# A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics

**Article** · March 2018

DOI: 10.1089/big.2017.0104

CITATIONS
0

READS
288

**3 authors:**

Floris Devriendt
Vrije Universiteit Brussel
**3** PUBLICATIONS **23** CITATIONS
SEE PROFILE

Darie Moldovan
Babeş-Bolyai University
**16** PUBLICATIONS **36** CITATIONS
SEE PROFILE

Wouter Verbeke
Vrije Universiteit Brussel
**54** PUBLICATIONS **815** CITATIONS
SEE PROFILE

Big Data
Volume 6 Number 1, 2018

DOI: 10.1089/big.2017.0104

ORIGINAL ARTICLE

# A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics

Floris Devriendt,[1,*] Darie Moldovan,[2] and Wouter Verbeke[1]



**Abstract**

Prescriptive analytics extends on predictive analytics by allowing to estimate an outcome in function of control variables, allowing as such to establish the required level of control variables for realizing a desired outcome. Uplift modeling is at the heart of prescriptive analytics and aims at estimating the net difference in an outcome resulting from a specific action or treatment that is applied. In this article, a structured and detailed literature survey on uplift modeling is provided by identifying and contrasting various groups of approaches. In addition, evaluation metrics for assessing the performance of uplift models are reviewed. An experimental evaluation on four real-world data sets provides further insight into their use. Uplift random forests are found to be consistently among the best performing techniques in terms of the Qini and Gini measures, although considerable variability in performance across the various data sets of the experiments is observed. In addition, uplift models are frequently observed to be unstable and display a strong variability in terms of performance across different folds in the cross-validation experimental setup. This potentially threatens their actual use for business applications. Moreover, it is found that the available evaluation metrics do not provide an intuitively understandable indication of the actual use and performance of a model. Specifically, existing evaluation metrics do not facilitate a comparison of uplift models and predictive models and evaluate performance either at an arbitrary cutoff or over the full spectrum of potential cutoffs. In conclusion, we highlight the instability of uplift models and the need for an application-oriented approach to assess uplift models as prime topics for further research.

**Keywords:** uplift modeling; prescriptive analytics; literature survey; experimental evaluation; performance measures; profit-driven analytics

## Introduction

Uplift modeling is a type of predictive modeling[1] that aims at predicting the net effect of performing some action on a certain outcome. So rather than the outcome, as in traditional predictive modeling, it is the difference in outcome as a result of an action that is estimated. This facilitates a further optimization of decision-making across various application fields in business as well as beyond. As it becomes apparent from the literature survey that is provided in this article, in recent years, uplift modeling has attracted a growing interest from the scientific community as well as from practitioners in the industry. Most of the developments that are proposed in the literature take place in the field of direct marketing, where uplift modeling is applied to optimize targeted marketing campaigns in terms of selected customers as well as the design of campaign. From a business perspective, uplift modeling can assist in achieving higher incremental sales and response rates than traditional predictive techniques, in turn leading to higher profits. However, it is to be noticed that current uplift modeling approaches, discussed in the next section, do not fully consider and take into account the actual business objective, that

[1]*Faculty of Economic and Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Brussels, Belgium.*
[2]*Business Information Systems Department, Babeş-Bolyai University, Cluj-Napoca, Romania.*

*Address correspondence to: *Floris Devriendt, Faculty of Economic and Social Sciences and Solvay Business School, Vrije Universiteit Brussel, Brussels 1050, Belgium,* E-mail: floris.devriendt@vub.be

is, maximizing the returns. Therefore, a stronger alignment of uplift modeling approaches with the actual business application's characteristics, with the aim to maximize returns and to support optimized decision-making, offers opportunities to further boost the creation of business value. Uplift modeling has a broad use and applicability, and plays a pivotal role in the evolution from descriptive and predictive toward prescriptive analytics. Whereas describing the observed behavior yields basic understanding and predicting future behavior provides actionable information, the aim of prescriptive analytics is to further extend these analyses by indicating or prescribing how to influence or attain a preferred future state. Essentially, instead of telling what happens (descriptive analytics) or what will happen (predictive analytics), prescriptive analytics aims to tell what should be done to make things happen. Prescriptive analytics allows to simulate the future in function of control variables, as such allows to optimize and prescribe the settings of control variables that maximize (or minimize) the expected outcome. Uplift modeling is a core approach in the field of prescriptive marketing and customer analytics, as it allows gazing at the effects of a marketing campaign on the customer's behavior. An enterprise, given limited resources, could therefore determine which course of action is best to undertake, that is, optimal in terms of return on marketing investment.[2] Often there are budget constraints and when multiple treatments are considered, it is important to establish which treatment works best for each customer. This can be formulated as an optimization problem, where there is an objective (i.e., maximizing return on marketing investment or profitability) and a set of constraints (e.g., a limited marketing budget or a limited number of opportunities to target a customer).[1]

In this article, we aim to provide a structured and complete overview of the state-of-the-art in uplift modeling, as well as an experimental evaluation of the principal approaches that have been proposed in the literature, allowing to gain deeper insight into the practical usage, challenges, and shortcomings of the current approaches, as such identifying research gaps. Both the literature survey and the reported findings of the experiments substantially contribute to the field by (1) providing a convergence point to the scientific community in terms of bringing together the available body of literature, as well as reviewing and assessing the state-of-the-art, which allows (2) identifying key directions for further research. In addition, (3) in extension to this article and building on the work by researchers who have published their work, the full code of the experimental setup is provided as a digital annex. This allows fellow researchers to replicate and extend on the experiments that are reported on in this article. As such, we envisage to facilitate and spur both further research in the field of uplift modeling and a broader adoption by practitioners.

In this article, we first provide a broad overview of the literature by defining uplift modeling, providing an overview of the state-of-the-art uplift modeling approaches, and by discussing performance measures for evaluating uplift models. Performance measures are essential to predictive and prescriptive modeling and therefore explicitly discussed in this article. The second part of this article reports on the results of the experiments that were conducted. Full details on the formalized methodology and experimental design are provided in the Experimental Setup section, and in the Empirical Results section the results are presented and discussed. The Conclusion and Future Research section concludes by highlighting challenges and providing detailed directions for further research.

## Literature Survey

Developments in the field of uplift modeling have been spurred by, and are closely intertwined with, the main field of application, which is response modeling. Also in this article, the discussions and examples are mainly situated in the field of response modeling, although uplift modeling most definitely has a broader use as illustrated by a limited number of alternative example applications. Therefore, in this section, we first briefly explain traditional, that is, predictive response modeling before introducing uplift response modeling as an improved approach. Subsequently, the use of uplift modeling is more generally discussed. Then, an exhaustive and structured survey of the literature on uplift modeling is provided, and the principal uplift modeling approaches, experimentally evaluated in the second part of this article, are explained in detail. Finally, evaluation approaches for assessing the performance of uplift models are reviewed.

### Predictive versus uplift modeling

**Predictive response modeling.** Response modeling[3,4] essentially concerns the identification of (prospective) customers who are likely to respond when targeted by a direct marketing campaign. Notice that *responding* may have a different meaning depending on the type of marketing campaign, but typically concerns purchasing a product or subscribing to a service. Ideally, for evaluating the effect of a direct marketing campaign, the

customer base is split into two groups: a control group and a treatment group. The control group is a randomly selected subgroup of the population, that is, the customer base that will not be targeted by the campaign, whereas the treatment group is a subgroup of the population that is targeted by the campaign. The term *treatment* is generally used to refer to a specific action or set of actions and is characterized or described by control variables. Since in most uplift data sets, there is only one control variable, indicating whether or not a treatment was applied, this control variable is referred to as the treatment variable. The use of random control groups is a common practice in various fields and applications and not exclusive to uplift modeling. In A/B testing,[5] two or more groups are randomly selected from a homogeneous population and each group then is subjected to a different treatment. This allows contrasting and testing the difference in performance or outcome of both treatments. A/B testing is a common practice in website and application development, where changes and variations in the layout are tested on subgroups of visitors or users, as such allowing to select the optimal design. Also, in the medical sciences, a similar practice is adopted in clinical trials, for instance, to test or proof the effect of a new treatment or drug. First, two random groups of test subjects are selected, one of which then receives the experimental treatment or drug and the other group receives a placebo. Again, this allows comparing and statistically testing the difference in outcome.

In marketing, the outcome or behavior of the customers in the treatment and control groups is observed in terms of response. This allows calculating and comparing the response rate for both groups, which is equal to the proportion of the number of customers who respond over the number of customers in the group. A campaign is considered successful if it succeeds in boosting the response rate of the treatment group compared with the response rate of the control group. The difference in response rate is the *uplift* due to the campaign. To further increase the returns of future similar direct marketing campaigns, a predictive response model can be developed using data related to the treatment group of the previous campaign. A *traditional* predictive response model estimates the probability to respond for customers targeted in the campaign. Such a response model can be used to select the customers with the highest probability to respond and therefore to be targeted in a new campaign, which should yield an improvement in the return on marketing investment, by generating an additional uplift in response rate due to selecting customers using the response model.

However, the development and adoption of traditional predictive response models lead to selecting and targeting customers who are similar to customers targeted in a previous campaign and were observed to respond. However, as demonstrated by several researchers,[6–8] by doing so, also customers who would respond when not targeted, so-called baseline responders, will be selected. Instead, only customers should be targeted who are likely to respond *because* of the marketing campaign, but unlikely to respond otherwise. These can be called *true* responders. Hence, an additional distinction between *true* versus *baseline* responders is required when developing a response model. The general drawback of predictive response models is that they are not designed to estimate net response or uplift and to maximize incremental impact. As pointed out, the core of the problem concerns the objective function that is adopted in learning a predictive response model.[6,8] The objective function in predictive analytics concentrates on the probability of response, which is not an indicator of the incremental effect but rather of gross response, whereas *uplift modeling* estimates the increase in probability of response,[8] that is, the *change* in behavior, because of the marketing campaign.

Uplift response modeling. A customer base can be classified along two dimensions (Fig. 1), that is, in function of response (yes or no) and treatment (yes or no), resulting in four groups[9]:

(1) *Sure things:* Customers who always respond (also known as *always-buy*). Targeting *sure things* does not generate additional returns but does generate additional costs, that is, the fixed costs of contacting a customer and possibly a cost related to a financial incentive offered to the targeted

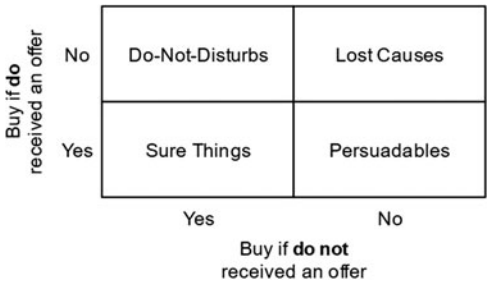


**FIG. 1.** Classification of customer base along two dimensions: response and treatment. Adapted from Kane et al.[9] and Siegel.[10]

customers. For example, coupons that are send to customers offering reduced prices aiming at convincing customers to purchase certain products.

(2) *Lost causes:* Customers who never respond (also known as *never-buy*). Similar to *sure things*, targeting *lost causes* will not generate additional revenues, but does involve additional costs. However, the additional costs are lower than for including *sure things*. Since *lost causes* do not respond, they do not take advantage of the offer, which would involve an additional cost.

(3) *Do-not-disturbs:* Customers who do not respond *only because* of a treatment (also known as *anti-persuadables* or *sleeping dogs*). They respond if not treated, but do not respond if they are treated. For example, customers who are targeted in retention campaigns can be triggered to churn because of the campaign and withdraw from current products or services. Hence, the opposite effect of what is aimed for is achieved. Clearly, including *do-not-disturbs* in a campaign generates no additional revenues but comes with large additional costs. When many do-not-disturbs are included in the campaign, it could even be better, in terms of returns, to not run a campaign at all.

(4) *Persuadables:* Customers who respond *only because* they are exposed to a campaign (also known as *influencables*). These are the net-responders who should be treated. They respond only when targeted and make the campaign to generate additional revenues, and as such a net profit, after subtraction of the costs generated by including the other types of customers. *Persuadables* are exactly the customers who need to be selected and who should be identified by an uplift model.

Uplift modeling aims at identifying *persuadables*, and at the same time aims at avoiding the treatment of *do-not-disturbs*. This classification, however, is campaign dependent. For instance, it is possible for one user to be a *lost cause* when the campaign offers a 5% reduction at a next purchase. However, when the campaign offers a 20% reduction at a next purchase, that same customer might be a *persuadable*. In other words, the classification depends on the campaign characteristics. Uplift modeling has been applied for response modeling to improve the effectiveness of direct marketing campaigns,[6,8,11] as well as in personalized medicine,[12] to determine which treatment a patient should receive. In clinical trial analysis, uplift modeling is capable of selecting subgroups of patients, or even individual patients, who benefit from an alternative treatment, even if the alternative treatment overall is worse than the standard treatment.[12] Uplift modeling has also been used for selecting customers to include in retention campaigns and as such to maximize the effect of a retention campaign, by identifying customers whose probability to churn is most reduced by offering an incentive to remain loyal.[13] Uplift modeling has also been adopted in politics,[14,15] for identifying swing voters to target in election campaigns. An overview of various applications of uplift modeling is provided in the work of Siegel.[10]

Uplift modeling is closely related to the aforementioned A/B testing and clinical trials, as in those approaches as well randomized groups are selected and receive a different treatment. The observed behavior for the different groups allows uplift modeling techniques to identify and predict the incremental effect of the treatment on the individual entity level. On the contrary, A/B testing is oriented toward decision-making at the group or segment level, by assessing whether a treatment performed (significantly) better. As an example, in clinical trials, typically the objective is to establish whether an experimental drug performs significantly better than the placebo in terms of observed outcome. Typically, such trials and A/B testing disregard subject heterogeneity and do not intend to predict the effect of a treatment at the individual level, for example, at the individual patient or website-visitor level. Experimental medicine that proves to be effective over the entire population may be ineffective or even detrimental for patients with certain conditions.[16] In contrast, uplift modeling is focused on decision-making at the individual level. It aims at supporting the selection of the most appropriate treatment for each individual customer, regardless whether or how well that treatment performs on the full customer base.

