University of Cologne

Faculty of Economics and Social Sciences

Information Systems Bachelor Thesis:

Leveraging Sleeping Dogs to Develop a Novel Uplift Modeling Approach for Direct Marketing

Henri Beyer; Patrick Seidel
Zülpicher Wall 46, 50674 Cologne; Champagneweg 1, 53332 Bornheim
hbeyer@smail.uni-koeln.de; pseidel3@smail.uni-koeln.de

Student ID: 7375073; 7380447

Advisor: Prof. Dr. Detlef Schoder
Second Advisor: Jannik Rößler

Cologne 15.08.2023

Table of Contents

Table of Contents I
Table of Figures II
Table of Tables III
Table of Abbreviations IV
1 Introduction 1
2 Background and Related Work 3
2.1 Prior Works 3
2.2 Gap in Related Literature 6
3 Methodology 8
3.1 Method 8
3.2 Dataset 10
3.2.1 Data preprocessing 11
3.3 Evaluation metrics 12
4 Results 15
4.1 Uplift Modeling Algorithm Evaluation 15
4.2 Classification Algorithm Evaluation 16
4.3 Combined Two-Step Approach Evaluation 17
5 Discussion 20
6 Conclusion 23
References 24

Table of Figures

Figure 1 Qini curves displaying negative uplift in the last deciles (De Caigny et al., 2021). 6

Figure 2 Example qini curve. 14

Figure 3 Uplift Modeling Algorithm Evaluation. 15

Figure 4 Average qini curve across 10-fold cross validation of X-learner and X-learner with two-step approach. 19

Table of Tables

Table 1 Studies examined for Qini curves with uplift values exceeding the ATE by more than 10%. 7

Table 2 Mean evaluation metrics for each model’s best performing threshold. 17

Table 3 Unscaled qini coefficients for traditional and two-step uplift modeling methods across all 10 folds of the cross-validation. 18

Table of Abbreviations

ATE	Average Treatment Effect
CATE	Conditional Average Treatment Effect
CRM	Customer Relationship Management
GBM	Gradient Boosting Machine
ITE	Individual Treatment Effect
LR	Logistic Regression
MLP	Multi-layer Perceptron
RF	Random Forest
ROI	Return on Investment
UQC	Unscaled Qini Coefficient

1 Introduction

Targeting policies play a central role in today's data-driven marketing, as they are vital for optimizing campaigns. Given the diverse preferences and behaviors of customers, it is crucial to customize targeting to suit individual responses (Ascarza, 2018). The goal for campaign managers is to enhance the return on investment (ROI) by identifying the most suitable and receptive audience for their marketing campaign (Strycharz et al., 2019). To accomplish this, practitioners are now focusing on predicting a customer's individual sensitivity to marketing efforts, commonly known as the individual treatment effect (ITE) (Simester et al., 2020).

Companies across various marketing domains, including churn management, customer acquisition, or product introductions, are increasingly leveraging ITEs to develop highly effective targeting policies. Booking.com showed that implementing a personalized targeting policy can transform an underperforming promotional campaign into a profitable one by substantially boosting the response rate (Goldenberg et al., 2020). Similarly, when launching new products, Uber successfully employs targeted cross-selling techniques. By focusing on the 30% of most receptive customers, as determined by ITE ranking, they minimize wasted targeting efforts while still achieving a conversion increase comparable to targeting all customers (Chen et al., 2020).

These optimized direct marketing policies are enabled by the emerging practice of uplift modeling. Uplift modeling is used to identify selected customers with the highest positive impact of a treatment denoted by the ITE (Olaya et al., 2020). It can distinguish between customer segments like "sure things" who consistently make purchases regardless of promotions (e.g. auto-response), as well as customers who are negatively affected by treatment, referred to as "sleeping dogs" (e.g. individuals prone to churn if reminded of a contract by a retention campaign) (Devriendt et al., 2018).

Ideally, to develop a model that predicts ITEs, we would need to simultaneously observe a customer receiving a treatment and not receiving a treatment, enabling us to measure the causal impact of the treatment. We could then train a machine learning model using this data. In reality, it is impossible to make both observations for each individual, which is known as the fundamental problem of causal inference (Holland, 1986). This problem arises from the fact that we can never definitively analyze the exact impact a treatment had on a single individual, lacking a reliable ground truth. Therefore, uplift modeling predicts ITEs by leveraging data from randomized controlled experiments (e.g., A/B tests) and employing causal inference techniques.

Multiple researchers showed that in direct marketing contexts like catalogue mailing or email promotions, uplift modeling has proven to enhance profits through the implementation of personalized targeting strategies (Hitsch et al., 2018; Rößler & Schoder, 2022). Additionally, Ascarza et al. (2018) emphasized the concept of sleeping dogs in churn management, showing that an uplift modeling-based targeting policy, which considers a treatment's negative effects, outperformed the traditional approach of targeting customers with the highest risk of churning.

However, despite research highlighting the vital role of sleeping dogs in contractual settings like churn management, they can often be disregarded in non-contractual direct marketing contexts, due to infrequent occurrences or minimal financial impact. The experimental data from multiple researchers indicates a minimal presence of sleeping dogs in non-contractual direct marketing contexts (Devriendt et al., 2018; Rößler & Schoder, 2022).

Thus, ignoring sleeping dogs allows us to introduce the presence of ground truth for a subgroup of individuals, thereby simplifying the problem by narrowing down the number of subjects to which the fundamental problem of causal inferences applies. Utilizing this assumption, we will develop a novel uplift modeling approach and demonstrate its effectiveness by applying it to a real-world dataset from a renowned international fashion brand. Through a comparison with other state of the art uplift modeling approaches, we will demonstrate the superior performance of our method.

We concentrate on data with a binary treatment indicator (i.e., customers receive a treatment or not) and a binary response variable (i.e., customers respond or do not respond) throughout the paper. Additionally, according to Athey and Imbens (2015), we assume that the data come from randomized controlled experiments (i.e., A/B testing) or meet the unconfoundedness and stable unit treatment value assumptions.

2 Background and Related Work

2.1 Prior Works

The evolution of targeting in marketing had its roots in the early 20th century when research revealed distinct reading patterns among men and women in newspaper sections (McDonough & Egolf, 2002). By the mid-20th century, the growing diversity in products prompted Smith in 1956 to challenge the then-prevailing mass marketing strategy, which viewed the market as a homogeneous entity. He advocated for a more nuanced approach, proposing a segmentation marketing strategy that acknowledged the market's heterogeneity and varying customer preferences (Smith, 1956). As product diversity further expanded and advancements in Customer Relationship Management (CRM) emerged, the focus shifted from a product-centric approach to a customer-centric one (Levitt, 1984).

As market segmentation has become widely recognized as a critical factor for successful advertising (Arens & Weigold, 2017), targeting policies, referring to matching different marketing actions to different customers, have become essential for the optimization of data-driven marketing campaigns. A consistent finding across studies is that more nuanced targeting policies tend to yield better performances if the goal is to maximize conversion rates (Hartmann, 2010; Li et al., 2022; Rossi et al., 1996). If enough data is available these policies can be optimized to tailor specific marketing actions to individual customers, such as serving digital advertisements to users, displaying different properties to various homebuyers, or offering free trials to new customers (Simester et al., 2020a; Yoganarasimhan et al., 2023).

