This is consistent with the results obtained for the Qini metric.

Figure 3 plots the expected responses of the MTUM approaches at different targeting levels for the Hillstrom data set. The horizontal axis indicates the percentage of the population targeted with the predicted optimal treatments, whereas the vertical axis

Table 7 Expected response

Data set	Hillstrom	Gerber	Bank	Bladder	Colon1	Colon2	Turnover	AOD
A. at 100%								
SMALR	**0.18**	**0.38**	0.16	0.79	0.54	0.56	0.99	0.51
SMARF	0.16	0.36	0.15	0.97	0.51	0.51	1.07	0.51
DIALR	**0.18**	**0.38**	0.16	0.79	0.54	0.58	1.01	0.46
DIARF	0.17	0.37	0.15	0.85	0.54	0.55	1.14	0.46
CKNN	–	–	–	0.86	0.49	0.45	0.92	0.48
NUARF	**0.18**	0.37	0.14	0.86	0.50	0.61	1.12	0.49
NUACCIF	**0.18**	**0.38**	**0.17**	0.84	0.44	**0.62**	1	0.53
MMOALR	**0.18**	**0.38**	0.12	1.12	0.47	0.56	1.32	0.53
MMOARF	0.17	0.37	0.12	**1.20**	0.55	0.48	**1.49**	**0.56**
CTS	**0.18**	**0.38**	0.16	0.83	0.48	**0.62**	0.96	0.50
ED	**0.18**	**0.38**	0.14	0.81	0.49	**0.62**	0.89	0.51
XLearner	**0.18**	**0.38**	0.14	0.84	0.55	**0.62**	0.95	0.54
RLearner	**0.18**	0.37	0.15	0.84	**0.57**	0.59	1.15	0.54
B. at 10%								
SMALR	**0.12**	**0.33**	0.08	0.79	0.39	0.44	0.90	0.49
SMARF	0.11	**0.33**	0.09	0.81	0.56	0.45	0.91	0.51
DIALR	**0.12**	**0.33**	0.08	0.81	0.39	0.44	0.90	0.51
DIARF	**0.12**	**0.33**	0.08	0.68	**0.57**	0.47	0.93	0.49
CKNN	–	–	–	0.79	0.53	0.42	0.93	0.51
NUARF	0.11	0.32	0.07	0.84	0.52	0.45	0.95	0.52
NUACCIF	**0.12**	0.32	**0.10**	0.84	0.51	0.45	0.91	0.53
MMOALR	**0.12**	0.32	0.06	0.94	0.52	0.43	0.99	0.51
MMOARF	0.11	0.32	0.07	**1.03**	0.56	0.44	**1.20**	0.53
CTS	**0.12**	0.32	0.03	0.79	0.55	0.44	0.95	**0.54**
ED	**0.12**	**0.33**	0.04	0.80	0.55	0.44	0.97	0.52
XLearner	**0.12**	**0.33**	0.08	0.84	0.54	0.38	0.80	**0.54**
RLearner	**0.12**	0.32	0.09	0.84	0.54	**0.48**	0.98	**0.54**

shows the expected response. As expected, optimal targeting positively influences the effect of treatments. The advantage of this visualization tool is that it supports decision-making in the sense that when confronted with resource constraints, one can select the model that yields the largest expected response for a given percentage of the targeted population. We observe that the ED and RLearner methods generally achieve the highest expected response, regardless of the proportion of targeted test cases.

5.4 Matched test cases and overall ranking of MTUM approaches

A final analysis of the results consists of assessing the matched test cases and contrasting the performance of each model according to the different evaluation metrics.