In general, uplift modeling is about estimating the causal effect of an action or treatment on an outcome and therefore allows determining which action to take or treatment to apply to optimize the outcome, that is, the result or effect of the action or treatment. In other words, uplift modeling allows to optimize the level or setting of control variables, concerning factors (such as actions to take or treatments to apply) that can be controlled or (to some extent) be decided about, and therefore can be optimized. Uplift models in essence allow simulating the future for various scenarios, with a scenario being defined in terms of the involved control variables. Clearly, this may have practical use in a wide variety of application settings. In line with the literature and given the available data sets for experimental

evaluation, the focus in this article is on optimizing the target population for marketing campaigns, that is, response modeling.

Let us assume the customer base is randomly divided into two groups, defined as a treatment group and a control group. A customer is either in the treatment group, that is, has been targeted in the campaign, or is in the control group, that is, has not been targeted. As a formal definition, let $X$ be a vector of independent or predictor variables, $X=\{x_1,\ldots, x_n\}$, and let $Y$ be the binary dependent or target variable, $Y\in\{0,1\}$, with $Y=1$ indicating response and $Y=0$ no response. In addition, the control or treatment variable $T$ denotes whether or not a customer is in the treatment group, $T=1$, or in the control group, $T=0$. Finally, $P$ denotes a probability as estimated by a model. Uplift is then defined as the probability of a customer to respond if treated minus the probability of the customer to respond when not treated, or:

$$U(x_i) := P(y_i=1|x_i; t_i=1) - P(y_i=1|x_i; t_i=0) \quad (1)$$

Hence, uplift is the difference in behavior, that is, the impact of the treatment.

Compared to predictive modeling, treatment and control groups are required for developing uplift models. This additional data *dimension*, however, may cause a new type of imbalance in the data set. In traditional predictive modeling often the class distribution is imbalanced, for instance, in response modeling often many more nonresponders are observed than there are responders. The imbalanced class distribution may cause predictive models to perform poorly, which is often addressed by applying sampling techniques or cost-sensitive learning approaches. In uplift modeling, we may have an additional imbalance in terms of the number of control group versus treatment group observations. Although this second type of imbalance in uplift modeling has not extensively been discussed or the impact experimentally evaluated in the literature, some uplift approaches do implicitly take it into account.[8,9]

## Literature survey on uplift modeling

A complete overview of research articles on uplift modeling is provided in Tables 1–3. For each article, the table summarizes the type of application, the applied techniques together, details on the data set, and the preprocessing as well as evaluation procedure that was applied. The table does not include a number of white papers on uplift modeling that have been published by companies, since they do not provide exhaustive and precise details on the data that were collected and analyzed, on the applied techniques or the reported results.[10,17,18] Uplift modeling in the literature is referred to as *differential response analysis*,[19] *true-lift modeling*,[6,9] *true response modeling*,[19] *net lift modeling*,[20] *persuasion modeling*,[14] *differential marketing*,[19] *incremental value modeling*,[21] *incremental impact modeling*,[21] and *personalized treatment selection*.[16] The term uplift modeling is most commonly used and therefore adopted in this article.

As can be seen from the table, in the majority of the reported articles, either tree- or regression-based approaches are adopted for developing uplift models. Tree-based approaches segment a population in smaller, homogeneous groups. This aligns with the objective of uplift modeling, that is, segmenting the population into the four groups identified in Figure 1. Regression on the contrary is a widely popular approach in science and industry, given the predictive strength and interpretability of the resulting model.

## Uplift modeling techniques

In this section, a detailed review of uplift modeling techniques is provided. The approaches are grouped into data preprocessing and data processing approaches. Whereas data preprocessing approaches essentially adopt traditional predictive analytics in an adapted setup for learning an uplift model, the data processing approaches concern adapted predictive analytics for developing uplift models. Within the data preprocessing approaches, we further distinguish between transformation approaches that redefine the target variable and approaches that essentially concern predictor variable selection procedures. Within the data processing approaches, a further differentiation is made between indirect and direct estimation approaches.

### Data preprocessing approaches.

*Transformation approaches.* In the Uplift Response Modeling section, the customer population is categorized into four groups based on whether a customer responds when treated or not treated: sure things, lost causes, persuadables, and do-not-disturbs (Fig. 1). Preferably, we would like to know for each customer individually to which group he or she belongs, which would allow us to treat all persuadables and to maximize the returns of a campaign. However, we do not have this information; we therefore develop uplift models with the aim of identifying the persuadables. What we do know, based on previous campaign or experimental data that were gathered to build an uplift model, is whether a customer was treated and whether the customer responded. Hence, a customer can be grouped into one of the following

**Table 1. Overview of literature on uplift modeling: part 1**

| *Authors* | *Title and journal* | *Year* | *What?* | *Techniques* | *Data set-# cust. -# vars. - public (1) or private (2) - type* | *Metrics - sampling - variable selection - validation* |
|---|---|---|---|---|---|---|
| Radcliffe and Surry | Differential Response Analysis: Modeling True Response by Isolating the Effect of a Single Action—*Credit Scoring and Credit Control* | 1999 | Introduction to differential response analysis along with explanation of decision tree algorithm | Tree-based approach | Telephone company - 100,000 cust. - unknown vars. - (1) - Development | Sign-up rate - no sampling - no var. selection - unknown |
| Chickering and Heckerman | A Decision Theoretic Approach to Targeted Advertising—*Uncertainty in Artificial Intelligence Proceedings* | 2000 | Introduces slightly modified decision tree to select the customers who would create more revenue | Tree-based approach | Microsoft network subscriptions - 110,000 - 15 - (2)- Development | Expected revenue and expected lift probability - no sampling - no var. selection - 70% train/30% test |
| Hansotia and Rukstales | Incremental Value Modeling—*Journal of Interactive Marketing Volume 16 Issue 3* | 2002 | Builds models to estimate incremental response and compares a modified CHAID with a logistic regression approach | Tree-based approach | Holiday promotion from unspecified major national retailer - 282277 - unknown vars. - (2) - Development | Incremental response rate - no sampling - no var. selection - cross-validation |
| Lo | The True Lift Model—A Novel Data Mining Approach to Response Modeling in Database Marketing - *ACM SIGKDD Explorations* | 2002 | Comparison of traditional response modeling methodology versus new methodology | Logistic regression, Neural network and Naïve Bayes | Simulated data set - 100,000 - 4 - (1) - Development | True lift, Lift chart - no sampling - no var. selection - Hold-out validation |
| Lai | Influential Marketing: A New Direct Marketing Strategy Addressing The Existence Of Voluntary Buyers—*Master Thesis of Computer Sience at Simon Fraser University* | 2006 | Comparison of traditional response modeling versus Lo's approach[6] and newly introduced four-quadrant approach | Association rule classifier, Decision tree, Logistic regression | CIBC financial - 304698 - 407 - (2) - Development | Positive influence curve, Undecided buyer rate - decision tree: undersampling; ARC: over-sampling - no var. selection - threefold Cross-validation |
| Radcliffe | Using Control Groups to Target on Predicted Lift: Building and Assessing Uplift Models—*Direct Marketing Analytics Journal* | 2007 | Compares tree-based approach versus Lo's approach[6]) on three different data sets and introduces the Qin coefficient metric. | Tree-based and regression-based approach | (i) Catalog mailing from retailer - 100,000 - unknown - unknown - (2) - Development (Deep-selling) (ii) Mobile phone company - unknown - unknown - unknown - (2) - Retention (iii) Bank sector - unknown - unknown - unknown - (2) - Development (cross-selling) | Qini coefficient - Unknown - Unknown |
| Rzepakowski and Jaroszewicz | Decision Trees for Uplift Modeling—*Proceedings of the 10th International Conference on Data Mining (ICDM)* | 2010 | Introducing a tree-based approach and a comparison with existing tree-based approaches. | Tree-based | Several data sets from UCI Repository - unknown - 6 to 61 - (1) - Clinical | AUUC - No sampling - Simple heuristic - 2×5 cross-validation |

AUUC, area under uplift curve; CHAID, chi-square automatic interaction detection; CIBC, Canadian Imperial Bank of Commerce; UCI, University of California, Irvine.

**Table 2. Overview of literature on uplift modeling: part 2**

| *Authors* | *Title and journal* | *Year* | *What?* | *Techniques* | *Data set-# cust. -# vars. - public (1) or private (2) - Type* | *Metrics - sampling - variable selection - validation* |
|---|---|---|---|---|---|---|
| Radcliffe and Surry | Real-World Uplift Modelling with Significance-Based Uplift Trees—*Stochastic Solutions White Paper* | 2011 | Well-documented explanation of uplift modeling and a description of the significance-based uplift trees. | Tree-based | Synthetic - 64,000 - (2) - Development | Qini - Bagging - Pessimistic Qini-based - unknown |
| Rzepakowski and Jaroszewicz | Uplift Modeling In Direct Marketing—*Knowledge and Information Systems* | 2012 | Tree-based techniques for uplift modeling that are used on real marketing data and compared with traditional response models and existing uplift modeling techniques. | Tree-based | Online merchandise - 64,000 - 12 - (1) - Development | Cumulative percent of total visits, AUUC - unknown - no var. selection - 10×10 cross-validation |
| Rzepakowski and Jaroszewicz | Decision Trees for Uplift Modeling With Single And Multiple Treatments—*Knowledge and Information Systems* | 2012 | Explaining a tree-based approach and applying it on both a single treatment and multiple treatments (to choose from). | Tree-based | Several data sets from UCI Repository - unknown - 6 to 61 - (1) - Clinical | AUUC - no sampling - var. selection with simple heuristic - 2×5 cross-validation |
| Jaśkowski and Jaroszewicz | Uplift modeling for clinical trial data—*ICML 2012 Workshop on Machine Learning for Clinical Data Analysis* | 2012 | An approach that allows for application of standard classification models built on the treatment and control data sets similar to semisupervised learning to improve accuracy. | Logistic regression | (i) Bone marrow transplant data (ii) tamoxifen data (iii) hepatitis data - unknown - unknown - (1) - clinical | AUUC - no sampling - no var. selection - 10×10 cross-validation |
| Zaniewicz and Jaroszewicz | Support Vector Machines for Uplift Modeling—*The First IEEE ICDM Workshop on Causal Discovery* | 2013 | Introduction of an uplift SVM that classifies instances in negative, positive, or neutral. Comparison with existing techniques. | Support vector machine | (i) Online merchandise (ii) bone marrow transplant (iii) tamoxifen - (i) 64,000 (ii–iii) not mentioned - (i) 12 (ii–iii) not mentioned - (1) - (i) Development (ii–iii) Clinical | AUUC - no sampling - no var. selection - 128×80/20 training/test |
| Kane et al. | Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods—*Journal of Marketing Analytics* | 2014 | Literature review of regression approaches along with the proposal of a new method and metric. Comparison of new and current techniques. | Gradient-boosted decision tree and multinomial logistic regression | (i) Financial services (ii) Online merchandise (iii) Retail office supplies - (i) 1,000,000 (ii) 64,000 (iii) 435,000 - (i) unknown (ii) 12 (iii) unknown - (i) (2) (ii) (1) (iii) (1) (i–iii) Development | Gini, Gini Top 15%, Gini repeatability - (i–iii) no sampling - (i–iii) no var. selection - Hold-out cross-validation |
| Guelman et al. | Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study—*Research Group on Risk in Insurance and Finance (Working Paper)* | 2014 | Introduces random forests for uplift modeling, analyzes existing techniques, and benchmarks them in R | Random forests, logistic regression | Royal Bank of Canada, RBC Insurance - 24362 - 50 - (2) - Retention, Development | Spearman's correlation rank, average treatment effect - random sampling - LASSO (10×cross-validation) - 70/30 training/test |
| Soltys et al. | Ensemble methods for uplift modeling—*Data Mining and Knowledge Discovery* | 2014 | Applies the idea of an ensemble of methods to uplift modeling and compares it with some other techniques in the literature. | Ensemble of tree-based techniques | (i) Online merchandise (ii) bone marrow transplant (iii) tamoxifen (iv) survival (v) Kmsurv - (i) 64,000 (ii–v) not mentioned - (i) 12 (ii–v) not mentioned - (1) - (i) Development (ii–v) Clinical | AUUC - no sampling - var. selection with simple heuristic - 128×80/20 training/test |

**Table 3. Overview of literature on uplift modeling: part 3**

| *Authors* | *Title and journal* | *Year* | *What?* | *Techniques* | *Tech- Data set-# cust. -# vars. - public (1) or private (2) - Type* | *Metrics - sampling - variable selection - validation* |
|---|---|---|---|---|---|---|
| Kuusisto et al. | Support Vector Machines for Differential Prediction—*Machine Learning and Knowledge Discovery in Databases* | 2014 | Applies several SVM-models, designed to maximize uplift, on medical data | SVM | (i) Breast cancer (ii) COX-2 - (i) 907 (ii) 3920 - (i) 55 (ii) Not mentioned - (2) – Clinical | AUUC - no sampling - no var. selection - 10 cross-validation |
| Shaar et al. | Pessimistic Uplift Modeling—*ACM SIGKDD* | 2016 | Introduces a novel uplift modeling technique to deal with the issue where uplift models are highly sensitive to noise and disturbance | Tree-based | (i) Splice (ii) breast cancer (iii) tamoxifen (iv) online merchandise (v) bone marrow transplant (vi) simulation - (i–iii) not mentioned (iv) 64,000 (v) not mentioned (vi) 200 - (i–iii) not mentioned (iv) 12 (v) not mentioned (vi) 20 - (1) - (i–iii) Clinical (iv) Development (v–vi) Clinical | Real uplift curve - no sampling - no var. selection - 100× repeated experiments |
| Cao et al. | Untangle Customers Incrementality Using Uplift Modeling with a Case Study on Direct Marketing—*MWSUG* | 2017 | Presents a case study using the Dummy treatment approach in combination with SAS software | Logistic regression | Furniture store - 427,559 - 50+ - (2) - Development | Uplift chart - unknown - unknown - 70/30 training/test |
| Zaniewicz and Jaroszewicz | $L_p$-Support vector machines for uplift modeling—*KAIS* | 2017 | Adapts SVM for uplift modeling | SVM | (i) Online merchandise (ii) bone marrow transplant (iii) tamoxifen (iv) survival (v) Kmsurv - (i) 64,000 (ii–v) not mentioned - (i) 12 (ii–v) not mentioned - (1) - (i) Development (ii–v) Clinical | AUUC - no sampling - var. selection with simple heuristic - 128× 80/20 training/test |
| Zhao et al. | Uplift Modeling with Multiple Treatments and General Response Types—*SIAM SDM* | 2017 | Introduces a new technique to handle the multiple treatment scenario | Tree-based | (i) Synthetic (ii) priority boarding data (iii) seat reservation data - (i) 32,000 (ii) 600,000 (iii) 765,216 - (i) 50 (ii) 9 (iii) 12 - (i) (1) (ii) (2) (iii) (2) - (i–iii) - Development | Modified uplift curve - unknown - no feat. selection - 50/30/20 training/validation/test |
| Zhao et al. | A Practically Competitive and Provably Consistent Algorithm for Uplift Modeling—*ICDM* | 2017 | Introduces an updated technique to handle multiple treatment scenario | Tree-based | (i) Synthetic (ii) Priority boarding data - (i) 32,000 (ii) 600,000 - (i) 50 (ii) 9 - (i) (1) 2) (2) - (i–ii) - Development | Modified uplift curve - unknown - no feat. selection - unknown |
| Michel et al. | Effective customer selection for marketing campaigns based on net scores—*JRIM* | 2017 | Introduces a new technique focused on statistical tests for the split criterion | Tree-based | Stocks - 125,636 - 34 - (2) - Development | Monetary metric - unknown - feat. selection - eightfold cross-validation |

four categories as discussed in Kane et al.[9] and Lai's article[22] and shown in Figure 2:

- Control responders (CR) are the customers who responded without having received a treatment. Note that we do not know whether a CR would have bought when treated. Therefore, we do not know whether a CR is either a sure thing or a do-not-disturb.
- Control nonresponders (CN) are the customers who did not receive a treatment and did not respond. Because they might have responded had they been treated, we cannot tell whether a CN is either a lost cause or a persuadable.
- Treatment responders (TR) are customers who were treated and responded. They might or might not have responded without treatment; therefore, a TR can be either a persuadable or a sure thing.
- Treatment nonresponders (TN) are customers who received a treatment but did not respond. They might be either lost causes or do-not-disturbs because they might or might not have responded had they not received treatment.