Our research primarily focuses on the application of targeting policies in a direct marketing context. In this regard, targeting policies aim to influence customer response to marketing initiatives. They play a role in enhancing customer retention in churn management (Ascarza, 2018), optimizing the overall conversion rate, and thereby improving the ROI (Goldenberg et al., 2020; Liu, 2022; Musalem et al., 2008). They can also be instrumental in identifying customers particularly likely to be receptive to a new product launch (Chen et al., 2020).

Traditionally, marketers derived targeting policies by predicting a customer’s response probability (Coussement et al., 2015; Guido et al., 2011). However, recent research indicates that a customer's likelihood to respond may not be the most effective criterion for determining which customers to target. Instead, customer-specific sensitivity to a particular treatment, in our case a marketing initiative, has been identified as a superior factor (Ascarza, 2018). This is due to the varied responses across the customer base to certain marketing efforts, demonstrating the heterogeneity of customer behavior. As a measure for the customer's sensitivity to a treatment, researchers and practitioners use the ITE, i.e., the causal effect that a treatment has on the customer's response probability (Devriendt et al., 2018). Using ITEs instead of response probabilities enables marketers to consider potentially futile or adverse effects of marketing efforts (Ascarza, 2018) and to save resources on customers with a high response probability but a low treatment sensitivity (Musalem et al., 2008).

When optimizing a targeting policy, the ITE is the difference in response of a customer when targeting them compared to not targeting them. As it is impossible to make both observations, true ITEs cannot be measured. This is referred to as the "fundamental problem of causal inference" (Holland, 1986). Thus, researchers have opted to approximate the ITE with a model based on Rubin's (1974) model to approximate the difference in causal effect of two treatments by taking the difference of two average treatment effects (ATE). The ITE can be approximated by enriching Rubin's model with attribute or feature vectors of the individuals exposed to the treatment to measure the difference in ATE of similar subgroups of individuals. This more specific approximation is called conditional average treatment effect (CATE) (Imbens & Rubin, 2015) and is used by many marketing researchers to approximate ITEs.

Ellickson et al. (2022) used the CATE to predict customer responses and increase the profitability of promotional E-Mail campaigns and Zantedeschi et al. (2017) employed an ad-stock model to evaluate customer responses in a multi-channel setting involving email and catalog mailing. Smith et al. (2023) used the CATE to estimate the effect different targeted pricing policies have per customer to determine an optimal pricing policy. Ascarza (2018) predicted customers sensitivity to a retention incentive. This approach proved successful in optimizing two retention campaigns to minimize customer churn. Yoganarasimhan et al. (2023) used the CATE to approximate customers' responses to different trial lengths of free trail promotions in the "Software as a Service" market. These approximations allowed them to design and optimize personalized treatment assignment policies.

Utilizing ITEs enables the evaluation of an unlimited number of targeting policies on a single randomized dataset, as highlighted by Hitsch et al. (2018). This technique significantly minimizes expenses when contrasted with conducting distinct field experiments for each policy. Hitsch et al., (2018) further showcased the applicability of their framework by maintaining consistent results in a campaign for the consecutive year using the initial training data, thereby demonstrating its robustness and transportability.

One of these ways to algorithmically estimate ITEs on customer level using CATE is the emerging field of uplift modeling. This concept, initially termed "differential response modeling," was introduced by Radcliffe and Surry in 1999 (Radcliffe & Surry, 1999). Unlike traditional response modeling, uplift modeling differentiates between responses induced by the treatment and baseline responses independent of it, thereby providing a more nuanced representation of treatment-induced changes in response likelihood (Rößler et al., 2021). It achieves that by leveraging randomized controlled experiments (A/B tests) to estimate a customer's ITE based on the CATE (Gubela et al., 2019). This enables the calculation of the CATE by comparing the response of very similar customers from the control and treatment group to get a more accurate ITE prediction.

Uplift modeling approaches can mainly be categorized into three categories: two-model, class transformation and direct (Gutierrez & Gerardy, 2016). While benchmarkings show that no individual algorithm is universally superior recent research indicates that in general direct estimation methods yield the best results (Devriendt et al., 2018; Hitsch et al., 2018). In contrast to the two-model approach which estimates the uplift by building separate models for the treatment and control group (Kane et al., 2014), and the class transformation approach which estimates the uplift by transforming the problem into a binary classification task (Kane et al., 2014), the direct approach builds upon existing machine learning approaches by modifying them to train the models uplift directly on the CATE, mostly using random forests (RF) as a base model (Guelman et al., 2014; Hansotia & Rukstales, 2002; Sołtys et al., 2015).

In essence, uplift modeling strives to accurately identify persuadable customers while actively steering clear of treating sleeping dogs or sure things. As Devriendt et al. previously asserted, categorizing customers into these groups can greatly depend on the specific campaign (Devriendt et al., 2018). Notably, their benchmarking study revealed a lack of downlift (negative uplift) when contacting a larger customer segment in specific marketing campaign data. This suggests that in certain campaign contexts "there are no or few do-not-disturbs in the customer population" (Devriendt et al., 2018, p. 37). We aim to expand on this observation and leverage it to develop a new, more effective uplift modeling approach tailored to specific campaign settings.

2.2 Gap in Related Literature

To the best of our knowledge, no scientific work has further investigated the claim of a lack of sleeping dogs or another group in a dataset. This establishes a gap in the existing uplift modeling literature. To verify the research gap, we conducted an examination of the qini curve evaluations across a diverse range of uplift papers, encompassing a sample of 133 models trained on both contractual (6 datasets) and non-contractual (8 datasets) contexts from 5 distinct studies. We assessed the qini curves by examining the uplift values in each decile, noting that a decile can only have negative uplift if the control response rate exceeds treatment response rate. Based on the assumed quality of the algorithms in the studies, we expected sleeping dogs to be positioned towards the curve's end, leading to uplift values exceeding the ATE, with a subsequent decline as the sleeping dogs appear, as shown in Figure 1. Recognizing the inherent volatility often associated with uplift algorithms (Rößler et al., 2021), a threshold was applied, whereby only models exceeding the ATE by more than 10% were classified as detecting negative uplift.