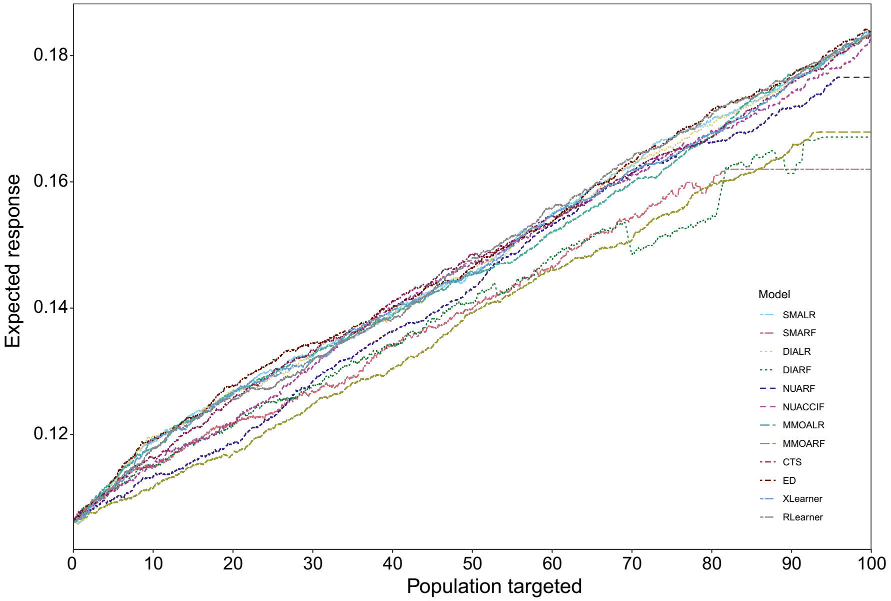

Fig. 3 Expected response as a function of the population targeted for the Hillstrom data set. The curves correspond to the 12 different experimentally evaluated MTUM approaches

We emphasize in Sect. 3 that evaluating the performance of MTUM approaches can be challenging. Particularly, test set cases receive treatments at random, and hence their predicted optimal treatments do not necessarily match their observed treatments. In order to assure a correct interpretation of the findings, the Qini metric and the expected response only consider the test cases with the same predicted and observed treatments. The major drawback of this method is that it results in discarding a considerable quantity of data points. Figure 4 shows the cumulative proportion of matched test cases as a function of the percentage of the population targeted for each MTUM approach. As expected, due to the random allocation of treatments, performance metrics use approximately half of the total test samples for evaluation.

On the other hand, we also rank the MTUM approaches based on the different performance metrics. This is illustrated in Fig. 5. The horizontal axis displays the different models, and the vertical axis shows their average ranking according to performance metrics (Qini and expected responses with 10 percent and 100 percent targeting). The shapes represent the evaluation metrics, and the lengths of the vertical lines represent the dispersion of the ranks. The average rank of a model is calculated based on its performance with respect to each data set (i.e., the model with the best performance is ranked first). Later, ranks are averaged among the 8 data sets with respect to the evaluation metrics.

It is observed that most of the MTUM approaches do not consistently outperform the others. The MMOARF generally achieves satisfactory results for all data sets and, therefore, is similarly ranked by the evaluation metrics. Remarkably, the CKNN algorithm performs poorly for all data sets and holds the worst position in the ranking.

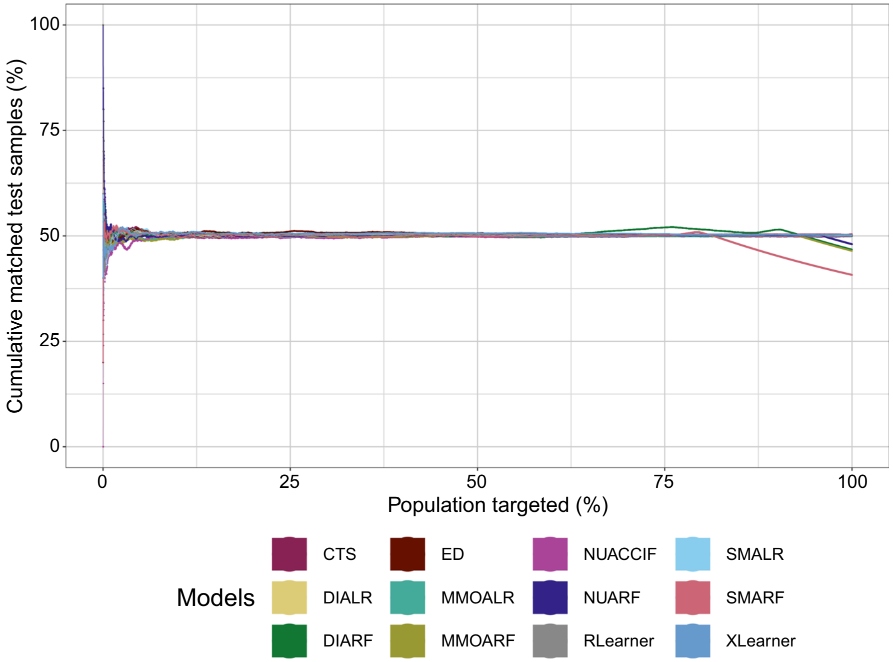

Fig. 4 Percentage of matched test samples as a function of the population targeted for the Hillstrom data set. The different curves correspond to the 12 different experimentally evaluated MTUM approaches