Lai's approach then labels CN and treated responders as positive targets, because both groups together contain all persuadables and do not contain do-not-disturbs, who should not be treated. In addition to persuadables, these groups (CN and TR) also contain lost causes and sure things. These groups should optimally not be treated, but only involve a minor cost compared with treating do-not-disturbs. CR and treated nonresponders are labeled negative targets because these groups only contain do-not-disturbs, lost causes, and sure things. Targeting these groups always involves costs but never generates additional revenues. Hence, CR and treated nonresponders should not be treated. As such, the target variable is transformed into a binary target variable, and the uplift modeling problem is converted into a binary classification problem to which logistic regression or any other classification technique can be applied. The uplift or net effect of a treatment is then obtained by subtracting the probability to be negative from the probability to be positive, as estimated by the resulting model:

$$Uplift_{Lai}(x) = P(TR \text{ OR } CN|x) - P(TN \text{ OR } CR|x) \quad (2)$$

Alternatively, a multiclass classification model can be developed to estimate CR, CN, TR, or TN group membership, and aggregate the resulting probability estimates to calculate uplift as follows, as proposed by Kane et al.[9]:

$$Uplift_{Kane}(x) = [P(TR|x) + P(CN|x)] - [P(TN|x) + P(CR|x)] \quad (3)$$

Finally, as shown by Kane et al.,[9] to correct for a bias introduced in the uplift formula in case of unequal control and treatment group sample sizes (as mentioned in the

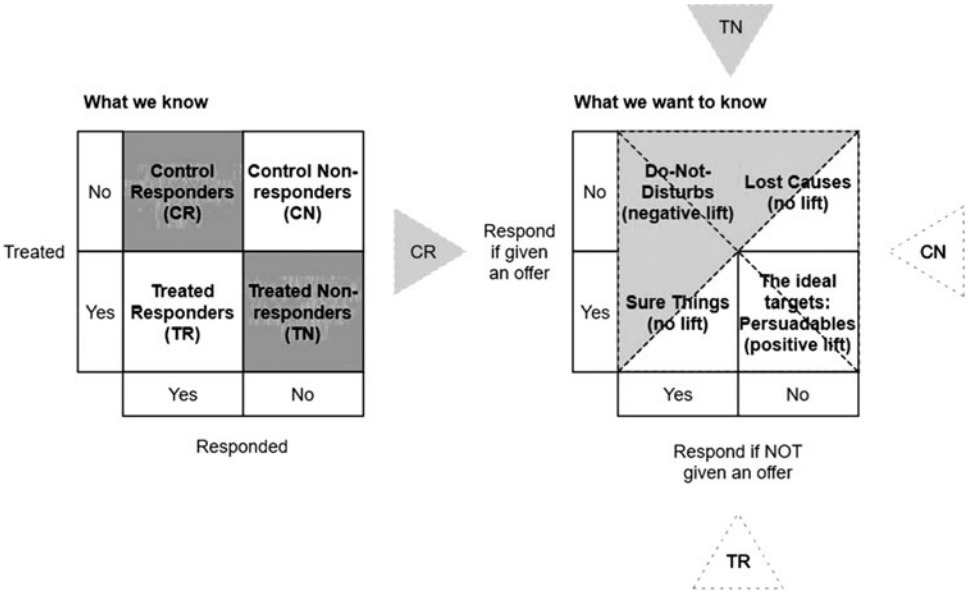


**FIG. 2.** Conceptual table. Adapted from Kane et al.[9]

Uplift Response Modeling section), *Kane's variation* of Lai's approach needs to be corrected as follows, leading to the *generalized Lai approach*[9]:

$$Uplift_{GeneralizedKane}(x) = \frac{P(TR|x)}{P(T)} + \frac{P(CN|x)}{P(C)} - \frac{P(TN|x)}{P(T)} - \frac{P(CR|x)}{P(C)} \quad (4)$$

Shaar et al.[23] state that uplift modeling is unstable and sensitive to noise and disturbance, and develop a technique based on Lai's approach to consider the noise in the data. Just like stated by Kane et al.,[9] they have weighted Lai's approach, although slightly different:

$$Uplift_{WeightedLai}(x) = P(Positive \mid x) * P\left(\frac{positive}{population}\right) - P(Negative \mid x) * P\left(\frac{negative}{population}\right) \quad (5)$$

Afterward, instead of predicting the probability a person will respond given a treatment, they predict the probability of a person being treated given he or she has responded.[23] To do this they take on a two-model approach, one model $M_r$ for the responders and one model $M_n$ for the nonresponders. Once both models have been developed they calculate the *reflective uplift* as follows:

$$Uplift_{Reflective}(x) = P_{Reflective}(Positive|x) - P_{Reflective}(Negative|x) \quad (6)$$

With $P_{Reflective}$ being calculated, using both models as follows:

$$P_{Reflective}(Positive|x) = P_{M_r}(T|R) * P(TR) + P_{M_n}(C|N) * P(CN) \quad (7)$$

$$P_{Reflective}(Negative|x) = P_{M_n}(T|N) * P(TN) + P_{M_r}(C|R) * P(CR) \quad (8)$$

Then, finally the reflective uplift and Lai's weighted approach are combined to get the *Pessimistic Uplift* to have more precision, robustness, and reliability[23] as follows:

$$Uplift_{Pessimistic} = \frac{1}{2} * (Uplift_{WeightedLai} + Uplift_{Reflective}) \quad (9)$$

The $Uplift_{Reflective}$ works as a stabilizer on the $Uplift_{WeightedLai}$'s approach. The authors report similar performance when compared to *causal conditional inference tree* as discussed below.

A similar target variable transformation was proposed by Jaśkowski and Jaroszewicz.[12] The target variable is transformed by defining a new target variable $Z \in \{0, 1\}$ as follows:

$$Z = \begin{cases} 1 & \text{if treated and } Y = 1. \\ 1 & \text{if not treated and } Y = 1. \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

with 1 being the positive outcome. The resulting model then allows estimating uplift as the difference between the success probabilities in treatment and control groups:

$$Uplift_{Jaskowski} = P(response|X, treament) - P(response|X, control) = 2P(response|X) - 1 \quad (11)$$

Again, the uplift modeling problem is converted into a binary classification problem allowing to adopt any traditional classification technique. The approach has been tested on three medical data sets. The first data set has information on two types of bone marrow transplant, the second focuses on treatment of breast cancer with tamoxifen, and the last looks at whether or not steroids were used to treat patients with hepatitis. When applying the above approach for uplift modeling, the authors indicate uplift approaches are applicable to medical sets to find subgroups of patients where the treatment is beneficial.[12] This approach is referred to as the *Modified Outcome Method*,[24] whereas in Pechyony et al.[25] an extension is made to simultaneously maximize the uplift *and* to maximize the response is presented in the context of online advertising.

*Variable selection procedures.*

*Treatment dummy approach.* In Lo,[6] an uplift modeling approach is proposed, which groups the treatment and the control group into a single development sample for estimating a response model. The *group origin*, that is, control or treatment group, is indicated by a treatment dummy variable $T$, with $T=1$ for observations from the treatment group and $T=0$ for observations from the control group. A response model is then developed with candidate predictor variables $X$, $T$, and $X^*T$, with $X$ the predictor variables, $T$ the treatment dummy variables, and $X^*T$ interaction variables. Although this setup can be adopted in combination with any predictive modeling approach, the original approach estimates a logistic regression model as follows:

$$P_i = E(Y_i|X_i) = \frac{exp(\alpha + \beta' X_i + \delta T_i + \gamma' X_i T_i)}{1 + exp(\alpha + \beta' X_i + \delta T_i + \gamma' X_i T_i)} \quad (12)$$

where $\alpha$ represents the intercept, $\beta$ denotes the main effects of the independent predictor variables, $\delta$ captures the main treatment effects, and $\gamma$ represents the

additional effects of the predictor variables due to the treatment.[6] Then, uplift can be calculated as follows:

$$Uplift_{Lo} = P(response|campaign) - P(response\,|\,control) \quad (13)$$

$$= \frac{exp(\alpha+\gamma+\beta' X_i+\delta' X_i)}{1+exp(\alpha+\gamma+\beta' X_i+\delta' X_i)} - \frac{exp(\alpha+\beta' X_i)}{1+exp(\alpha+\beta' X_i)} \quad (14)$$

The predicted uplift is essentially the difference of the probability a user would respond when treated (i.e., receiving the campaign) minus the probability a user would respond when not treated. Drawbacks of this approach are the potentially large compound errors resulting from subtracting two model scores, and the multicollinearity between predictor variables serving both as baseline and interaction variable,[9] potentially leading to instability and overfitting. Another example where this approach has been applied on a case study in direct marketing can be found in Cao et al.[26]

*Net weights of evidence and net information value.* The net weights of evidence (NWOE) and net information value (NIV) as proposed by Larsen[20] do not present a stand-alone uplift modeling approach, but can be adopted in combination with any other uplift modeling approach for coding and selecting predictor variables, respectively. The NWOE and NIV are adaptations for uplift modeling of weights of evidence (WOE) and information value (IV), as used in predictive analytics for categorizing, coding, and selecting predictor variables.[27] The WOE describes the relationship between a predictor variable $X$ and the target variable $Y$, whereas the IV describes the strength of the relationship.[20] The WOE of a value $i$ of variable $X$ is defined as follows:

$$WOE_{X=i} = ln\left(\frac{P(X=i\,|\,Y=1)}{P(X=i\,|\,Y=0)}\right) \quad (15)$$

The WOE indicates the differential effect in terms of log-odds of a variable $X$ taking value $i$, compared with the average log-odds observed for the full sample, since

$$ln\left(\frac{P(Y=1|X=i)}{P(Y=0|X=i)}\right) = ln\left(\frac{P(Y=1)}{P(Y=0)}\right) + ln\left(\frac{P(X=i|Y=1)}{P(X=i|Y=0)}\right) \quad (16)$$

The IV is the weighted sum of the WOE values and expresses the strength of the relationship between predictor variable $X$ and target variable $Y$:

$$IV = \sum_{i=1}^{n} (P(X=i|Y=1) - P(X=1|Y=0)) \times WOE_i \quad (17)$$

The WOE and IV can be used to select predictor variables. In Larsen,[20] they have been extended for application in uplift modeling to the NWOE and NIV. Notice that net, here, refers to netlift modeling,[20] as a synonym of uplift modeling. The NWOE is defined as the difference in the WOE of the treatment and control group:

$$NWOE = WOE_T - WOE_C \quad (18)$$

The NIV is defined as follows:

$$NIV = 100 \times \sum_{i=1}^{n} (P(X=i|Y=1)_T P(X=i|Y=0)_C - P(X=i|Y=0)_T P(X=i|Y=1)_C) \times NWOE_i \quad (19)$$

The NIV allows to select predictor variables with strong uplift prediction. However, the NIV as defined in the equation above has been found not to be robust, that is, is found to be dependent on the train set of observations used to calculate the NWOE and NIV. Therefore, the NIV is penalized for instability, with the penalized NIV equal to the NIV minus the sum of the differences between the NWOE on the train and validation set.

## Data processing approaches.

*Two-model approach.* Uplift modeling approaches can be categorized into indirect estimation approaches and direct estimation approaches. Whereas direct estimation approaches develop models that directly estimate uplift, a simple and intuitive approach to uplift modeling is to create two separate predictive models, one for the treatment group, $M_T$, and one for the control group, $M_C$. Uplift can then be estimated by subtracting the estimate from the control group model from the estimate from the treatment group model.

This approach is called the two-model approach and has the significant benefit of being straightforward to implement and allowing to adopt traditional predictive modeling approaches, such as logistic regression or decision trees. However, the two-model approach has been shown to only work well in the simplest of cases.[21,28] Radcliffe and Surry[8] offer a detailed and illustrative explanation on why this approach typically does not work well in practice. A key concern is that both classification models are built independently on the treatment and the control group, disregarding one another, and estimating uplift by subtracting the estimates resulting from both models. Notice that as a result of this approach, predictor variables that are directly related to uplift are unaccounted for, because the constituting two models have not been developed explicitly with the aim to estimate uplift, but rather response within the

treatment or control group. Hence, categorization as an indirect estimation approach, in contrast to direct estimation approaches, is discussed below. Possibly, the resulting two models include a different set of predictor variables, and different as well from a model that directly estimates uplift. In addition, to accurately predict uplift, both underlying models have to be accurate. Errors can be amplified when aggregated to predict uplift, leading to poor predictive performance of the aggregated model.[8] The two-model approach is also known as the *naive* or *difference score* method, and has been applied in Robins,[29] Hansotia and Rukstales,[21] Vansteelandt and Goetghebur,[30] and Robins and Rotnitzky.[31]

The two-model approach is to some extent similar to Lo's approach, with two scores being calculated and subtracted to estimate uplift. However, in the two-model approach, two separate models are developed, whereas Lo's approach yields a single, integrated model oriented toward estimating uplift rather than response. The two-model approach, generalized Lai's and Lo's approach, has been evaluated on three data sets.[9] Overall, the *generalized Lai's approach* performed best, whereas *Lo's approach* appeared to perform well for some data sets, but not all.