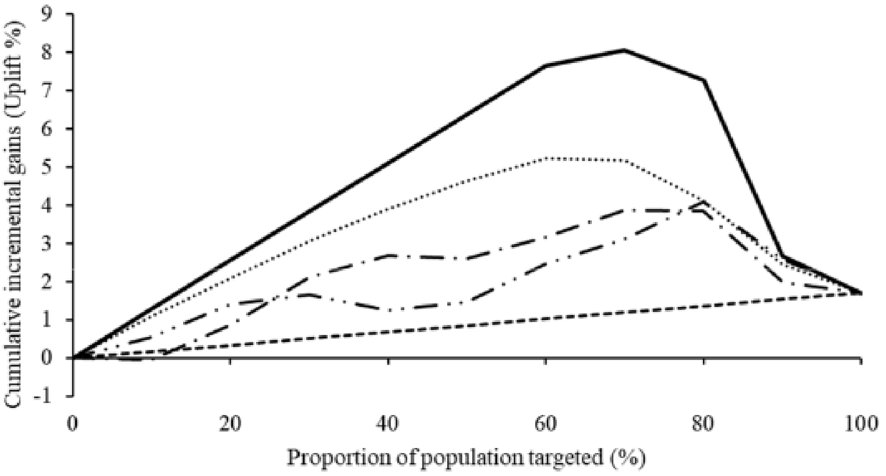


Figure 1 Qini curves displaying negative uplift in the last deciles (De Caigny et al., 2021).

Remarkably, only one out of 90 models trained and evaluated on non-contractual datasets exhibited negative uplift prediction, while all 43 models trained and evaluated on contractual datasets predicted a significant negative uplift in the last deciles. This asymmetry reinforces the idea that sleeping dogs might be absent in certain settings, particularly non-contractual ones. Our findings further suggest that the observation by Devriendt et al. (2018) about the absence of sleeping dogs is not exclusive to their dataset but might be a consistent pattern in non-contractual environments. By addressing this gap with our proposed method, we believe the performance of current techniques can be improved.

Reference	Dataset Description	Type	Negative Uplift
(Ascarza, 2018)	Free Credit when Recharging SIM Card	contractual	x (1/1)
(Ascarza, 2018)	Discount for Subscription-Based Membership	contractual	x (1/1)
De Caigny et al., 2021	Churn Prevention Campaign	contractual	x (4/4)
(Devriendt et al., 2018)	Insurance Campaign	contractual	x (11/11)
(Devriendt et al., 2018)	Financial Services Retention Campaign	contractual	x (11/11)
Rößler & Schoder, 2022	Churn Prevention Campaign	contractual	x (15/15)
Devriendt et al., 2018	Online Merchandise	non-contractual	(0/11)
Devriendt et al., 2018	Retailer E-Mail Campaign	non-contractual	(0/11)
Rößler & Schoder, 2022	SMS Marketing Campaign	non-contractual	x (1/15)
Rößler & Schoder, 2022	Email Marketing Campaign	non-contractual	(0/15)
Rößler & Schoder, 2022	Promotional Campaign via Mobile App	non-contractual	(0/15)
Rößler & Schoder, 2022	Email Marketing Campaign	non-contractual	(0/15)
Rößler et al., 2022	Print Marketing Campaign Offering Discounts	non-contractual	(0/4)

Rößler et al., 2022	Print Marketing Campaign Offering Discounts	non-contractual	(0/4)

Table 1 Studies examined for Qini curves with uplift values exceeding the ATE by more than 10%.

3 Methodology

Thus, we developed a new uplift modeling algorithm under the assumption of the absence of sleeping dogs within the dataset. The minimal probability of sleeping dogs existing within the dataset was corroborated by both the marketing managers of the fashion brand and the analysis of traditional uplift modeling qini curves. Based on their business knowledge, the marketing managers asserted that the lack of negative treatment effects was probably due to the customers not having a direct negative effect from a discount coupon. Furthermore, none of the 15 traditional uplift modeling algorithms, which were trained on the dataset, demonstrated any decrease in uplift when higher percentages of the customer base were contacted. This suggests that sleeping dogs were absent (Devriendt et al., 2018).

This fundamentally shifts the underlying assumption of uplift modeling regarding the total absence of ground truth when modeling customer sensitivity. For example, in uplift modeling, it is uncertain whether a customer that was treated with the marketing intervention but didn’t respond is a lost cause, meaning he is indifferent to the treatment, or a whether he is a sleeping dog, meaning he is adversely impacted by it. Disregarding sleeping dogs however, we can certainly identify these treatment non-responders as lost causes. Using the same logic, we can also identify control responders as certain sure things and not potentially sleeping dogs. However, for the treatment responders and control non-responders remains the problem of causal inference, meaning we can infer ground truth only for a subset of the data.

This introduction of ground truth enables the use of both traditional supervised machine learning techniques and uplift modeling specific techniques in a two-step approach. In the following we describe the uplift modeling approach implemented based on this assumption, as well as the underlying datasets and evaluation metrics.

3.1 Method

Our proposed uplift modeling method consists of two steps: First, we use a classifier to predict lost causes in our dataset. Second, we utilize an uplift modeling algorithm to predict the uplift of each individual. Finally, we combine these predictions by adjusting the sensitivity values of identified lost causes thus representing their low treatment sensitivity.

In the initial stage of our approach, we employ a RF to train a classifier. This classifier's purpose is to identify treatment non-responders, which, in the context of our study, can be referred to as a "lost causes classifier". Although we have information about the ground truth for sure things to the same extent as for the lost causes, it is not feasible to train a classification model on the data in our case, due to the low number of sure thing occurrences in the entire dataset. For a given train dataset, we create a target variable denoted as “lost cause” which labels each treatment non-responder as 1 and every other combination of the treatment and response variable as 0. This data is used to train an RF model using scikit-learn’s (Pedregosa et al., 2011) implementation of the algorithm. In our hyperparameter configuration, we set a maximum depth of 20 nodes per tree to avoid overfitting and maintain a manageable computational complexity. Moreover, we use 100 estimators, fix the random state at 0 to produce reproducible results and employ the “balanced_subsample” method for the class weight, as it is specifically designed to account for label imbalances. Following an array of testing iterations, we determined that a higher threshold of 0.65, which ensures a higher true positive rate, enhances the overall performance of the combined model in the two-step approach. Consequently, the trained RF predicts the lost cause class probabilities for each record of a given dataset and designates each record with a probability of greater or equal to 0.65 as a lost cause.

For the second step, we use the X-Learner, which was introduced by Künzel et al. (2019). It extends the two-model ITE estimation to a three-step ITE estimation. First, the method models the conditional expectations of the outcomes for individuals subject to a treatment $\mu_1(x) = E[Y_i(1)|X_i = x]$ and not subject to a treatment $\mu_0(x) = E[Y_i(0)|X_i = x]$ separately, using the treatment and control group, respectively. Second, for each individual it imputes the treatment effect. That is, for the individuals in the treatment group, the difference between the true outcome $Y(1)$

and the estimated outcome using the control estimator $\hat{\mu}_0(x)$ is calculated: $D^1 = Y_i(1) - \hat{\mu}_0(x)$. For the individuals in the control group, the difference between the estimated outcome using the treatment estimator $\hat{\mu}_1(x)$ and the true outcome *Y(0)* is calculated: $D^0 = \hat{\mu}_1(x) - Y_i(0)$.

The imputed treatment effects D^1 and D^0 are then used as outcome variables to estimate $\hat{\tau}_1(x) = \mathrm{E}[D^1|X = x]$ and $\hat{\tau}_0(x) = \mathrm{E}[D^0|X = x]$ using any type of base learner (e.g., decision tree, regression model). Third, the ITE is predicted by weighting the estimates from the second step with a weight function $g \in [0, 1]$: $\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$ where g typically is the propensity score (Künzel et al., 2019). We use AutoUM's (Rößler & Schoder, 2022) implementation of the X-Learner, which uses a RF algorithms for all base learners.