Moreover, recent algorithms such as RLearner and XLearner perform competitively when the metric of evaluation is the expected response. For the Qini metric, the methods that employ decision trees, such as SMARF, ED and CTS, exhibit better results.

The methods proposed in this study are competitive in terms of performance compared to current MTUM techniques. Irrespective of the size of the data sets, they achieve the best results in relation to the Qini metric and expected response (at 100 percent) in five and seven out of the eight data sets, respectively. Their estimations are also more stable, since they have smaller variations. For example, MMOALR is consistently among the best performers for the Hillstrom data set, as observed in the plots of the Qini curves and the expected responses. It is a simple, easily interpretable and computationally inexpensive approach.

In summary, the primary advantage of our methods is their ease of implementation, since they are based on existing algorithms that are readily available and generally known. Moreover, they are built upon conventionally accepted binary uplift modeling approaches that have been previously evaluated by several studies.

6 Conclusion

Predicting treatment effects at the individual level supports decision-makers in the allocation of scarce resources, since it facilitates the identification of the individuals

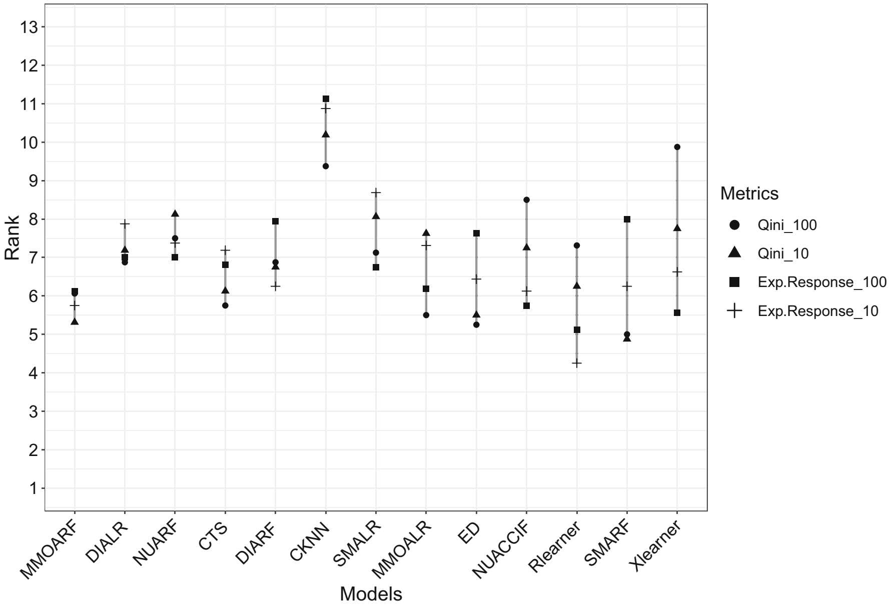

Fig. 5 Overall ranking of the different MTUM approaches by performance metrics. The shapes indicate the performance metric, whereas the lines show the ranking dispersion of each model given the performance metrics

most likely to respond to particular actions. In this regard, uplift modeling serves as a tool to investigate and anticipate the effects of treatments in diverse contexts. Conventional uplift techniques are mostly limited to queries involving the effect of a single treatment. Situations in which more than one treatment alternative is at hand are rarely considered. Therefore, there exists only a vague understanding of which MTUM techniques are available, as well as little evidence regarding the cases which have been elaborated.

We contribute to the state-of-the-art in the field of uplift modeling by: (1) providing an exhaustive survey of the literature on MTUM and applying a framework to classify these methods; (2) proposing two new MTUM techniques; and (3) presenting the results of an extensive benchmarking study, and thus providing ample empirical evidence with respect to the performances of 13 MTUM methods for eight multitreatment uplift data sets. The experiments are performed on data sets from diverse domains such as marketing, political behavior, personalized medicine and human resources. The performances of the models are evaluated by means of the Qini metric and the expected responses in order to facilitate their comparison.