The modified outcome method was applied on three medical data sets[12] and compared with the two-model approach and the approaches introduced in Lo[6] and Vansteelandt and Goetghebur.[30] The reported findings support the so-called no-free-lunch theorem to apply to uplift modeling, since there was no unique winner across the data sets that were included in the experiments. On the contrary, Lo's approach was found to perform poorly on all three data sets, contradicting the findings of Kane et al.[9] reported above. An important conclusion from these experimental studies is that the performance of uplift models appears to heavily depend on the nature of the data set and application.

*Direct estimation approaches.*

*Tree-based approaches.* Most tree-based approaches for uplift modeling are adaptations from well-known decision-tree induction algorithms such as classification and regression trees (CART),[32] C4.5,[33] or chi-square automatic interaction detection (CHAID).[34] To tree-based approaches, the splitting criteria and/or the pruning techniques involved in building the model are usually modified.

A first approach for developing uplift trees[19] is similar to CART and C4.5 decision tree induction techniques, with candidate splits being evaluated in terms of a quality measure indicating the *goodness* of a split, and iteratively the best split being selected until a stopping criterion is met. The best split in an uplift tree is defined as both maximizing the difference in uplift between the resulting child nodes, with uplift measured as the difference in response rates between the treatment group and control group, and minimizing the difference in size between child nodes. The latter objective expresses a preference for splits that result in groups of equal size and a dispreference for splits leading to highly imbalanced groups. Radcliffe and Surry[8] report the optimal approach for balancing between the importance of both split characteristics to be application dependent, as such reducing the practical use of this approach.

An alternative split evaluation criterion is proposed by Radcliffe and Surry,[8] that is, the *Significance-Based Splitting Criterion*. The goodness of a split is evaluated in terms of the *significance* of the difference in uplift between the resulting child nodes, as expressed by the related coefficient in a linear regression model that is fitted to estimate response in terms of child node as well as treatment or control group membership. The significance of the coefficient is evaluated by a *t*-test. In addition to this splitting criterion, a variance-based pruning approach is introduced for deciding on the optimal size of the tree.

Chickering and Heckerman[28] adapt a CART technique by ensuring a split on the treatment variable (i.e., whether or not the person received a treatment) on every path from the root to its leaf nodes. To accomplish this, they enforce the final split of any path to be on the treatment variable. Afterward, a postprocessing step is performed, removing splits that do not result in a statistically significant difference in terms of the observed uplift. Other than this modification, the decision tree works like traditional decision trees.[32]

Hansotia and Rukstales [11,21] introduces two decision tree-based approaches. The first approach in essence boils down to the *Two-Model Approach* described before, but is reported to perform poorly compared with the second approach that is introduced, which is a CHAID-like decision tree induction In each step, the difference in uplift between the child nodes is maximized. Compared with the above-reported approach proposed of Radcliffe and Surry,[19] this approach is simpler and less refined since it does not aim at simultaneously minimizing the difference in size of the resulting child nodes when selecting splits.

Rzepakowski and Jaroszewicz[7,35] draw from the field of *information theory* to design divergence-based splitting criteria for growing uplift trees. The objective of a split in an uplift tree is considered to be maximizing the distance in the class distributions of the response

between treatment and control groups in the child nodes. In other words, the fractions of responders and nonresponders (i.e., the class distribution of the target variable) in the treatment groups should be as different as possible from the fractions of responders and nonresponders in the control group. Divergence measures allow measuring the difference in class distribution in both child nodes. Three different distribution divergence measures are adopted: the Kullback–Leibler divergence, the squared Euclidean distance, and the chi-squared divergence. The *Divergence Gain* measure, which serves a similar role as the information gain measure in CART and C4.5, for a split $A$ and divergence measure $D$, is then defined as follows:

$$D_{gain}(A) = D(P_T(Y) : P_C(Y)|A) - D(P_T(Y) : P_C(Y)) \quad (20)$$

The Divergence Gain is the decrease in divergence in the parent node, $D(P_T(Y) : P_C(Y))$, and the weighted average divergence in the child nodes $D(P_T(Y) : P_C(Y)|A)$ resulting from a split $A$. Notice that this measure accounts for imbalanced child nodes by calculating the weighted average divergence in the child nodes as follows:

$$D(P_T(Y) : P_C(Y)|A) = \frac{n_L \times D(P_{L,T}(Y) : P_{L,C}(Y)) + n_R \times D(P_{R,T}(Y) : P_{R,C}(Y))}{n_L + n_R} \quad (21)$$

with $n_L$ and $n_R$ the number of observations in the left ($L$) and right ($R$) child nodes, respectively, and subscripts $T$ and $C$ indicating treatment and control group, respectively. The adopted divergence measure $D$ can be any of the previously mentioned tests and therefore can be regarded as a parameter of the divergence-based uplift tree approach. A similar approach has been implemented in Michel et al.[36] making use of the chi-squared divergence for the splitting criteria.

*Ensemble approaches.* Ensemble approaches for uplift modeling have been proposed only recently.[8,13,24,37,38] In Radcliffe and Surry,[8] an ensemble of uplift models is developed by adopting the bagging meta learning approach,[39] with the aim to improve the stability of the final uplift model, that is, to improve the generalization power. In Guelman et al.,[13] decision trees are indicated to suffer from high variance because of the hierarchical nature of the splitting process, since every error in the top node of a tree will eventually be propagated to all of the leaf nodes. Therefore, a *random forests*-based approach is proposed,[40] called *uplift random forests*. Combining a group of uplift trees in an ensemble setup provides a means to reduce the variance because of the reduction of the correlation between the trees.[13] Just as in Rzepakowski and Jaroszewicz,[35] the splitting criteria of the ensemble of uplift trees are based on a conditional divergence measure. Further adjustments are proposed[24] to avoid multivalued predictor variables to be favored by the splitting criterion, and with a built-in stopping criterion to avoid overfitting. This results in *causal conditional inference trees* and *causal conditional inference forests*, with the estimated uplift calculated as the average uplift over the trees in the ensemble.[13]

In Soltys et al.,[37] an extensive analysis is performed regarding ensemble approaches for uplift modeling, including bagging and random forests. A bagging approach is proposed similar to the original bagging approach for classification,[41] but adapted toward the uplift case. For the random forest techniques, the authors compared the techniques proposed in Guleman et al.[13] to the *double uplift random forests* (double randomized decision trees, one for the treatment data set and one for the control data set) on a number of data sets. The main conclusion is that ensembles are powerful uplift modeling approaches able to achieve excellent results.[37]

More recently, Zhao et al.[16] proposed a tree-based ensemble called the context treatment selection (CTS). The novelty of this approach is situated in the support for analyzing and developing uplift models for multiple treatments, which so far has received limited attention in the literature.[35] In essence, the CTS approach selects in each leaf node the most optimal treatment, which allows for new instances to be classified with the treatment that is most successful according to the selected performance metric. In a later version they adapted the CTS algorithm to the unbiased contextual treatment selection, which eliminates the estimation bias of leaf responses by using separate data sets for partition generation and leaf estimation.[42]

## Evaluation

In this section, we review evaluation approaches for assessing the performance of uplift models. Both visual approaches and performance metrics are discussed and illustrated.

In line with standard predictive modeling practices,[43] for uplift model development and evaluation, the data set can be randomly split into a separate train and test set, respectively. Typically, the train set consists of 70% and the test set of 30% of the observations in the data set. By evaluating the uplift model on a test set that was not used for developing the model, an

unbiased estimate of the performance of the model is obtained. Notice that for parameter tuning purposes, the train set may be further split into a train and validation set, which is typically done as well in a proportion of 70% versus 30%. Alternatively, a cross-validation evaluation approach can be adopted, as detailed in the experimental section, which basically concerns a repeated split sample evaluation.

In predictive modeling (e.g., classification or regression modeling), performance metrics typically summarize the accuracy of model estimates for each observation in the test set. That is, the predicted and actual outcome or observed outcome for each observation is compared. In uplift modeling, however, the predicted outcome is not observed, that is, we cannot know the net effect of a treatment, for example, whether a customer is a persuadable, lost cause, sure thing, or do-not-disturb. Only one treatment can be applied and the outcome observed. The outcome for alternative treatments cannot be applied, so the net difference in outcome compared with applying alternative treatments is unobservable. This problem is known as the fundamental problem of causal inference.[44] Therefore, for evaluation, the difference in outcome for different treatments is compared across *similar* groups of observations, that is, equivalent segments of the population that are estimated by the model to experience the same or similar net effect of the treatment. Subsequently, the difference in outcome for the groups that received a different treatment, for example, the treatment and control, can be calculated and compared to the predicted effect of the treatment by the model.[8]

Keep in mind that both training and test sets have a treatment and control group. All models produce an uplift score, which is usually some form of Equation (1), that is, the probability of a person responding favorably, given the treatment, minus the probability of person responding favorably, while not being treated.

Uplift chart. One of the most common ways to evaluate uplift is visually through the means of uplift charts. After a model is built, it is used to score each individual of the test set with an uplift score. These scores are ordered from high to low and binned together in (semi-)deciles. This way, all the individuals the model perceives as being persuadable will have a high uplift score and thus be in front. For each decile, there are individuals who either belong to the treatment group or the control group. The incremental response rate of a decile is calculated by subtracting the response rate of the control group from the response rate of the treatment group. A successful model will accomplish ranking the responders of the treatment group high in the first deciles, whereas the responders of the control group should be ranked low in the last deciles. Theoretically, an ideal uplift chart should look like the left chart in Figure 3. However, uplift models are generally not that stable[9] and the lift charts in practise usually look like the chart on the right side of Figure 3. With stability is meant how well the models perform in multiple situations. Every scenario is different and no uplift model is ideal for each case.

Although visualizations have advantages such as offering a clear immediate impression of the performance of the model, it is not ideal when using them to compare different models with each other on the same data set. Numerical metrics are more precise and usually more preferred.

Gini coefficient and gains chart. The Gini coefficient is a measure of goodness-of-fit and one of the measures typically used in direct marketing as a way to measure the traditional response models. It is a measure of goodness-of-fit and is connected to the Lorenz curve. Just as with the uplift charts, the scores are ordered from high to low. The Lorenz curve represents the cumulative lift (Fig. 4), which is calculated by accumulating the responses and dividing by the total amount of responses. If an optimal model was represented, we would see a steady Lorenz curve rising steeply upward. This means the model is successful in ranking the responding individuals first and thus always targets responders. When all responders have been accounted for, we would see a horizontal line, that is, no extra lift is achieved, on the graph all the way to the end.

The Gini coefficient is computed by measuring the ratio between two areas as seen in Figure 4. Area A corresponds to the area between the Lorenz curve and the diagonal line (which corresponds to random targeting). Area B is the area between the optimal curve and the diagonal line.

The Gini coefficient is defined between 0 and 1. A value of 1 represents a perfect model that ranks the customers in such a way that the positive class is separated from the negative class. A value of 0 represents a model that is not performing any better than random ranking of customers (i.e., the responses are equally distributed over all deciles).

The Gini coefficient is a good metric as it is a single value and can thus be easily used to compare the performance of other models on a particular data set.

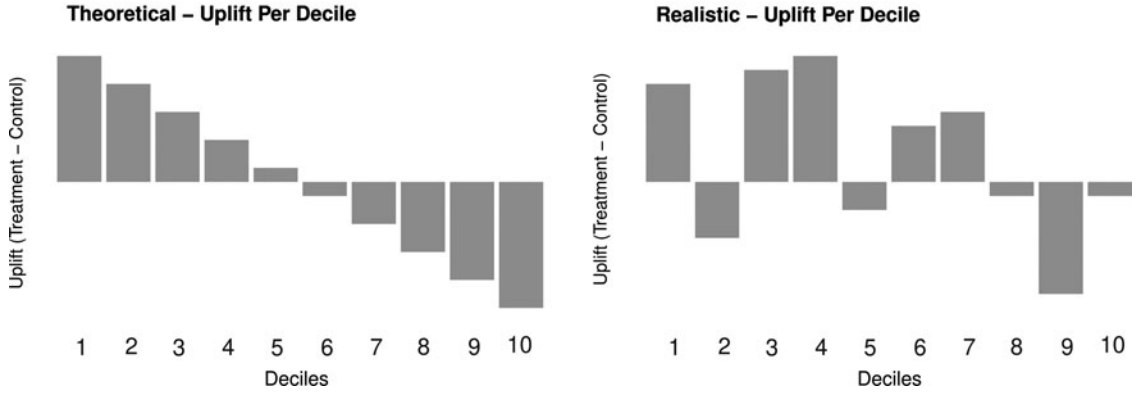


**FIG. 3.** On the left-hand side, an uplift chart of a good uplift model is shown where the first deciles accumulate the biggest incremental response rate. On the right-hand side, an uplift chart of a bad uplift model is displayed, showing negative uplift for the second decile and high uplift for lower ranked customers. Figures adapted from Kane et al.[9]

This metric was sometimes used in traditional response models. The Gini coefficient is a very useful metric, however, the Gini coefficient cannot be readily applied in uplift modeling because of having both the treatment and control groups. It is not possible to know how an individual would respond in both groups.

In the normal Gini calculations, the value depends on having an idea of what the optimal curve is. The denominator usually represents the maximum possible value the numerator can take. However, in uplift modeling, the maximum value is data dependent.[9] There is also the possibility of negative uplift, which the Gini coefficient does not take well into account. Therefore, in the literature, some adjustments have been made to the Gini coefficient to take these issues into account and these are presented next.