In our two-step model, the X-Learner initially predicts the uplift scores. Subsequently, for each individual classified as a lost cause by the RF classifier, we post hoc adjust their uplift, setting it to a value below the minimum predicted uplift by the X-Learner.

3.2 Dataset

Our dataset originates from a global fashion brand which regularly conducts randomized controlled trials to track the performance of their marketing campaigns. We analyze the data of a mail discount campaign conducted in 2020. It contains a total of 200,282 data points across 152 distinct features. These features encompass a wide range of purchasing behaviors, including order counts, turnover segmented by product categories, season-based turnover, and return rates. Additionally, the dataset provides metrics such as open rate, click rate and channel affinity, indicating the responsiveness to prior marketing initiatives.

The ratio of treated to control was about 4:1, with 160,004 customers receiving a mail coupon as part of the treatment group, while the control group consisted of 40,278 customers who received no marketing intervention. A positive response was indicated by any recorded purchase made within the four weeks following the treatment. In total, 42,834 customers made a purchase during this period, whereas 157,448 customers did not. Within the treatment group, the response rate was approximately 21.79%, compared to 19.78% in the control group. This resulted in a relatively modest ATE of 2.01% in the initial campaign.

To ensure the randomization of features across the treatment and control groups we test for equal means for the features across the treatment and control groups via t-tests. 8.2% of features exhibited a statistically significant difference between the treatment and control groups, indicated by a p-value of less than 0.05. This is slightly higher than the 5% of features we would expect due to variance in the data with an alpha of 0.05. To ensure robustness in our study, we presented the data scientist at the fashion company with these findings. This consultation provided assurance regarding the randomized assignment of the treatment and control conditions given their experience in conduction randomized A/B tests.

To further assess the quantitative significance of these findings we examine the covariate balance between the treatment and control samples by calculating the standardized mean difference (1) per feature (Hitsch et al., 2018).

$$\frac{\overline{X_k}(1)-\overline{X_k}(0)}{\sqrt{\frac{s_k^2(0)+s_k^2(1)}{2}}} \tag{1}$$

$\overline{X_k}(1)$ represents the mean of the sample for the treatment group ($w = 1$), and $\overline{X_k}(0)$ represents the mean of the sample for the control group ($w = 0$) with $w \in \{0, 1\}$. Similarly, $s_k^2(0)$ and $s_k^2(1)$ are the variances of the treatment and control groups, respectively. The median standardized difference of means, calculated with the formula representing their corresponding variances, was found to be 0.014. This result is specific to the features where we rejected the equality of means using the t-test. This suggests that the quantitative difference between the treatment and control samples in terms of mean is rather small. Therefore, these minor differences can be disregarded for the purpose of our analysis.

3.2.1 Data preprocessing

Since our dataset is organic and stems directly from a conducted marketing campaign, we had to perform various preprocessing steps to make it suitable for machine learning. First, we split the data into training and testing subsets in a randomized 80-20 split, while ensuring stratified sampling based on the treatment and outcome variable. In the training dataset, we manually removed features which were not useful, such as customer IDs and redundant information. Next, we removed all features with a standard deviation of 0. During these steps, we removed 10 features. In addition, we deleted 13 features that had more than 50% missing values. Up to this point we mirrored the preprocessing steps onto the testing dataset, specifically features removed from the training set were correspondingly removed from the testing set. In the next step we manually evaluated for each numerical feature, whether to treat the missing values as 0, the mean or the maximum of the feature's records and imputed the training dataset based on these rules. Missing values of categorical features were imputed with a generic string ("unknown").

In our evaluation of potential outliers, we detected that 6.45% of the observations featured values outside the range of $[0, 1]$ for variables representing turnover shares in the training set. Following a consultation with a Data Scientist from the fashion brand, it was confirmed that these were indeed erroneous values. Consequently, we made the decision to remove these from the dataset.

For the encoding of categorical features, we employed backward difference encoding which compares the mean of a target variable of each manifestation of the feature to the manifestation adjacent to it (Potdar et al., 2017). In our context the response column is the target variable, subsequently we trained a backward difference encoding model on the training dataset and used this model to encode the nominal features for the training and testing dataset. Although backward difference encoding is designed for ordinal features, iterative tests showed that our models perform best when we choose an arbitrary order for nominal features and encode them with backward difference encoding as well.

Lastly, we employed a ridge regression to assess the relevance of the features within our dataset. Based on this analysis, we selected the 60 most relevant features.

3.3 Evaluation metrics

Initially, we train and evaluate traditional uplift modeling methods on the train dataset to identify a baseline of uplift modeling performance. During this initial training and evaluation, we used a stratified 10-fold cross-validation on the training dataset. To ensure the maximum performance of the uplift modeling algorithms we conducted hyperparameter tuning using a grid search approach.

Our uplift modeling approach consists of two progressive modeling stages, necessitating both an intermediate for performance evaluation and a final evaluation to compare it with traditional uplift modeling methods. Initially, we must assess the performance of our "lost cause classifier". As previously mentioned, we can assume the existence of a definite truth in the target variable within a data subset. Therefore, traditional performance evaluations, which involve comparing predicted outcomes with actual ones, are applicable. To measure the performance of our classifier, we utilize the F1-score class-wise and calculate the weighted F1-score over all class labels. The F1-score is the harmonic mean of precision and recall for a specific class label (Chinchor & Sundheim, 1993). It is commonly used in the evaluation of classification algorithms because it balances the trade-off between precision and recall, providing a comprehensive measure of a model's accuracy, especially in cases like ours where the class distribution is imbalanced. The weighted F1-score is the weighted average of all class label's F1 scores, with the weights being the number of true instances for each class.

Finally, we evaluate the performance of our two-step approach in comparison with other uplift modeling approaches. In this setting we cannot compare actual versus predicted outcomes, due to the absence of ground truth caused by the fundamental problem of casual inference, as outlined in the related work section. To circumvent this problem researchers commonly evaluate the performance by comparing groups of customers rather than individuals (Gubela et al., 2019). This is typically achieved with a decile-based qini curve or the unscaled qini coefficient (UQC) as an aggregated measure (Ascarza, 2018; Devriendt et al., 2018; Gubela et al., 2019; Rößler & Schoder, 2022).