Current multitreatment uplift approaches are classified into two main categories: data preprocessing and data processing approaches. The former learn an uplift model by means of conventional machine learning algorithms. They redefine before training the original outcome variable or extend the input space by adding dummies and interaction terms. In contrast, data processing approaches separately train standard predictive algorithms or adapt their internal functioning. As a result, the uplift can be computed indirectly or directly. Indirect estimation separately processes the information contained in each treatment group, whereas direct estimation uses a multitreatment uplift algorithm that includes all treatments during training.

This paper extends the modified outcome method originally proposed for binary uplift modeling to the MTUM case. The MMOA directly estimates the uplift by means of any standard multiclass probabilistic classification algorithm. Moreover, the NUA takes advantage of existing binary uplift modeling machine learning algorithms. As opposed to the SMA, fewer models are trained, and each treatment is directly contrasted with the control group.

Evaluating the performance of MTUM techniques is challenging due to the fundamental problem of causal inference. Estimating the true uplift is an impossible task in reality, since an individual cannot be simultaneously exposed to all treatments. Therefore, the different counterfactual scenarios are unobservable. In this article, conventional uplift evaluation methods (i.e., uplift curve and Qini metric) are implemented and adapted to the multitreatment case and contrasted with the expected response approach recently proposed by Zhao et al. (2017a) and Zhao et al. (2017b). Given that treatments are randomly assigned to test cases, the predicted optimal treatments do not necessarily match the observed treatments. As such, only matched test cases are considered in evaluating the performances of models. Although it is expected and observed that such strategy implies a considerable data loss of approximately 50 percent, it assures a correct evaluation of the performance of MTUM techniques.

The experimental setup includes an inventory of eight data sets from various domains. This facilitates testing uplift techniques in diverse multitreatment scenarios. In addition, studies where selection bias is tested and controlled are rare in the uplift literature. Therefore, we verify and, if needed, correct the imbalance among the pretreatment characteristics of the treatment groups by applying matching. We apply PSM to four data sets where the chi-square test detected imbalance. However, this does not necessarily eliminate the risk of selection bias, nor does it aim to improve the performances of the models.

Different MTUM approaches are considered in the experimental evaluation. The Friedman test confirms that none of the evaluated techniques consistently outperform other techniques in terms of the Qini metric and the expected response. Therefore, we conclude that the two techniques proposed in this study are competitive. They achieve similar performances as current MTUM techniques. In addition, the proposed approaches can be easily implemented, since the required algorithms are readily available in standard software packages. Generally, the study shows that the performance of an uplift multitreatment technique is highly context-dependent.

On the other hand, we observe that the size of the uplift data set has implications for the capacity of a model to compute reliable estimates. Small data sets such as Bladder, Colon1, Colon2 and AOD present high volatility in the uplift predictions among different folds in the cross-validation evaluation.

This study includes certain limitations, which can serve as a motivation for future research. First, optimal matching with propensity scores leads to an important loss of information when treatment groups are not of equal size. In addition, this technique is highly dependent on the correct specification of the set of observable characteristics. Other methods for correcting selection bias could offer more reliable uplift estimates. Second, to ensure a correct evaluation, our study does not consider test cases for which the predicted and the observed treatment do not match. Consequently, a significant amount of data is obviated. Other solutions may consider all test cases, wherein mismatches are penalized but not removed from the analysis. Finally, the level of the analysis can be enriched by discriminating among types of treatments and individuals. Inexpensive and effective treatments should be privileged over less effective and costly treatments. Analogously, some customers are more valuable than others.