Gini coefficient by Kane. In Kane et al.,[9] the authors have chosen to make a small alteration in the Gini coefficient calculation. The average uplift is calculated for every decile and is used to calculate the cumulative lift and the cumulative sample. The Gini coefficient is then calculated as follows:


**FIG. 4.** Incremental gains or Qini curve. Color images available online at www.liebertpub.com/big

$$\text{Gini Coefficient} = \sum_{i=1}^{n} \left( \frac{\sum_{j=1}^{i} lift(j) n_{tj}}{\sum_{j=1}^{n} lift(j) n_{tj}} - \frac{\sum_{j=1}^{g} n_{tj}}{n_t} \right) \quad (22)$$

with $t$ indicating the treatment group. The only requirement for this metric is that it can only be used to compare models on the same data set. For more information on the calculations, see Kane et al.[9]

Qini coefficient by Radcliffe. Radcliffe[45] present a generalization of the Gini coefficient, called the Qini coefficient. This is based on the "Gains Chart for Uplift," or the Qini curve, which is similar to the regular gains curve. The Qini curve plots the cumulative difference in response rate between treatment and control test sets as a function of a selected fraction x of the entities as ranked by the uplift model from high to low uplift. This curve is also called the cumulative uplift or the cumulative incremental gains.[45] An important reminder is that the performance is evaluated by comparing groups of observations rather than individual observations. The Qini value $Q$ is the ratio of the actual uplift gains curve (or the incremental gains) above the diagonal, representing the incremental gains achieved through random targeting, to that of the optimal Qini curve. More information can be found in Radcliffe.[45]

In Surry and Radcliffe (Surry PD, Radcliffe NJ. Quality measures for uplift models; submitted to KDD 2011; unpublished data), the authors have proved mathematically that the Qini curve is equivalent to the Gini curve of the treatment group minus the Gini curve of the control group:

$$Qini = Gini_{Treatment} - Gini_{Control} \quad (23)$$

Area under uplift curve. In Rzepakowski and Jaroszewicz,[7] the authors use the area under uplift curve (AUUC) as a metric to evaluate the performance of the models. This metric calculates the area underneath the curve compared with the optimal curve. The metric is very similar to the Qini value.[45] The difference is that the Qini measure also takes into account the random incremental gains through random targeting, whereas the AUUC does not.

## Conclusions of the literature survey

In this section, a comprehensive survey is provided of the literature on uplift modeling by means of Table 1. The most prominent approaches, experimentally evaluated in the second part of this article, are discussed in detail, and evaluation approaches for assessing the performance of uplift models are reviewed.

In conclusion, we find that in recent years an increasing, although still relatively limited, number of articles are devoted to developing and applying uplift modeling. The main field of application remains response modeling, although a clear need and use for these approaches exists in other application fields. Hence, further research, developing applications and case studies elaborating the adoption of uplift modeling in different fields, is needed. In addition, we find that a broad and thorough experimental evaluation and comparison of the various approaches that have been introduced are missing. Several studies present the results of limited experimental studies, typically comparing the performance of a newly introduced approach to some baseline approaches. We find that these results, when compared across articles, are often contradictory rather than complementary. Which is no surprise, since in predictive modeling, the *no-free-lunch* theorem states that no single technique always performs best, that is, the performance of predictive models depends on application-specific problems and data set characteristics. Hence, there is an actual need for a broad experimental evaluation, as presented in the second part of this article, providing insight into the variation in performance across applications and approaches, as well as providing a benchmark to both the scientific community and practitioners. Such a general reference point facilitates comparison and supports practitioners to select approaches for practical implementations, spurs further development, and allows the scientific community to reach consensus with respect to the appropriateness and power of various modeling approaches. Notice that, in addition, such broad experimental studies provide further insight into the general shortcomings as well as the practical challenges in elaborating actual applications, leading to the identification of research gaps and objectives for further research. As such, the state-of-the-art is characterized and advanced.

The main research gaps identified as a result of the literature survey that was performed concern the following:

- Many studies report the resulting uplift models to be unstable in terms of performance. To a large extent, this is considered to be a result of the (often strongly) imbalanced class distribution, for example, there are typically much less responders than nonresponders. In addition, in uplift modeling there may be an additional imbalance, that is, between the treatment and control group, both in terms of size and in terms of

observed class distributions for these groups, and as such causing instability. Although ensemble approaches have been developed and shown to address to some extent this issue, further research should focus on developing or extending approaches that address the instability issue. Inspiration may be found in cost-sensitive learning approaches and sampling approaches as adopted in the field of predictive analytics to improve learning when the class distribution is heavily imbalanced.

- Most, if not all, approaches and case studies that are developed in the literature concern a binary experimental design, that is, either a treatment is applied or no treatment is applied, leading to the control group and treatment group. Further differentiation in terms of treatments should be accommodated, subsequently facilitating a further optimization of decision-making in terms of the optimal treatment to apply for maximizing or minimizing the outcome. However, this as well increases the applicable requirements in terms of data that are needed for developing uplift models. Given the involved costs for setting up experiments in which random groups are treated in various manners, the data requirements can be expected to be restrictive toward the development of multiple treatment uplift approaches.
- Finally, further research may extend uplift modeling for regression, that is, for continuous target variables. Example applications involving a continuous target variable include customer lifetime value, loss given default, and customer demand modeling.

In conclusion of the section on uplift model evaluation, we find that previous works mainly adopt visual approaches such as response or uplift-by-decile graphs[6,21] or an adapted version of the Gini coefficient.[9,45] Although the literature has not grown to a general consensus and has not developed mature evaluation approaches, the Qini metric, as discussed before, has recently gained popularity in scientific publications. Therefore, we claim there is a need for improved, appropriate performance metrics and evaluation procedures, allowing to assess and intuitively interpret model strength from an application perspective, as well as to handle the imbalanced class distribution and the instability of uplift models, as reported by several authors. Therefore, we highlight the development of flexible, user-friendly, and accurate evaluation procedures as a key topic for future research.

## Experimental Setup

A selection of uplift modeling approaches as presented in the previous section is experimentally evaluated on four uplift data sets. In this section, the overall experimental setup is detailed, including the overall experimental procedure, the adopted variable selection procedure, and the setup of uplift modeling techniques.

### Experimental procedure

To achieve consistent results that allow to correctly compare the performance of uplift modeling techniques across the four data sets, a formalized experimental procedure has been designed and implemented for testing the various techniques on the four data sets. The most important steps in the procedure, as fully detailed in Algorithm 1, are the following:

- Seeds: the seed required for random sampling and random initialization is set at the start of each run of the experimental procedure, allowing to rerun experiments and validate outputs. For example, when applying cross-validation, an identical random split-up of the data set in train and test set is made when using the same split-up. As such, a potential impact of randomness is eliminated when comparing results across levels of experimental factors such as uplift modeling technique.
- *N*-fold stratified cross-validation: cross-validation is a procedure to achieve stable and valid experimental results in terms of the performance of a modeling technique, by repeatedly splitting the data set in test and train sets in a random manner, without overlap across the *N* test sets or folds.[27] Given the typical imbalanced class distribution, the folds are stratified, that is, randomly sampled so as to ensure an identical distribution of the target variable across the folds. In addition, as well an identical distribution of treatment and control group observations in the folds is enforced, which is a specific requirement that applies to uplift modeling.
- Variable selection: a number of data sets include many variables, among which many irrelevant variables. This has been found to sometimes hinder modeling techniques to achieve good performance,[46,47] as well as to slow down development of models. Therefore, given the extensive experimental design, to speed up the experiments and to achieve near-optimal performance, a variable selection procedure is implemented in the experimental procedure as outlined below.

**Algorithm 1: Experimental Procedure**

```
1: Set seed determining randomness.
2: Load preprocessed data.
3: Optional: Take sub-selection of the data.
4: Create n stratified folds.
5:   Each fold has the same distribution of:
6:     treated versus nontreated and responders versus
         nonresponders.
7: for i=1 to n do
8:   Test set is fold i
9:   Training set consists of n−1 folds, excluding fold i
10:  Split the training set randomly into two sets:
11:    A tuning set: used to build the model (2/3 of the training set).
12:    A validation set: used to test the tuned model (1/3 of the
         training set).
13:    Tuning and validation sets have the same distribution of:
14:      treated versus nontreated and responders versus
           nonresponders.
15:  Perform variable selection procedure on the tuning set:
16:    if number of variables >50 then
17:      Perform NIV variable selection procedure.
18:    Perform a wrapper variable selection procedure.
19:      Develop the models on the tuning set.
20:      Test the models on the validation set.
21:    Select the optimal number of variables maximizing the
         performance (Qini) on the validation set.
22:  Perform parameter tuning:
23:    Set up a tuning grid of the selected parameters.
24:    for all possible combinations do
25:      Develop the models with the selected set of variables on the
           tuning set.
26:      Test the models with the selected set of variables on the
           validation set.
27:    Select the optimal set of parameters.
28:  Run experiment.
29:    With:
30:      The optimal number of variables.
31:      The optimal set of parameters.
32:    Build on the full training set.
33:  Save uplift model.
34:  Evaluate performance on test set.
35:  Return performance metrics.
36: Average the metrics across folds and calculate standard deviation.
```



## Experimental design

The experimental design includes two factors tested to assess their impact on the performance of the resulting uplift model. The first factor concerns the uplift modeling technique that is used in the model. The levels of this factor are the various techniques that are evaluated and discussed in detail in the first part of this article.

The second factor is the variable selection procedure that is applied. Each uplift modeling technique without an internal variable selection procedure is applied in combination to an external wrapper-based variable selection. This factor has two levels, that is, with (level 1) or without (level 0) variable selection. Full details on the variable selection procedure are provided below.

### Uplift modeling techniques.

The uplift modeling techniques, discussed in detail in the Uplift Modeling Techniques section under Literature Survey, are evaluated in the experimental study. Table 4 provides a full overview of the various approaches and variations of approaches that have been implemented in the open-source statistical software R. Available R-packages implementing a number of the tested approaches were used, that is, Uplift, Information, and Caret. The full code of the implementations and of the experimental procedure is made available as a digital appendix to this article,* to provide full details on the experiments and to allow reproduction of the results of the experimental study. This is permitted since we mostly make use of publicly available data sets. in addition, by providing such a platform, we aim to facilitate further research to expand on this study by evaluating the performance of new techniques on new data sets in new domains. As such, the state-of-the-art in the field is advanced and further research on the topic of uplift modeling is spurred.

When applicable, parameter tuning has been applied to optimize performance, involving experimental evaluation on a validation set of various parameter values, or combinations of parameter values if more than one parameter governs the development of a model. For this, a tuning grid is to be set up, with a sufficiently wide range of possible values to be determined as well as the stepsize for varying the parameter value within the specified range.[43] Notice that, the wider the range and the smaller the stepsize, that is, the higher the resolution, the more models need to be developed and as such the more time it takes for optimizing the parameters. Given the large number of uplift modeling approaches tested in this study on four data sets, a balance was kept between aiming for optimal performance and limiting the magnitude of the experiments, in line with the objectives of the experimental study. In addition, for tuning, an appropriate performance measure is to be selected, which for evaluating uplift models is the Qini metric, as discussed before. Examples of parameters that can be tuned are the number of trees in a random forest and the number of candidate predictor variables that are randomly selected for making splits in the nodes of the decision trees in a random forest.

### Variable selection procedure.

A variable selection procedure is typically adopted with the aim to reduce the complexity of the resulting model and as a result to increase the stability, that is, robustness or generalization

*Available at www.data-lab.be

**Table 4. Overview of the techniques used in the experiments**

| *Uplift modeling technique* | *Code* | *Classifier* | *References* |
|---|---|---|---|
| Lai's Method | Lai | Stochastic gradient boosting | 9,22 |
| Generalized Lai's Method | Glai | Stochastic gradient boosting | 9 |
| Pessimistic Uplift Modeling | Pes | Logistic regression | 23 |
| Response Variable Transformation For Uplift Modeling | Trans | Logistic regression | 12,24 |
| Dummy Treatment Approach | Dta | Logistic regression | 6,9 |
| Two-Model Approach | Tma | Logistic regression | 9,11,21 |
| Uplift Random Forest—Splitting Rule: Euclidean Distance | urf-ED | Random forest | 24,38 |
| Uplift Random Forest—Splitting Rule: Chi-squared Divergence | urf-Chisq | Random forest | 24,35,38 |
| Uplift Random Forest—Splitting Rule: Kullback–Leibler Divergence | urf-KL | Random forest | 24,35,38 |
| Uplift Random Forest—Splitting Rule: Interaction Method | urf-Int | Random forest | 8,24,38 |



power, as well as possibly the comprehensibility of a model, although a trade-off exists between the number of predictor variables that is selected (which is preferred to be small for the abovementioned reasons) and the predictive power, which typically depends on the number of included predictor variables to a certain extent.[47] Variable selection aims at reducing overfitting and avoiding correlation between predictor variables, that is, multicollinearity. An interesting discussion on variable selection for uplift modeling can be found in Radcliffe and Surry,[8] as well as in Larsen,[20] who in fact develops an uplift modeling approach based on variable selection (Net Weights of Evidence and Net Information Value section) making use of the weight of evidence and the IV. In addition to Larsen's approach, also a general wrapper-approach[47] for variable selection is adopted in combination with the Qini measure for assessing performance.

*General wrapper-style approach.* The general wrapper approach is a search paradigm for selecting an optimized subset of predictor variables following an iterative design. The process is formalized in Algorithm 2 and 3. When a data set has $n$ predictor variables, the wrapper approach will develop $n$ models, each model including all predictor variables except one. The models are then ranked according to a prespecified and appropriate performance metric and the best model is selected. This model includes $n-1$ predictor variables, and the predictor variable that was not included in this model is removed from the data set. This process is repeated until all predictor variables are removed. As a result, the trade-off between predictive power and number of predictor variables can be plotted and the optimal number of predictor variables can be determined, which is typically done in an unstandardized manner and reflecting the optimal trade-off for the developer of the model.