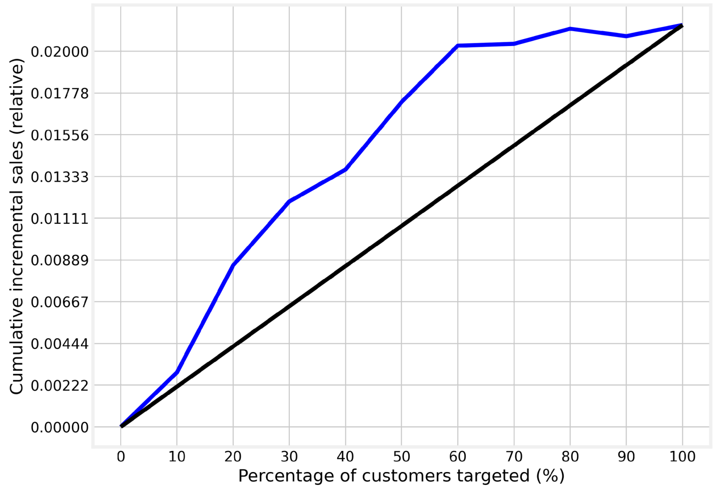
The qini curve sets the number of customers targeted in a relation to the cumulative incremental number of responses achieved (Radcliffe, 2007). To calculate the curve, we first split the entire population of customers S according to their treatment variable into two groups, $C, T \subset S$, corresponding to the treatment and control group. This is done to account for imbalances between the number of control and treatment samples during the calculation. We proceed by sorting both groups according to the customer's predicted ITEs and calculating for each decile the absolute cumulative uplift u, i. e. the additional cumulative responses in the given decile. To calculate u (2), we subtract $N\left(C_R\right)$, the number of responders in the control group's decile, from $N\left(T_R\right)$, the number of responders in the treatment group's decile. To account for potential imbalances in the sizes of the groups, the control group's responders $N\left(C_R\right)$ are scaled by the ratio of the total number of control samples $N(C)$ to the total treatment samples $N(T)$ within the decile.

$$u = N\left(T_R\right) - \frac{N\left(C_R\right)*N(T)}{N(C)} \tag{2}$$

To improve interpretability of the qini curve, we calculate the relative cumulative uplift u_r (3), by dividing the absolute cumulative uplift by the number of treated samples in the entire population, $N\left(S_T\right)$.

$$u_r = \frac{u}{S_T} \tag{3}$$

By plotting the deciles of targeted customers on the x axis and the corresponding u_r values on the y axis, we can draw the qini curve, as shown in Figure 2. The figure depicts two curves for comparison: The blue curve represents the performance of a sample uplift algorithm, while the black curve, drawing a diagonal, indicates the performance achieved through random targeting.



The UQC combines the evaluation based on deciles into a single metric. The UQC is calculated by dividing the area under the uplift curve by the area under the diagonal (Radcliffe & Surry, 2011). According to this definition, UQC values greater than one indicate superior performance compared to random targeting while those lower than one indicate inferior performance.

As we conduct a 10-fold cross-validation we essentially train 10 models on different parts of the data for each approach. Subsequently, we calculate the average qini curves and UQCs to assess the overall performance of the model.

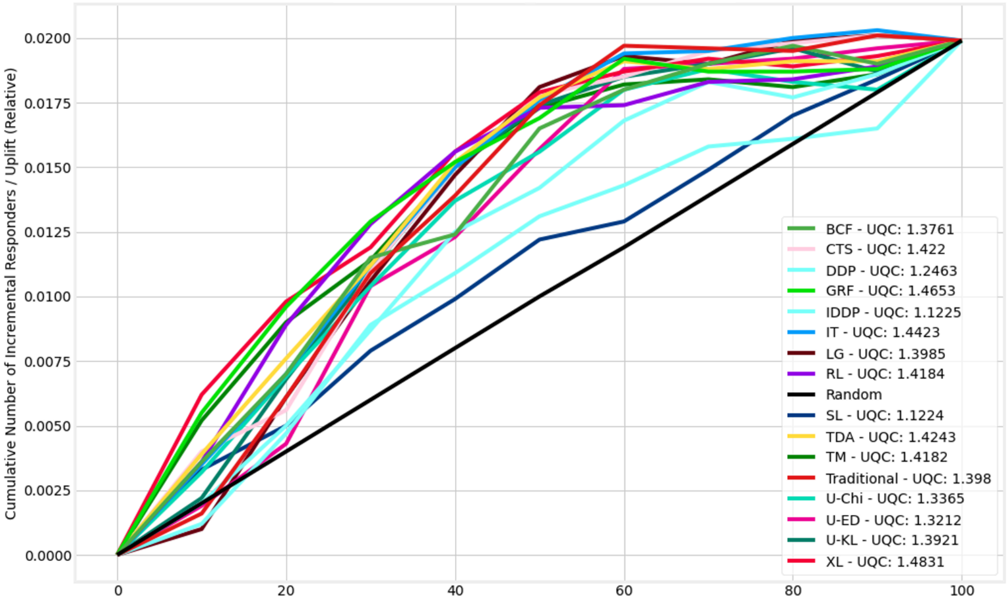
Subsequently, we conduct a final evaluation on the initially set aside test dataset to ensure an unbiased assessment. By training the models on the entirety of the training dataset, we aim to further improve the performance of both models. We then assess the UQC on the full test dataset as a final performance comparison between the two approaches.

4 Results

The two-step approach consists of two separate models: A supervised classification model and an uplift modeling model. To achieve the best possible performance, we determine for both models the individual best performing algorithm, before we evaluate the combined performance, resulting in three distinct evaluations. For each evaluation, we use the train dataset and evaluate using the very same stratified 10-fold cross-validation across all three evaluations.

4.1 Uplift Modeling Algorithm Evaluation

Initially, we evaluate traditional uplift modeling techniques to establish a performance baseline and to identify the most effective algorithms the uplift modeling step of the two-step approach. The average qini curves and UQC from the 10-fold cross-validation are illustrated in Figure 3. In total we evaluated 17 algorithms[1] on the originally separated train dataset. The best UQC was achieved by the X-learner with a score of 1.4831 while the S-Learner algorithm performed worst with a score of 1.1224. The top three algorithms by UQC were the X-learner, Generalized Random Forest (GRF) and Uplift Random Forest with the Interaction Tree (IT) with similar scores of 1.4831 and 1.4423.



To optimize our model's performance, we undertook a systematic hyperparameter tuning process via a grid search for the three algorithms with the highest performance. We optimized the following parameters: maximum tree depth, number of estimators in the RF, minimum samples for leaf nodes, and minimum treatment samples. This led us to assess 1260 unique hyperparameter combinations for each of the 17 uplift modeling algorithms. Despite this extensive search, we observed no persistent improvement in performance in the respective 10-fold cross validations. This suggests our initial parameter selections were quite effective. This is to be expected as these hyperparameters stem from the prior research project practically applying uplift modeling in cooperation with the global fashion brand. Based on these results, we chose the X-Learner for the further analysis.

4.2 Classification Algorithm Evaluation

Upon determining the most effective uplift modeling algorithm, we proceeded to evaluate the optimal classifier for the classification step of the two-step approach. We examine a range of supervised machine learning classification algorithms to pinpoint the optimal one for training on our ground truth data. To ensure a thorough assessment, we selected algorithms from three categories: ensemble methods, feedforward neural networks, and regression algorithms. For the ensemble category, we selected a Gradient Boosting Machine (GBM) model and a RF model. A Multi-layer Perceptron (MLP) model was our choice for the neural network category, while a Logistic Regression (LR) model was chosen for the regression category. The evaluation of these methods was conducted using the scikit-learn (Pedregosa et al., 2011) implementations.