Compliance with ethical standards

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

References

Alemi F, Erdman H, Griva I, Evans CH (2009) Improved statistical methods are needed to advance personalized medicine. Open Transl Med J 1:16

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. Annual Review of Economics 11

Athey S, Wager S (2017) Efficient policy learning. Papers 170202896, arXivorg, revised September 2019

Bertsimas D, Kallus N (2019) From predictive to prescriptive analytics. Management Science

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Chen X, Owen Z, Pixton C, Simchi-Levi D (2015) A statistical learning approach to personalization in revenue management. SSRN Electronic Journal

Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Trans Inf Theory 13(1):21–27

Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7(Jan):1–30

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

Diamond A, Sekhon JS (2013) Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Rev Econ Stat 95(3):932–945

Friedman M (1940) A comparison of alternative tests of significance for the problem of m rankings. Annals Math Stat 11(1):86–92

Gerber AS, Green DP, Larimer CW (2008) Social pressure and voter turnout: evidence from a large-scale field experiment. Am Political Sci Rev 102(1):33–48

Gross SM, Tibshirani R (2016) Data shared lasso: a novel tool to discover uplift. Comput Stat Data Anal 101:226–235

Gubela R, Lessmann S, Haupt J, Baumann A, Radmer T, Gebert F (2017) Revenue uplift modeling. In: Thirty eighth international conference on information systems, South Korea

Guelman L (2014) Uplift: uplift modeling. R package version 03:5

Guelman L (2015) Optimal personalized treatment learning models with insurance applications. Dissertation, Universitat de Barcelona

Guelman L, Guillén M, Pérez-Marín AM (2012) Random forests for uplift modeling: an insurance customer retention case. In: Engemann KJ, Gil-Lafuente AM, Merigó JM (eds) Modeling and simulation in engineering, economics and management. Springer, Berlin, pp 123–133

Guelman L, Guillén M, Pérez Marín AM (2014a) Optimal personalized treatment rules for marketing interventions: a review of methods, a new proposal, and an insurance case study. UB Riskcenter Working Paper Series, 2014/06

Guelman L, Guillén M, Perez-Marin AM (2014b) A survey of personalized treatment models for pricing strategies in insurance. Insur: Math Econ 58:68–76

Guelman L, Guillén M, Pérez-Marín AM (2015) A decision support framework to implement optimal personalized marketing interventions. Decision Support Syst 72:24–32

Hansen BB, Bowers J (2009) Covariate balance in simple stratified and clustered comparative studies. Qual Control Appl Stat 54(1):101–102

Hansotia B, Rukstales B (2002) Incremental value modeling. J Interact Mark 16(3):35–46

Hillstrom K (2018) The minethatdata e-mail analytics and data mining challenge. Minethatdata blog. http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html, Retrieved: 21.06.2018

Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81(396):945–960

Imai K, Ratkovic M et al (2013) Estimating treatment effect heterogeneity in randomized program evaluation. Annals Appl Stat 7(1):443–470

Jaskowski M, Jaroszewicz S (2012) Uplift modeling for clinical trial data. In: ICML 2012 Workshop on clinical data analysis

Kallus N (2017) Recursive partitioning for personalization using observational data. In: Proceedings of the 34th international conference on machine learning, ICML'2017, pp 1789–1798

Kallus N, Zhou A (2018) Confounding-robust policy improvement. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, inc., pp 9269–9279, http://papers.nips.cc/paper/8139-confounding-robust-policy-improvement.pdf

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 Mark Anal 2:218–238

Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. J Roy Stat Soc: Ser C (Appl Stat) 29(2):119–127

Kuhn M, Johnson K (2013) Applied predictive modeling, vol 26. Springer, New York

Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. In: Proceedings of the national academy of sciences 116(10):4156–4165. 10.1073/pnas.1804597116, https://www.pnas.org/content/116/10/4156.full.pdf

Kuusisto F, Costa VS, Nassif H, Burnside E, Page D, Shavlik J (2014) Support vector machines for differential prediction. In: Joint european conference on machine learning and knowledge discovery in databases, Springer, pp 50–65

Lai LYT (2006) Influential marketing: a new direct marketing strategy addressing the existence of voluntary buyers. Dissertation, Simon Fraser University School of Computing Science, Burnaby, BC, Canada

Leo B, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth Int Group 37(15):237–251

Li C, Yan X, Deng X, Qi Y, Chu W, Song L, Qiao J, He J, Xiong J (2018) Reinforcement learning for uplift modeling. arXiv:1811.10158

Liaw A, Wiener M (2002) Classification and regression by randomforest. R News 2(3):18–22. https://CRAN.R-project.org/doc/Rnews/

Lo VS (2002) The true lift model: a novel data mining approach to response modeling in database marketing. ACM SIGKDD Explor 4(2):78–86