**Algorithm 2: Creating the different predictor lists**

```
1: procedure CREATEPREDICTORLISTS (predictors)
2:   listPredictorSets ← list()
3:     for i in 1:length(predictors)) do          ▷ Go over each variable
4:       predictorSet ← predictors[-i]            ▷ Get all variables, except for i
5:       listPredictorSets[i] ← predictorSet
6: return listPredictorSets                       ▷ Return the list
```

**Algorithm 3: Variable Selection**

```
1: procedure VARIABLESELECTION(df.train,df.test,
   predictors,bestModelList)                      ▷ df = DataFrame
2:   listPredictorSets ← CreatePredictorLists(predictors)
3:   listOfQinis ← list()
4:   for i in 1:length(listPredictorSets)) do     ▷ Go over each predictorset
5:     QiniValue ← runExperiment(df.train, df.validation, predictors)
6:     listOfQinis[i] ← QiniValue
7:   bestQiniValueIndex ← which.max(listOfQinis)
8:   predictionsAndQini ← c(listPredictorSets[bestQiniValueIndex],
   listOfQinis[bestQiniValueIndex])
9:   bestModelList[length(listPredictorSets)] ← predictionsAndQini
10:  if length(listPredictorSets) == 1 then
11:    return bestModelList
12:  else
13:    VariableSelection(df.train, df.test,
   listPredictorSets[bestQiniValueIndex],bestModelList)
```

## Empirical Results

### Data sets

Four real-world data sets have been gathered for the experimental evaluation. Table 5 provides an overview of the key properties of these data sets. The first three data sets are publicly available. The first data set is part of the INFORMATION R-package.[†,20] The data relate to a marketing campaign in the insurance industry and the target variable indicates whether or not a purchase happened. The second data set is published on the website MineThatData[‡] and contains data from an e-mail marketing campaign concerning clothing merchandise. The data set includes three target variables: visit (yes/no), purchase (yes/no), and conversion (numerical; the amount of money spent). The data set includes 64,000 observations with 1/3 targeted with an e-mail

[†]https://cran.r-project.org/web/packages/Information/index.html
[‡]http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

**Table 5. Overview of data sets used in the experiments**

| *Variable* | *Data set 1* | *Data set 2* | *Data set 3* | *Data set 4* |
|---|---|---|---|---|
| Type | Development | Development | Development | Retention |
| Description | Insurance | Online merchandise | Retailer | Financial services (bank) |
| Channel | E-mail | E-mail | Catalog | Unspecified |
| Public or private | Public | Public | Public | Private |
| No. of observations | 10,000 | 42,693 | 100,000 | 200,903 |
| No. of treatment observations | 4972 | 21,387 | 50,000 | 82,094 |
| No. of control observations | 5028 | 21,306 | 50,000 | 118,809 |
| No. of variables | 68 | 10 | 95 | 160 |
| Response variable (binary) | Purchase | Visit | Response | Churn |
| Treatment-to-control size ratio | 0.99:1 | 1:1 | 1:1 | 0.69:1 |
| Treatment positive rate | 20.37% | 15.14% | 3.71% | 13.25% |
| Control positive rate | 19.55% | 10.62% | 3.09% | 25.51% |
| Uplift initial campaign | 0.82% | 4.52% | 0.62% | −12.27% |
| Signal-to-noise ratio | 4.21% | 42.61% | 19.92% | −48.08% |



campaign concerning men's merchandise, 1/3 targeted with an e-mail campaign concerning women's merchandise, and 1/3 customers who were not targeted. For this data set, the "visit" target variable was selected as the target variable of interest and the selected treatment is the e-mail campaign for women's merchandise, in line with a previous study[9] to facilitate comparison. The third publicly available data set was published on the Udemy-website.[§] The data set concerns a retailer's catalog marketing program. The target variable indicates whether or not there was a response, up to two months after customers were targeted with a marketing campaign.

The fourth data set concerns a customer retention campaign. At first sight, the objective of a retention campaign differs from the objective of a *response* marketing campaign, in the sense that instead of targeting customers who are likely to be influenced by the campaign to buy a product, retention campaigns target customers who are likely to be influenced by the campaign to remain loyal to the company, that is, not to churn. In this regard, for retention campaigns, uplift represents a *decrease* in propensity to churn. Therefore, from a conceptual and uplift modeling perspective, the objective is identical. Both for response and retention marketing campaigns, uplift modeling allows to decide about which treatment to apply to individual customers, based on individual customer's characteristics, by estimating the net effect of a *treatment* on the behavior of the customer, whether it be a net increase in probability to response or a net increase in probability to remain loyal. Practically, rather than defining a positive response as *churning*, responding to a retention campaign is defined as *not churning*, leading to a fully comparable analysis as for the three other data sets. The response rate, that is, the nonchurn rate, of the treatment group, should be higher than the response rate of the control group.

[§]https://www.udemy.com/uplift-modeling

### Variable selection

Considering time and resource constraints, for data sets 1, 3, and 4, the NIV was used to select the 50 most predictive variables before applying the general wrapper approach. Variable selection has been applied in combination with *tma*, *dta*, *pes*, and *trans* approaches, but not in combination with tree-based approaches since these internally incorporate a variable selection procedure. So there is no need from a practical perspective, and in addition, the performance of tree-based approaches may deteriorate when adding an external variable selection procedure.

As can be seen from the experimental results shown in Figure 5, generally, the best performance in terms of the Qini measure is achieved with a relatively small number of predictor variables in the model, that is, in between 5 and 15 predictor variables. On all four data sets, when adding predictor variables to the model, at first the performance of the model improves. When the optimal number of predictor variables resulting in maximum performance is reached and more predictor variables are added to the model, a downward trend in performance is observed. An explanation of the downward trend is that the effect of the campaign is captured or described by a limited number of predictor variables. When adding more predictor variables, which are either only weakly related to the effect of the campaign or strongly correlated with the predictors already in the model, in fact noise is

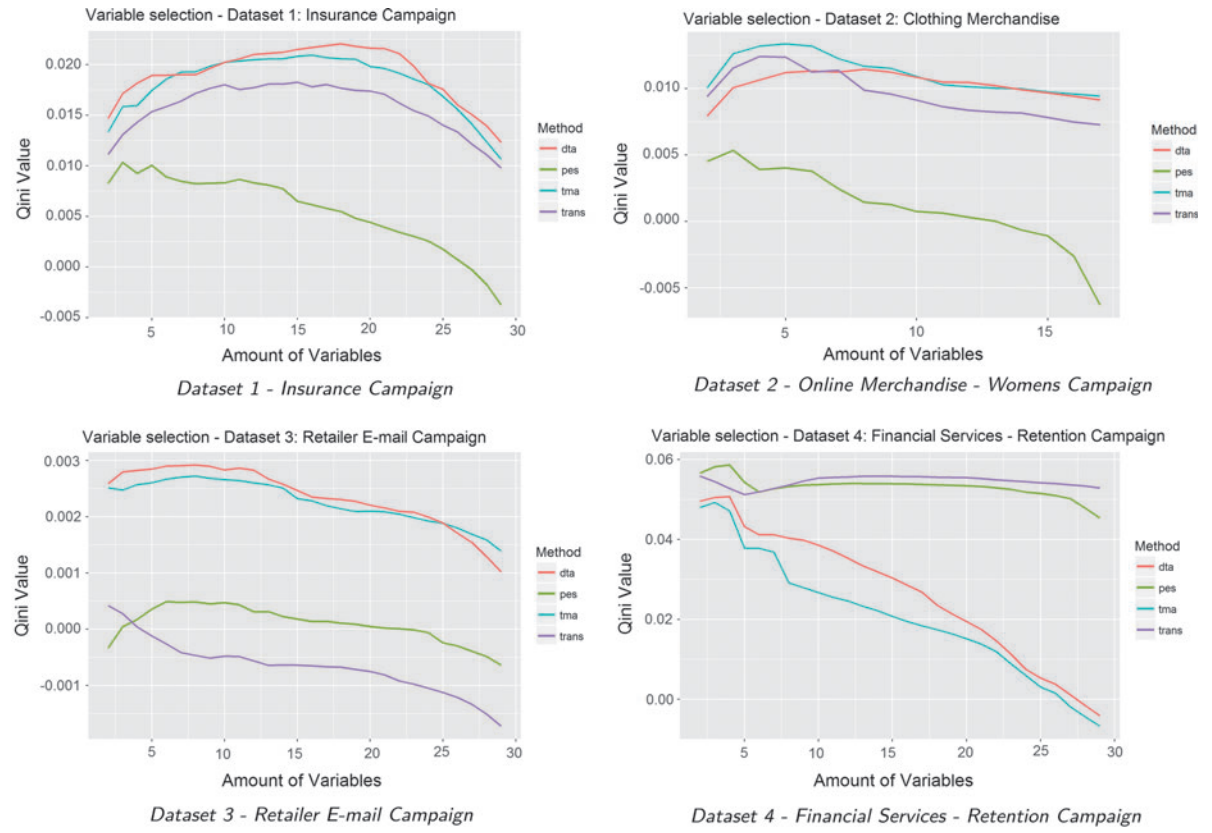


**FIG. 5.** Results of the wrapper variable selection procedure applied in combination with the methods tma, dta, pes, and trans on the four data sets. Color images available online at www.liebertpub.com/big



added to the data, leading to a decrease in terms of the performance and generalization power of the models, or an increase in terms of stability or overfitting.

## Results and discussion

In this section, we report the experimental results. First, the predictive performance of the various uplift modeling approaches across the four data sets is assessed by means of the Qini metric. Then, visual evaluation approaches are adopted, such as Qini-plots and uplift decile charts, to further analyze performance. Finally, the stability of the resulting models is discussed.

### Numerical analysis.

Tables 6 and 7 report the performance in terms of the Qini and Gini metrics, for the various approaches on the four data sets, as well as the Top 30% and Top 10% Qini and Gini. Notice that the Qini and Gini evaluate performance for 100% of the population being treated, that is, could be regarded as the Top 100% Qini and Top 100% Gini. Notice that, generally, the reported Qini > Qini Top 30% > Qini Top 10%. The same holds for the Gini, and makes perfect sense, since the area under the Qini and Gini curves can be expected to be larger for an increasing fraction of the population being selected. When looking at the results of data set 3, however, it can be observed that the performance of several techniques does not comply with this expectation. This is further analyzed below.

The standard deviation, as reported in between brackets next to the average Qini and Gini measures in Tables 6 and 7, indicates a substantial variation in performance across the validation folds. This may signal an issue with the stability of the models, as further discussed in the Stability section.

Tables 6 and 7 allow comparing the performance of various techniques across the four data sets. At first sight, it appears that uplift random forests consistently performs among the best techniques. Two more techniques, *dta* and *tma*, as well perform consistently well across multiple data sets.

**Table 6. Results of the experiments with the average Qini and standard deviation across the folds reported for each technique on each data set, with exception of urf-ED and urf-L1 for data set 3, which appeared to be computationally infeasible**

| | *Data set 1* | | | *Data set 2* | | |
|---|---|---|---|---|---|---|
| *Technique* | *Qini* | *Qini Top 30%* | *Qini Top 10%* | *Qini* | *Qini Top 30%* | *Qini Top 10%* |
| dta | 1.19804 (0.84582) | 0.24266 (0.16698) | 0.02688 (0.01704) | 0.73138 (0.30557) | 0.10887 (0.06159) | 0.01303 (0.00634) |
| tma | 1.03303 (0.50332) | 0.20718 (0.10525) | 0.01031 (0.01139) | 0.69436 (0.2506) | 0.08801 (0.06137) | 0.01175 (0.00761) |
| pes | −0.11695 (0.95243) | −0.05955 (0.18488) | −0.01258 (0.01452) | 0.10105 (0.26275) | 0.01443 (0.07041) | 0.00249 (0.00771) |
| trans | 0.7311 (0.64417) | 0.09419 (0.14353) | 0.0127 (0.03164) | 0.79716 (0.30777) | 0.15949 (0.07339) | 0.01903 (0.01019) |
| lai | 1.93541 (0.47193) | 0.31223 (0.16493) | 0.03652 (0.02978) | 0.68642 (0.33382) | 0.12095 (0.10856) | 0.01028 (0.01504) |
| glai | 1.94237 (0.47805) | 0.31368 (0.17317) | 0.03829 (0.0313) | 0.68225 (0.3315) | 0.12178 (0.10836) | 0.01046 (0.01535) |
| urf-KL | 1.7267 (0.59045) | 0.34581 (0.10118) | 0.05112 (0.01701) | 0.39527 (0.28409) | 0.1232 (0.08295) | 0.01845 (0.01639) |
| urf-Int | 1.95192 (0.54702) | 0.42458 (0.12839) | 0.05933 (0.02115) | 0.44974 (0.22472) | 0.14288 (0.05691) | 0.01795 (0.00778) |
| urf-Chisq | 1.79523 (0.68438) | 0.36613 (0.12326) | 0.05682 (0.02846) | 0.49707 (0.29362) | 0.14304 (0.09503) | 0.0187 (0.02223) |
| urf-ED | 1.80835 (0.64515) | 0.48863 (0.1866) | 0.06389 (0.03516) | 0.47637 (0.28437) | 0.12809 (0.07981) | 0.01431 (0.01542) |
| urf-L1 | 1.78538 (0.69904) | 0.36757 (0.2209) | 0.06255 (0.03317) | 0.50502 (0.2748) | 0.10424 (0.09049) | 0.0136 (0.01261) |
| knn | 1.0342 (0.32868) | 0.25798 (0.14607) | 0.04126 (0.03229) | | | |
| | *Data set 3* | | | *Data set 4* | | |
| dta | 0.146 (0.08334) | 0.05145 (0.02798) | 0.00883 (0.00463) | 0.94056 (0.15816) | 0.26635 (0.06307) | 0.04752 (0.01094) |
| tma | 0.14425 (0.10052) | 0.04751 (0.02401) | 0.00871 (0.0038) | 1.07133 (0.18251) | 0.36707 (0.05736) | 0.05849 (0.0117) |
| pes | −0.07505 (0.09691) | −0.0014 (0.02577) | 0.00018 (0.00716) | −0.25887 (0.57819) | 0.0216 (0.08783) | 0.00142 (0.01242) |
| trans | −0.10314 (0.09896) | −0.00998 (0.01324) | −0.00095 (0.00172) | 0.18679 (0.51995) | 0.12897 (0.06931) | 0.01876 (0.01212) |
| Lai | −0.01951 (0.05423) | 0.0165 (0.02682) | 0.00454 (0.00593) | 5.78086 (0.06663) | 1.11791 (0.02553) | 0.12608 (0.00517) |
| Glai | −0.01951 (0.05427) | 0.0165 (0.02682) | 0.00454 (0.00593) | 4.52845 (0.04611) | 0.40963 (0.02167) | 0.01013 (0.0191) |
| urf-KL | 0.14884 (0.03308) | 0.05944 (0.01493) | 0.01019 (0.00307) | 4.96948 (0.51917) | 1.47415 (0.12569) | 0.20598 (0.00939) |
| urf-Int | 0.14362 (0.06251) | 0.05392 (0.02712) | 0.00931 (0.00501) | 5.27961 (0.1459) | 1.38022 (0.04877) | 0.19943 (0.00718) |
| urf-Chisq | 0.087 (0.07059) | 0.03933 (0.02698) | 0.00647 (0.00671) | 5.23958 (0.44828) | 1.61705 (0.12678) | 0.23197 (0.01323) |
| urf-ED | | | | 5.1834 (0.42689) | 1.52665 (0.1369) | 0.22164 (0.01696) |
| urf-L1 | | | | 4.35877 (0.41794) | 1.14972 (0.10475) | 0.16863 (0.01757) |

The standard deviation is expressed in between the brackets. Results are multiplied with 100 to facilitate comparison.