Again, we evaluated the algorithms using the same stratified 10-fold cross-validation as in the previous step. The task was a binary classification for the lost cause customer group, either classifying a customer to be a lost cause or not. In each fold, we calculated the F1-score for both labels, as well as the weighted F1-score for each model. The mean evaluation metrics across all folds are displayed in Table 2. As can be observed from the table, each algorithm showcases its strength in different categories. The models performed comparably, with the weighted F1-score ranging from 0.6469 (MLP) to 0.6581 (RF). Notably, the LR model yielded the highest F1 score on the positive label (0.7782), while the RF model achieved the highest F1-score on the negative label (0.5095) and the highest weighted F1-score. Thus, the RF classifier was used in the further evaluation.

Algorithm	F1 0	F1 1	Weighted F1
GBM	0.443994	0.772726	0.649280
LR	0.442943	**0.778173**	0.652286
MLP	0.459151	0.759850	0.646930
RF	**0.509456**	0.747489	**0.658102**

Table 2 Mean evaluation metrics for each model's best performing threshold.

4.3 Combined Two-Step Approach Evaluation

Subsequently, we proceeded with the final evaluation and comparison of our two-step approach with traditional uplift modeling techniques. We build upon the results from the uplift modeling and classifier evaluation to maximize performance. The classifier was applied in the preliminary stage before implementing the traditional uplift modeling algorithm. For the uplift modeling algorithm, we employed the X-Learner, which had shown the best performance in the initial benchmarking. The focus of this evaluation was not solely on average performance but also on robustness, as indicated by the standard deviation measurements.

As outlined by Table 3, which displays the average UQC across all 10-fold in the final column, the two-step approach outperforms the traditional uplift modeling approach. The two-step X-learner achieved an average UQC of 1.5624, compared to 1.5198 of the traditional X-learner. This results in an average performance increase of about 4.3% across all folds. Furthermore, on average the two-step approach exhibited a reduced volatility with a standard deviation of 0.136 compared to the standard deviation of the traditional approach of 0.168. This constitutes a reduction in volatility across the folds of about 19% on average. These findings suggest a higher robustness in the performance of the two-step approach, underscoring its potential utility in real-world applications.

The fold-wise performance of the traditional and two-step approaches are summarized in Table 3. The two-step approach yields superior performance in 70% of the folds with worse performance than the traditional X-Learner only in folds 1, 6 and 7. As is common in uplift modeling the individual fold performances are quite volatile. The UQC of the traditional approach range from 1.2673 to 1.777 across the 10 folds, while those of the two-step approach range from 1.3286 to 1.783. Consequently, the overall performance spread of the traditional approach is 0.5097, higher than the spread of the two-step approach of 0.4544. The minimum and maximum performance across all folds of the two-step approach is also slightly higher than those of the two-step approach.

Algorithm	1	2	3	4	5	6	7	8	9	10	Avg.
Traditional	**1.635**	1.267	1.472	1.504	1.754	**1.777**	**1.561**	1.281	1.382	1.560	1.520
Two-Step	1.627	**1.329**	**1.565**	**1.611**	**1.783**	1.615	1.486	**1.414**	**1.454**	**1.742**	**1.563**

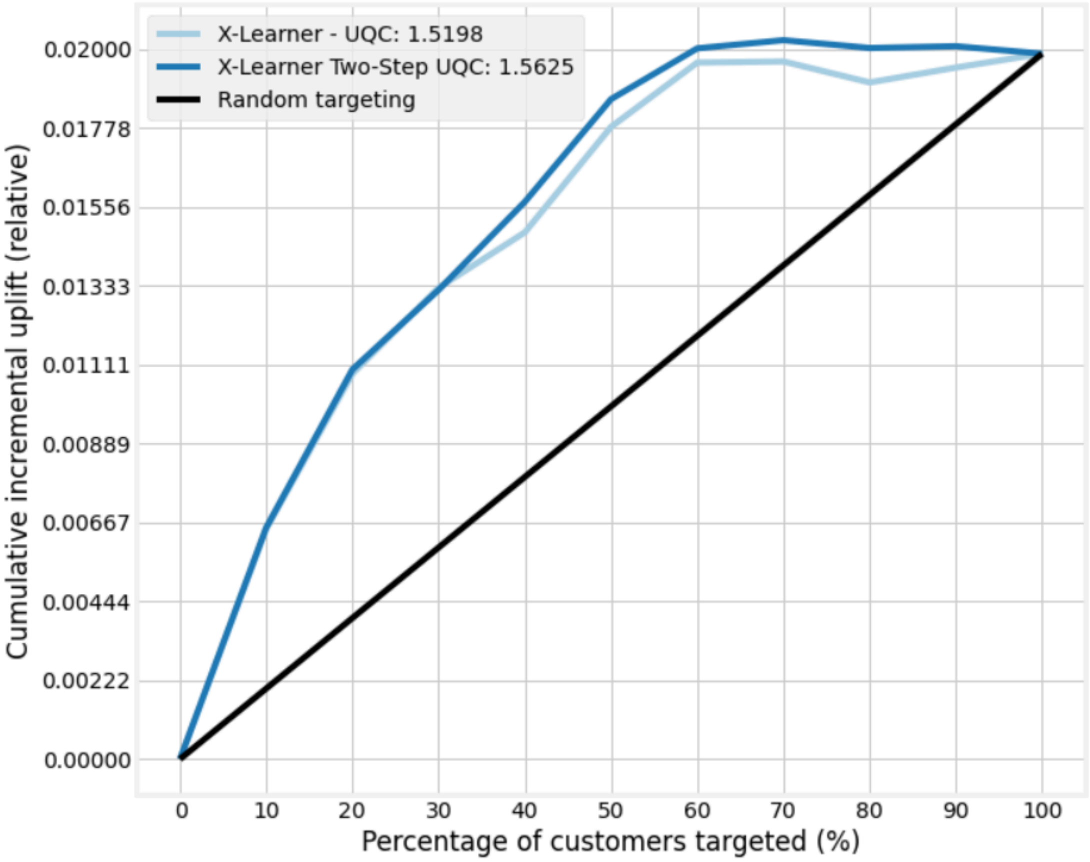
Table 3 Unscaled qini coefficients for traditional and two-step uplift modeling methods across all 10 folds of the cross-validation.

An analysis of the qini curve illustrated in Figure 4 clarifies the origin of the increased UQC performance. Within the top 30% of the customers with the highest predicted sensitivity to treatment both approaches yield almost the same uplift curve. This indicates that the initial classification approach does not falsely classify persuadable customers with a sensitivity to treatment as lost causes. For the following percentages of customers treated, the cumulative relative incremental uplift of the two-step approach is consistently higher than the traditional approach.

Furthermore, the two-step approach achieves the ATE within the dataset of 0.02 by contacting only 60% of the total customer base while the traditional approach does not achieve these scores until the entire customer base is contacted. Both approaches the optimal targeting policy in terms of the balance of uplift and percentage of customers targeted when treating 60% of the total customer base. Based upon the campaign size of 200,000 customers and a profit per conversion of €100[2] the increased uplift of the two-step approach leads to a modest increase of campaign returns of €8000.