Lo VS, Pachamanova DA (2015) From predictive uplift modeling to prescriptive uplift analytics: a practical approach to treatment optimization while accounting for estimation risk. J Mark Anal 3(2):79–95

Lopez MJ, Gutman R et al (2017) Estimation of causal effects with multiple treatments: a review and new ideas. Stat Sci 32(3):432–454

McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF (2013) A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med 32(19):3388–3414

Michel R, Schnakenburg I, von Martens T (2017) Effective customer selection for marketing campaigns based on net scores. J Res Interact Mark 11(1):2–15

Morgan SL, Winship C (2015) Counterfactuals Causal Inference. Cambridge University Press, Cambridge

Moro S, Laureano RMS, Cortez P (2011) Using data mining for bank direct marketing: an application of the CRISP-DM methodology. In: Proceedings of the European simulation and modelling conference, Guimaraes, Portugal, pp 117–121

Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. Decis Support Syst 62:22–31

Nemenyi P (1963) Distribution-free multiple comparisons. Dissertation, Princeton University

Nie X, Wager S (2017) Quasi-oracle estimation of heterogeneous treatment effects. arXiv:1712.04912

Pearl J (2009) Causality. Cambridge University Press, Cambridge

Peters J, Janzing D, Schölkopf B (2017) Elements of causal inference: foundations and learning algorithms. MIT press, Cambridge

Quinlan J (1993) C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Francisco

R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

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, Surry PD (1999) Differential response analysis: modeling true response by isolating the effect of a single action. Credit scoring and credit control IV Edinburgh, Scotland

Radcliffe NJ, Surry PD (2011) Real-world uplift modelling with significance-based uplift trees. Stochastic solutions white paper(1):1–33

Ripley B, Venables W (2011) nnet: Feed-forward neural networks and multinomial log-linear models. R package version 7(5):

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688

Rudaś K, Jaroszewicz S (2018) Linear regression for uplift modeling. Data Min Knowl Disc 32(5):1275–1305

Rzepakowski P, Jaroszewicz S (2010) Decision trees for uplift modeling. In: 2010 IEEE International conference on data mining, IEEE, pp 441–450

Rzepakowski P, Jaroszewicz S (2012) Decision trees for uplift modeling with single and multiple treatments. Knowl Inf Syst 32(2):303–327

Sawant N, Namballa CB, Sadagopan N, Nassif H (2018) Contextual multi-armed bandits for causal marketing. In: Proceedings of the 35th international conference on machine learning, Stockholm, Sweden, PMLR 80

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

Sutton RS, McAllester DA, Singh SP, Mansour Y (2000) Policy gradient methods for reinforcement learning with function approximation. In: Solla SA, Leen TK, Müller K (eds) Advances in neural information processing systems 12. MIT Press, Cambridge, pp 1057–1063

Therneau TM (2015) A Package for survival analysis in S. https://CRAN.R-project.org/package=survival, version 2.38

Tian L, Alizadeh AA, Gentles AJ, Tibshirani R (2014) A simple method for estimating interactions between a treatment and a large number of covariates. J Am Stat Assoc 109(508):1517–1532

Van Rossum G, Drake FL Jr (1995) Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands

Zhao Y, Fang X, Simchi-Levi D (2017a) A practically competitive and provably consistent algorithm for uplift modeling. In: 2017 IEEE International conference on data mining (ICDM), IEEE, pp 1171–1176

Zhao Y, Fang X, Simchi-Levi D (2017b) Uplift modeling with multiple treatments and general response types. In: Proceedings of the 2017 SIAM International conference on data mining, SIAM, pp 588–596

Zhao Z, Harinen T (2019) Uplift modeling for multiple treatments with cost optimization. arXiv:1908.05372

Zhou Z, Athey S, Wager S (2018) Offline multi-action policy learning: generalization and optimization. arXiv:1810.04778

Affiliations

Diego Olaya[1] · **Kristof Coussement**[2] · **Wouter Verbeke**[1]

✉ Diego Olaya
diego.olaya@vub.be

Kristof Coussement
k.coussement@ieseg.fr

Wouter Verbeke
wouter.verbeke@vub.be

[1] Data Analytics Laboratory, Faculty of Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[2] IESEG School of Management, Rue de la Digue 3, 59000 Lille, France