A more exhaustive comparison of the ranking of performance of the various techniques across the four data sets, both in terms of Qini and Gini, is provided in Figure 6. A first observation from this figure is that the Qini and Gini result in similar rankings for the individual data sets, except for data set 1 where the Qini and Gini appear to yield considerably different rankings. Notice that for data set 2, the ranking is identical, and for data sets 3 and 4, the rankings are only different in terms of the rank of one or two techniques, respectively. For example, for data set 3, the urf-Int technique is ranked fourth following the Qini, and second following the Gini. Second, there is no specific technique that consistently performs among the top techniques across all four data sets. However, it can be seen that most techniques that perform well on data set 1, also perform well on data set 4. Similarly, a technique that performs well on data set 2 also appears to perform well on data set 3. As already mentioned, uplift random forests generally performs well on most data sets, with the only exception the retention campaign data set, that is, data set 4. Also, lai, glai, tma, and dta perform well in comparison. Surprisingly, the tma technique (i.e., the two-model approach), which is often referred to as the naive approach since relatively simple in terms of setup, appears nonetheless to perform well in specific cases (e.g., data sets 2 and 3). A third observation from Figure 6 is that two techniques consistently show up at the bottom end of the rankings, that is, the modified outcome method (*trans*) and pessimistic uplift modeling approach (pes). An exception to this is the performance of the trans-technique on data set 2, where it ranks among the best techniques.

In line with and extending on the discussion in the Evaluation section, we want to highlight a number of limitations with respect to the evaluation metrics. First of all, the Qini and the Gini measures only provide an indication how well a model is performing when compared with other models on the same data set. A comparison across data sets is not supported, since these measures are not normalized and therefore depend on characteristics of the application. in addition, the Qini and Gini metrics evaluate how well an uplift model ranks the full population. However, when uplift models are developed, inherently only a fraction of the population is expected to be targeted. The aim of uplift modeling exactly is to rank the population and select a subset for which the treatment is expected to have an

**Table 7. Results of the experiments with the average Gini and standard deviation across the folds reported for each technique on each data set, with exception of urf-ED and urf-L1 for data set 3, which appeared to be computationally infeasible**

| | *Data set 1* | | | *Data set 2* | | |
|---|---|---|---|---|---|---|
| *Technique* | *Gini* | *Gini Top 30%* | *Gini Top 10%* | *Gini* | *Gini Top 30%* | *Gini Top 10%* |
| Dta | 8.78 (4.8) | 2.45 (1.58) | 0.47 (0.5) | 1.63 (0.66) | 0.32 (0.18) | 0.06 (0.03) |
| Tma | 10.71 (6.88) | 2.98 (2.06) | 0.21 (0.31) | 1.55 (0.53) | 0.27 (0.16) | 0.05 (0.03) |
| Pes | −3.04 (11.91) | −1.32 (2.95) | −0.3 (0.35) | 0.23 (0.58) | 0.04 (0.21) | 0.01 (0.04) |
| Trans | −56.79 (145.11) | −4.64 (14.01) | −1.09 (3.36) | 1.77 (0.64) | 0.46 (0.2) | 0.08 (0.04) |
| Lai | 64.05 (78.6) | 14.94 (20.84) | 3.1 (4.85) | 1.49 (0.71) | 0.35 (0.29) | 0.04 (0.06) |
| Glai | 54.69 (54.71) | 12.36 (14.71) | 2.63 (3.5) | 1.49 (0.7) | 0.36 (0.29) | 0.05 (0.07) |
| urf-KL | 30.62 (21.76) | 7.26 (4.38) | 2.07 (1.63) | 0.87 (0.61) | 0.34 (0.23) | 0.08 (0.07) |
| urf-Int | 16.21 (3.29) | 4.38 (0.65) | 0.91 (0.12) | 0.97 (0.49) | 0.4 (0.16) | 0.08 (0.03) |
| urf-Chisq | 20.73 (13.05) | 4.95 (2.44) | 1.17 (0.68) | 1.1 (0.65) | 0.4 (0.25) | 0.08 (0.1) |
| urf-ED | 20.25 (8.7) | 6.33 (2.33) | 1.29 (0.73) | 1.03 (0.59) | 0.37 (0.2) | 0.06 (0.07) |
| urf-L1 | 16.65 (6.62) | 4.06 (2.83) | 1.14 (0.84) | 1.13 (0.62) | 0.3 (0.24) | 0.06 (0.05) |
| | *Data set 3* | | | *Data set 4* | | |
| Dta | 2.62 (1.46) | 1.16 (0.58) | 0.34 (0.16) | 0.47 (0.09) | 0.25 (0.05) | 0.08 (0.03) |
| Tma | 2.6 (1.94) | 1.16 (0.59) | 0.36 (0.14) | 0.61 (0.08) | 0.32 (0.05) | 0.09 (0.02) |
| Pes | −1.08 (1.66) | 0 (0.51) | 0.02 (0.25) | −0.1 (0.22) | 0.02 (0.07) | 0.01 (0.04) |
| Trans | −1.5 (1.77) | −0.17 (0.3) | −0.02 (0.06) | 0.15 (0.07) | 0.12 (0.09) | 0.04 (0.04) |
| Lai | 0.5 (1.22) | 0.51 (0.59) | 0.19 (0.22) | 0.86 (0.01) | 0.4 (0.01) | 0.09 (0.01) |
| Glai | 0.5 (1.22) | 0.51 (0.59) | 0.19 (0.22) | 0.38 (0.06) | 0.01 (0.06) | −0.05 (0.05) |
| urf-KL | 2.83 (0.71) | 1.29 (0.36) | 0.36 (0.12) | 2.09 (0.03) | 0.87 (0.02) | 0.22 (0.02) |
| urf-Int | 2.77 (1.28) | 1.13 (0.6) | 0.32 (0.18) | 2.19 (0.03) | 0.89 (0.02) | 0.22 (0.01) |
| urf-Chisq | 1.8 (1.28) | 0.84 (0.54) | 0.23 (0.23) | 2.12 (0.05) | 0.91 (0.03) | 0.23 (0.02) |
| urf-ED | | | | 2.05 (0.03) | 0.89 (0.02) | 0.23 (0.01) |
| urf-L1 | | | | 1.56 (0.05) | 0.62 (0.02) | 0.17 (0.01) |

The standard deviation is expressed in between the brackets. Results are multiplied with 100 to facilitate comparison.



increased effect. It is typically not efficient, nor optimal and possible, to treat the full population. If it were, there would be no need for uplift modeling. For example, a marketing campaign may only target 10% of the customer base. In such a setting, we are less concerned about the accuracy of the uplift model in ranking the full customer base from high to low uplift, but rather we are interested in the uplift that is achieved among the 10% highest ranked since this will be the actual outcome when implementing and operating the uplift model in practice.

The current measures used in evaluating the performance of uplift models in scientific studies do not take into account the specific *targeting depth* that applies.

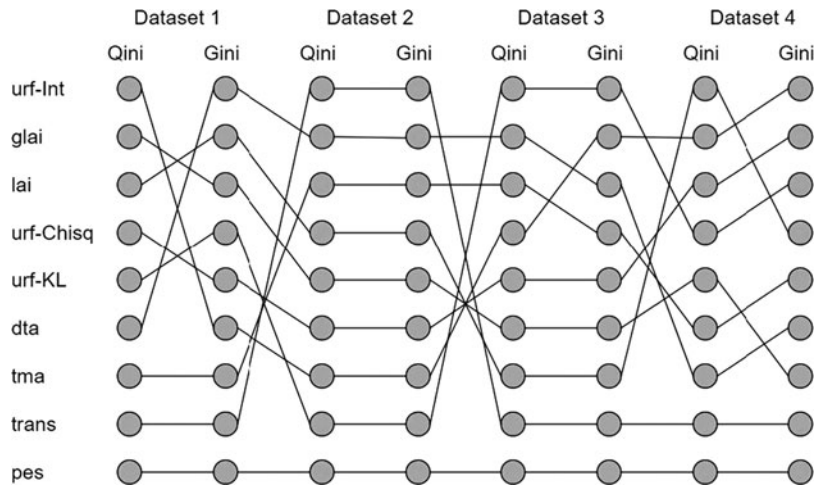


**FIG. 6.** Ranking of the uplift modeling techniques according to the Qini and Gini metrics across all four data sets.

However, in fact, neither do the existing uplift modeling techniques. This may be an important drawback and therefore represents a significant opportunity for further improving the performance. In addition, these measures arguably do not provide an intuitively interpretable indication on the performance of an uplift model from a practitioner's perspective. For example, it is not evident to explain what a Gini score of 0.5 means. The only intuitive understanding that is provided is that a higher Gini score is better, but for making business decisions these metrics give little insight into the actual use and utility an uplift model offers. Hence, an additional topic for further research is the development of intuitive and appropriate evaluation measures.

Visual analysis. In this section, an analysis using visual evaluation approaches is performed to provide further details with respect to the performance and relative comparison of performance of uplift models. Figure 7 shows the Qini plots for all techniques on all data sets. On data sets 1 and 4, some of the uplift techniques manage to achieve a considerable uplift improvement over targeting the entire population. As an example, an uplift of 3.5% can be observed on data set 1 when 60% of the population is targeted, whereas an uplift of 0.9% is achieved when targeting the full population. Notice that an improvement over targeting the full population can be achieved only when there are so-called do-not-disturbs in the population.

The following observations result from inspecting the various visual plots separately for the four data sets:

- Data set 1: The achieved uplift is considerably high for most techniques, but there is a varying degree of difference in performance between the techniques. Some techniques score less than others in terms of the maximum achieved uplift, whereas some achieve a high uplift but only when targeting a large fraction, that is, more than 50%, of the population. From a practical perspective, it is more interesting to realize as much uplift as possible for the smallest possible fraction. For large fractions of customers being selected, a substantial decrease in uplift, that is, downlift, is observed, which indicates

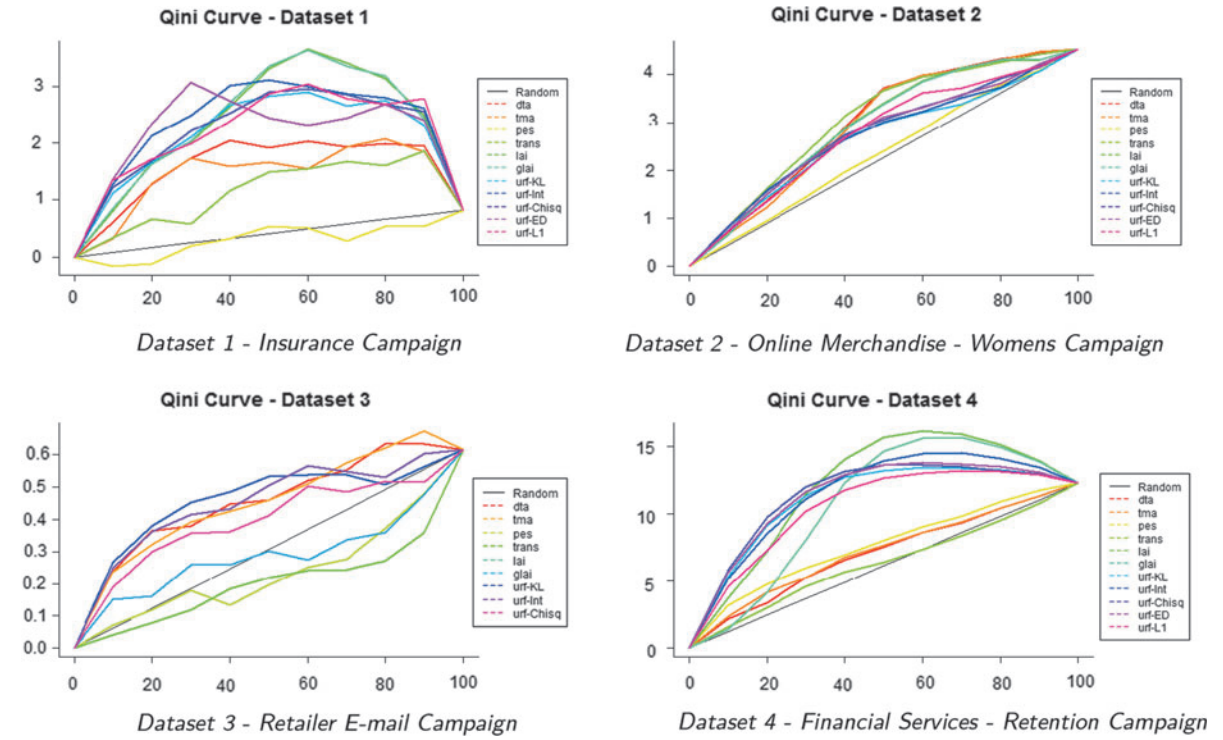


**FIG. 7.** Qini curves for four selected techniques on all data sets. Color images available online at www.liebertpub.com/big

a large number of *do-not-disturbs* to be among the customers. This is an important finding, as it warns us that targeting customers with this specific campaign may resort in significant losses, so care is to be taken in selecting customers.