To ensure an unbiased evaluation of the performance and robustness of the compared approaches, we conducted a final evaluation using the initially set aside testing dataset. For optimal performance, algorithms were trained using the entirety of the training data set. In these evaluations, the traditional X-Learner yielded an UQC of 1.414. In contrast, our two-step approach surpassed this with an UQC of 1.435. This modest improvement of approximately 1.5% is consistent with the superior performance observed during our 10-fold cross-validation.

[2] Altered by factor x for confidentiality reasons.


X-Learner - UQC: 1.5198
X-Learner Two-Step UQC: 1.5625
Random targeting
Cumulative incremental uplift (relative)
0.02000
0.01778
0.01556
0.01333
0.01111
0.00889
0.00667
0.00444
0.00222
0.00000
0
10
20
30
40
50
60
70
80
90
100
Percentage of customers targeted (%)

5 Discussion

In this study, we introduce a new uplift modeling theory challenging the assumption of the absence of ground truth currently prevalent in uplift modeling research (Radcliffe & Simpson, 2008). Specifically, this theory is based on the systematic absence of "sleeping dogs". We demonstrate that our proposed two-step approach leveraging this theory by utilizing established supervised machine learning algorithms, yields improved performance.

We found that in recent uplift modeling literature, most uplift modeling methods are not able to exceed the ATE on non-contractual datasets. This means there is only a negligible number of cases in which the treatment group's response rate is lower than the control group's response rate. Consequently, this evidence suggests a scarcity or even absence of sleeping dogs in non-contractual settings. Representatives of the fashion brand our dataset originates from confirmed to us that this is due to the minimal adverse reactions to marketing efforts such as coupons and discounts in their non-contractual settings, as these treatments present no potential downside for the customer.

Furthermore, we showed that the absence of sleeping dogs enables the identification of a subset of lost causes and sure things in the dataset. By using this information as ground truth, we partially changed the underlying problem type to a classification problem that can be addressed using supervised classification algorithms. We capitalized on this by building a lost cause classifier to identify the lost causes in a separate step before we apply an uplift modeling model. Our findings show that the UQC of the best performing uplift modeling algorithm on our dataset can be improved when using the lost cause classification to adjust the predicted uplift scores. The weighted F1-score of approximately 0.66 of the classifier leaves room for even more precise lost cause classification, suggesting that our approach can be further enhanced to yield even better performance.

In line with our initial assumption, our results validate that introducing partial ground truth to the uplift modeling problem enhances performance. Moreover, traditional uplift modeling algorithms may not be optimally suited for non-contractual settings. In these particular scenarios, our proposed method demonstrates superior performance, emphasizing the importance of tailoring methodologies to the specific characteristics of the problem domain.

Or study investigates Devriendt et al.'s (2018) statement about a possible absence of sleeping dogs in a dataset, extends it, and generalizes it to an entire category of marketing settings. This stands contrary to the assumption of four groups being present, which underlies uplift modeling research (Radcliffe & Simpson, 2008). Furthermore, it adds two new aspects to Devriendt et al.'s (2018) statement that, in uplift modeling, the classification is dependent on the campaign characteristics. First, our findings show that the campaign characteristic can have a significant impact on the uplift prediction, as it can lead to fewer customer groups and second, it can cause more information to be available during training, due to the inferred ground truth about lost causes and sure things.

In a practical dimension, our research mostly impacts marketing managers and data scientists as we provide a new model to create targeting policies. Our method increases the targeting performance, especially in non-contractual direct marketing contexts. Thus, to achieve optimal targeting performance via algorithmic targeting policies the campaign manager must first identify the setting in which the campaign will be conducted. The setting of non-contractual direct marketing provides a strong indication that no sleeping dogs are present within the targeted customer group, thus enabling two-step uplift modeling to provide maximum performance. Additionally, the data scientist creating sensitivity ranking for the targeting policy can examine the dataset for negative lift with traditional uplift modeling algorithms to assess the applicability of two-step uplift modeling.

When then applying two-step uplift modeling in a suitable environment, campaign managers can then expect higher response rates within the treated customer group as well as smaller overall treatment group sizes as evidenced by the evaluation above. Thus, both campaign conversions and contact costs are reduced, increasing the overall ROI of the campaign.

We first introduce the new theory of the absence of sleeping dogs in non-contractual direct marketing settings and indicate that building upon this theory increases targeting performance. This theory yields potential for further technical exploration, as well as application and evaluation in other marketing contexts and datasets. In terms of technical exploration, the assumption of ground truth introduces a large variety of new machine learning methods to the domain of uplift modeling. Therefore, many opportunities exist for fellow scholars to build upon the no sleeping dogs assumption by applying different machine learning methods and architectures to increase targeting performance.

We are aware that our research exhibits some limitations that provide opportunities for future research. First, the training and evaluation of the two-step approach was conducted on a single dataset in the fashion industry. Thus, generalizability of the findings is limited in the current state of evaluation. This presents the opportunity to apply and evaluate the two-step approach to more non-contractual datasets both within and beyond the fashion industry to benchmark performance and establish robustness of the novel approach.

Second, in this paper, we had to omit the classification of sure things as the infrequent frequency did not allow for accurate identification. The field of sure thing classification exhibits significant potential for evaluation on a different dataset. On a dataset with more occurrences of sure things, this classifier might however be feasible and further increase performance.

Third, our current methodology is limited to specific input data, namely binary treatment, and response variables. We encourage other researchers to adapt and expand this approach to accommodate continuous response variables and multiple treatment variables.

6 Conclusion

This study addressed a specific gap in uplift modeling within non-contractual direct marketing contexts. Our findings confirmed the negligible presence of sleeping dogs in such settings, which aligns with prior research from Devriendt et al. (2018). This confirmation allowed us to introduce the presence of ground truth for a designated subgroup, offering a potential resolution to the fundamental problem of causal inference for that group.

Based on this insight, we developed a new uplift modeling approach. When tested on a real-world dataset from a notable international fashion brand, our approach demonstrated superior performance compared to other current uplift modeling methods. This result highlights the importance of adapting modeling methodologies to specific characteristics of different marketing contexts.

In the broader context of data-driven marketing and its emphasis on personalized targeting strategies, uplift modeling has become increasingly vital for achieving improved ROI (Goldenberg et al., 2020). By contributing a new method and providing empirical evidence of its efficacy, our study offers a valuable resource for researchers and practitioners in the field.

In summary, our research provides both a deeper understanding of uplift modeling in non-contractual settings and a practical advancement in the methodology, aligning closely with the objectives set out in the introduction of this paper.

References

Arens, W. F., & Weigold, M. F. (2017). *Contemporary advertising and integrated marketing communications* (Fifteenth edition). McGraw-Hill Education.

Ascarza, E. (2018). Retention Futility: Targeting High-Risk Customers Might be Ineffective. *Journal of Marketing Research*, *55*(1), 80–98. https://doi.org/10.1509/jmr.16.0163

Athey, S., & Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *Stat*, *1050*(5), 1–26.

Chen, H., Harinen, T., Lee, J.-Y., Yung, M., & Zhao, Z. (2020). *CausalML: Python Package for Causal Machine Learning* (arXiv:2002.11631). arXiv. http://arxiv.org/abs/2002.11631

Chinchor, N., & Sundheim, B. M. (1993). MUC-5 evaluation metrics. *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. *Expert Systems with Applications*, *42*(22), 8403–8412. https://doi.org/10.1016/j.eswa.2015.06.054

De Caigny, A., Coussement, K., Verbeke, W., Idbenjra, K., & Phan, M. (2021). Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach. *Industrial Marketing Management*, *99*, 28–39. https://doi.org/10.1016/j.indmarman.2021.10.001

Devriendt, F., Moldovan, D., & Verbeke, W. (2018). A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics. *Big Data*, *6*(1), 13–41. https://doi.org/10.1089/big.2017.0104

Ellickson, P. B., Kar, W., & Reeder, J. C. (2022). Estimating Marketing Component Effects: Double Machine Learning from Targeted Digital Promotions. *Marketing Science*. https://doi.org/10.1287/mksc.2022.1401

Goldenberg, D., Albert, J., Bernardi, L., & Estevez, P. (2020). Free Lunch! Retrospective Uplift Modeling for Dynamic Promotions Recommendation within ROI Constraints. *Proceedings of the 14th ACM Conference on Recommender Systems*, 486–491. https://doi.org/10.1145/3383313.3412215

Gubela, R., Bequé, A., Lessmann, S., & Gebert, F. (2019). Conversion Uplift in E-Commerce: A Systematic Benchmark of Modeling Strategies. *International Journal of Information Technology & Decision Making*, *18*(03), 747–791. https://doi.org/10.1142/S0219622019500172

Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2014). A survey of personalized treatment models for pricing strategies in insurance. *Insurance: Mathematics and Economics*, *58*, 68–76. https://doi.org/10.1016/j.insmatheco.2014.06.009

Guido, G., Prete, M. I., Miraglia, S., & De Mare, I. (2011). Targeting direct marketing campaigns by neural networks. *Journal of Marketing Management*, *27*(9–10), 992–1006. https://doi.org/10.1080/0267257X.2010.543018

Gutierrez, P., & Gerardy, J.-Y. (2016). *Causal Inference and Uplift Modeling A review of the literature*.

Hansotia, B., & Rukstales, B. (2002). Incremental value modeling. *Journal of Interactive Marketing*, *16*(3), 35–46. https://doi.org/10.1002/dir.10035

Hartmann, W. R. (2010). Demand Estimation with Social Interactions and the Implications for Targeted Marketing. *Marketing Science*, *29*(4), 585–601. https://doi.org/10.1287/mksc.1100.0559

Hitsch, G. J., Misra, S., & Zhang, W. (2018). *Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation* (SSRN Scholarly Paper 3111957). https://doi.org/10.2139/ssrn.3111957

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

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

Kane, K., Lo, V. S. Y., & Zheng, J. (2014). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods. *Journal of Marketing Analytics*, *2*(4), 218–238. https://doi.org/10.1057/jma.2014.18

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Li, L., Li, X., Qi, W., Zhang, Y., & Yang, W. (2022). Targeted reminders of electronic coupons: Using predictive analytics to facilitate coupon marketing. *Electronic Commerce Research*, *22*(2), 321–350. https://doi.org/10.1007/s10660-020-09405-4

Liu, X. (2022). Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to Livestream Shopping. *Marketing Science*. https://doi.org/10.1287/mksc.2022.1403

Musalem, A., Bradlow, E. T., & Raju, J. S. (2008). Who's Got the Coupon? Estimating Consumer Preferences and Coupon Usage from Aggregate Information. *Journal of Marketing Research*, *45*(6), 715–730. https://doi.org/10.1509/jmkr.45.6.715

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift Modeling for preventing student dropout in higher education. *Decision Support Systems*, *134*, 113320. https://doi.org/10.1016/j.dss.2020.113320

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Potdar, K., S., T., & D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, *175*(4), 7–9. https://doi.org/10.5120/ijca2017915495

Radcliffe. (2007). Using Control Groups to Target on Predicted Lift: *Direct Marketing Analytics Journal*, 14–21.

Radcliffe, & Simpson, R. (2008). Identifying who can be saved and who will be driven away by retention activity. *Journal of Telecommunications Management*, *1*(2), 168–176.

Radcliffe, & Surry, P. (1999). Differential Response Analysis: Modeling True Responses by Isolating the Effect of a Single Action. *Credit Scoring and Credit Control IV*. https://www.research.ed.ac.uk/en/publications/differential-response-analysis-modeling-true-responses-by-isolati

Radcliffe, & Surry, P. D. (2011). Real-World Uplift Modelling with Significance-Based Uplift Trees. *Stochastic Solutions*.

Rossi, P. E., McCulloch, R. E., & Allenby, G. M. (1996). The Value of Purchase History Data in Target Marketing. *Marketing Science*, *15*(4), 321–340. https://doi.org/10.1287/mksc.15.4.321

Rößler, J., Guse, R., & Schoder, D. (2022). *The Best of Two Worlds–Using Recent Advances from Uplift Modeling and Heterogeneous Treatment Effects to Optimize Targeting Policies*.

Rößler, J., & Schoder, D. (2022). Bridging the Gap: A Systematic Benchmarking of Uplift Modeling and Heterogeneous Treatment Effects Methods. *Journal of Interactive Marketing*, *57*(4), 629–650. https://doi.org/10.1177/10949968221111083

Rößler, J., Tilly, R., & Schoder, D. (2021, January 5). *To Treat, or Not to Treat: Reducing Volatility in Uplift Modeling Through Weighted Ensembles*. https://doi.org/10.24251/HICSS.2021.193

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Simester, D., Timoshenko, A., & Zoumpoulis, S. I. (2020). Targeting Prospective Customers: Robustness of Machine-Learning Methods to Typical Data Challenges. *Management Science*, *66*(6), 2495–2522. https://doi.org/10.1287/mnsc.2019.3308

Smith, A. N., Seiler, S., & Aggarwal, I. (2023). Optimal Price Targeting. *Marketing Science*, *42*(3), 476–499. https://doi.org/10.1287/mksc.2022.1387

Sołtys, M., Jaroszewicz, S., & Rzepakowski, P. (2015). Ensemble methods for uplift modeling. *Data Mining and Knowledge Discovery*, *29*(6), 1531–1559. https://doi.org/10.1007/s10618-014-0383-9

Strycharz, J., Van Noort, G., Helberger, N., & Smit, E. (2019). Contrasting perspectives – practitioner's viewpoint on personalised marketing communication. *European Journal of Marketing*, *53*(4), 635–660. https://doi.org/10.1108/EJM-11-2017-0896

Yoganarasimhan, H., Barzegary, E., & Pani, A. (2023). Design and Evaluation of Optimal Free Trials. *Management Science*, *69*(6), 3220–3240. https://doi.org/10.1287/mnsc.2022.4507