- Data set 2: The uplift techniques are all close in terms of performance. Compared to the Qini curves of data set 1, it is striking that not a single technique achieves an uplift higher than the overall uplift. In other words, there is no downlift for higher fractions of customer being selected. This might indicate there are no or few *do-not-disturbs* in the customer population. A potential issue is that most uplift is obtained for larger fractions of selected customers, that is, for larger targeting depths. This is typically undesirable and may even be prohibitive for achieving profits, depending on the costs and benefits involved in targeting customers. Notice that these costs and benefits depend on the type of customers as discussed in the Uplift Response Modeling section, that is, whether a targeted customer is a *persuadable*, *sure thing*, *lost cause*, or *do-not-disturb*.
- Data set 3: The performance in terms of the Qini curve on data set 3 is highly variable and strongly depends on the technique. The dta and urf-int techniques are performing relatively well, however, the trans and glai techniques fail to do better than randomly targeting customers, as represented by the diagonal Qini curve. Notice that the techniques that are performing well appear to capture most of the achievable uplift, rank most of the *persuadables*, among the 20% highest ranked customers, that is, at 20%, 0.4 of the overall 0.6 uplift is already captured. Remember from the earlier discussions that we are uncertain about the number of *persuadables* and *do-not-disturbs*, and therefore, we cannot know how much uplift theoretically could be at most achieved (Surry PD, Radcliffe NJ; unpublished data).
- Data set 4: Similar to data sets 1 and 3, the performance of the models is highly variable. The dta and trans techniques barely achieve improvement over random targeting, whereas glai and urf-int perform considerably better. A maximum of 15% uplift is achieved when targeting around 60% of the population, which again is already a substantial fraction of the customer base. The Qini plots also clearly show that while the glai technique achieves the highest uplift, the urf-int technique achieves larger uplift values for smaller fractions, that is, at smaller targeting depths. Whereas glai achieves the maximum uplift at 60%, at 20% the technique does not perform much better than random targeting. On the contrary, the urf-int achieves about 8% uplift at only 20% of the population. This final observation indicates that the fraction of customers who should be targeted needs to be optimized, for maximizing the utility of the campaign.[47,48]

Supplementary Figure S1 shows the uplift per decile-charts for every technique (Supplementary data are available online at www.liebertpub.com/big). Ideally, these charts display strong uplift to the left and downlift to the right, indicating *persuadables* to be captured in the top deciles and *do-not-disturbs in the bottom deciles*. There is a direct relationship between the Qini curve and the uplift per decile plot. The higher the bars in the bar plot, the higher the Qini curve above the diagonal, since more uplift is achieved. If the uplift is high in the first decile and low in the last decile, then the Qini curve steeply increases for small selected fractions, because of the positive uplift in the first decile, and steeply decreasing for large selected fractions, because of the negative uplift or downlift in the last decile.

On the contrary, some techniques fail to include the *persuadables* in the top deciles of highest ranked customers, which are instead distributed across the ranking and deciles (Supplementary Fig. S1) leading to flat uplift per decile chart. Supplementary Figure S1 shows several such flat uplifts per decile plot, indicative for poor performance, although generally showing higher uplift to be achieved in the first deciles. This is true for all data sets, however, on data sets 1 and 4 these effects are more pronounced. For these data sets, we clearly observe a large uplift for the first deciles and even a downlift in the last deciles, indicating a good performance of the uplift model. Data sets 2 and 3 also show positive results, however, uplift is more distributed over the various deciles, instead of being concentrated in the top deciles. This means the uplift modeling techniques and models experience difficulties in accurately identifying persuadables. Finally, notice that in interpreting uplift per decile chart, it is not the absolute total amount of uplift that is of importance, rather the distribution of the uplift that is achieved as a result of treatment across the deciles. Also, if appropriate and permitted by the data, instead of uplift per decile as well uplift per quintile or any other percentile can be plotted. As such, the visual analysis of the

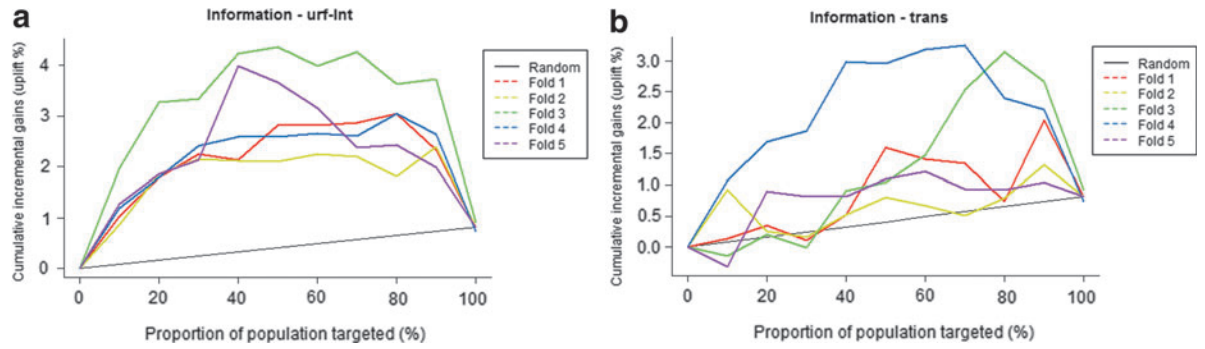


**FIG. 8.** Qini graph—performance of all folds for the urf-int and trans techniques. **(a)** Qini graph of the "uplift random forest" technique with the interaction-method as splitting method on dataset 1. Results on Fold 3 are significantly better than the other folds. **(b)** Qini graph of the "transformation" technique on dataset 1. Folds 1 and 3 perform worse than random targeting whereas Fold 5 achieves a good performance. Color images available online at www.liebertpub.com/big



performance can be further refined by increasing the resolution of the uplift chart.

Stability. In the Numerical Analysis section, large standard deviations were reported with respect to the average Qini values over the validation folds in the experiments. Notice that, in the experiments, a fivefold cross-validation setup was adopted allowing inspecting and analyzing the variation of the Qini curves across the various folds. A large standard deviation highlights a potential issue in terms of stability of the obtained uplift models, which directly relates to the generalization power and therefore is of crucial importance toward the quality and practical use of an uplift model.

For example, Figure 8a plots the five Qini graphs resulting from the fivefold cross-validation procedure of the uplift random forests approach with the interaction measure as splitting criterion (urf-int) for data set 1. It can be seen the Qini curve on fold 3 is well above the curves for the other folds, with the uplift around the fifth decile double the uplift on folds 2, 3, and 4. The curve on fold 5 displays a sudden peak around the fourth decile. Notice that the Qini curves are not smooth, although cumulative in nature, since calculated per decile to allow comparison between a sufficiently large set of observations from both the control and treatment groups.

Figure 8b plots the five Qini graphs of the transformation (trans) technique for data set 1. Here, the Qini curve on fold 4 is well above the curves on the other folds. Notice that, as can be seen from the curves on folds 1, 2, 3, and 5, the achieved uplift by selecting

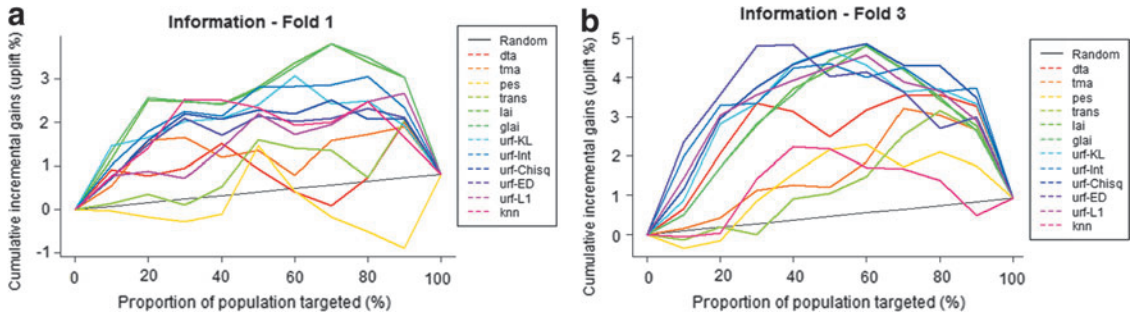


**FIG. 9.** Qini graph—performance of all techniques on a separate fold. **(a)** Qini graphs of all techniques on Fold 1 on dataset 1. **(b)** Qini graphs of all techniques on Fold 3 on dataset 1. Color images available online at www.liebertpub.com/big

customers using the uplift models does not lead to a higher uplift than achieved when randomly targeting the top 20% to 30% of the customers, as represented by the diagonal Qini curve. In general, for 8 of the 12 techniques that are applied on data set 1, the performance on at least onefold is worse than the baseline, that is, than random targeting.

Figure 9a and b provides complementary insight by showing the Qini curves of all techniques, on folds 1 and 3 of data set 1. There is a substantial variation within each plot, ranging from a performance that is worse than the baseline to substantial uplift being achieved, confirming the large variation in performance of the different techniques that was reported before. When comparing performance across the folds, we see that on fold 3, the performance generally is better, with different techniques on top compared with fold 1. Finally, notice that the technique achieving the highest uplift depends not only on the fold but also on the fraction of selected customers on the x-axis.

When analyzing the stability across data sets, it was observed that specifically for data sets 1 and 3, the instability is substantial and potentially prohibitive toward practical implementation and operation of the resulting models. A hypothesis that is to be tested in future research is that the instability of the uplift models is caused by a combination of three factors, that is, the size of the data, the size of the overall uplift observed when comparing the control and treatment groups, and the imbalanced class distribution that is typically observed in response as well as customer churn prediction, with typically many more nonresponders and nonchurners than responders and churners, respectively. In addition, the second type of imbalance between the control and treatment group, as discussed before, may add to the instability. Moreover, also the combinations of both imbalances may be a factor that needs to be considered (e.g., contacted customers who responded). In Lo's article,[49] the authors indicate that a large signal-to-noise ratio (i.e., uplift over the control response rate) is an important condition toward successfully applying uplift modeling. This statement, although not supported by empirical or theoretical evidence, aligns with the above hypothesis.

Notice that data set 1 contains only 10,000 observations and is the smallest data set that was analyzed. The uplift is equal to 0.82%, and the signal-to-noise ratio (i.e., the uplift over the control response rate) equals 4.21% (Table 5). Considering the fivefold cross-validation setup, each fold only contains 2000 observations, leading to a training set of 8000 observations and a test set with 2000 observations, of which half are control group and half are treatment group observations, of which 20% are responders, leading to 400 responders in the test control and test treatment groups, which clearly is few. Data set 3 contains much more observations, that is, 100,000, however, the initial uplift is very low as well and equal to 0.62. The signal-to-noise ratio is equal to 19.92%, which is well above the ratio for data set 1 (4.21%), but still considerably lower than the ratios for data sets 2 (42.61%) and 4 (−48.08%). Hence, the experimental results regarding the instability of uplift models support the above hypothesis. Obviously, further research is required into the nature and causes of instability, in function of problem and data set characteristics, as well as the inherent stability of uplift modeling approaches is to be further investigated. In addition, the development of *more* stable uplift modeling approaches is of crucial importance and a critical challenge toward a further practical adoption of these approaches in the industry. Very often only a limited amount of data are available, with an imbalanced class distribution and a limited uplift effect of a treatment. In such conditions, still uplift modeling should be applicable since it is more appropriate.

## Conclusion and Future Research

In recent years, driven by the limitations and shortcomings of predictive modeling approaches and looking to further improve the alignment with business needs and profitability, various uplift modeling techniques have been developed. Uplift models aim to estimate the net effect of a treatment on an outcome, as such allowing to optimize and prescribe the actions that should be undertaken to optimize the outcome. Uplift modeling, along with kindred approaches, has given rise to the field of prescriptive analytics. When applied in marketing for response modeling, uplift modeling aims at separating *true* from *baseline* responders. Whereas baseline responders will always respond, regardless of receiving a promotional offer, true responders only respond *because* of being targeted in a direct marketing campaign.

In this article, we presented an extensive survey of the literature on uplift modeling and proposed a framework that groups and characterizes the various uplift modeling approaches. We reviewed and contrasted regression-based and tree-based uplift modeling approaches, as well as straightforward transformation and indirect approaches. In addition, we discussed the evaluation metrics as adopted in the literature for assessing the performance of uplift models. In conclusion of the

literature study, we find there is a lack of intuitive measures that effectively provide insight into the performance and use of an uplift model from a business perspective, and that align with the business application and support decision-making. Specifically, the available evaluation approaches either assess the performance of an uplift model at an arbitrary cutoff or over the full spectrum of potential cutoffs.

A selection of the most prominent and powerful techniques, which were readily available or could be implemented following the technical details provided in the literature, were experimentally evaluated on four data sets with varying characteristics. All techniques were evaluated with standard evaluation metrics as used in the literature, such as the Qini coefficient,[45] the modified Gini coefficient [9] and the AUUC.[12] The results of the experiments, as presented in this article, highlight several issues and shortcomings, and also indicate that uplift modeling has great potential for increasing the returns on marketing investments. In line with the well-known no-free-lunch theorem, the experimental results indicate a large variability in terms of performance of the various uplift modeling approaches that were tested, with no clear winner across the four data sets. These techniques therefore can be concluded to be strongly data and application dependent, leading to the practical recommendation to test multiple approaches when developing an uplift model for practical business applications. One group of techniques that consistently perform well concern the ensemble approaches, such as the uplift random forests techniques. This is no surprise, since as well for predictive modeling it is found that ensemble approaches typically perform well and even best. Therefore, ensembles should be among the prime candidate approaches to test when developing an uplift model. It needs to be noted that a substantial number of ensemble approaches have been proposed in the literature, which are mainly different in terms of the adopted splitting criterion for learning the base uplift trees of the ensemble. Again, the experiments were inconclusive and no clear winner was found among the ensemble approaches.

As an important finding of the experimental study, we pinpointed the instability of uplift models as observed in the experiments for some of the data sets, and enlisted possible causes, among which the imbalanced class and group distribution as prime suspects. Particularly, the results for data sets 1 and 2 display a large variability, in terms of the performance that is observed across the folds in the fivefold cross-validation setup that was adopted. A key characteristic of both data sets 1 and 2 is the low initial uplift that is achieved by the campaign. In data set 1, the campaign is found to have achieved an uplift of 0.82, whereas in data set 3, an uplift of 0.62 is observed. A formal and deeper analysis of the potential causes for the instability is of key importance toward understanding and resolving the instability of uplift models. This is of essential importance toward the practical use of uplift models and therefore marked as a prime topic for future research. As a side note, although the overall uplift that was observed in data set 1 was small, the amount of uplift that could have been achieved in retrospect by adopting uplift modeling, however, was substantial. This is illustrative for the potential of uplift models for increasing returns on marketing investments, despite the described issues.

Finally, further research is required as well to extend and evaluate current uplift modeling approaches for multiple treatment setups (as mentioned in Radcliffe and Surry,[8] Rzepakowki and Jaroszewicz,[35] and Zhao et al.[16]), as well as for continuous target variables. In addition, the business value of uplift models may be further boosted by accounting for the actual use in terms of costs and benefits resulting from the prescriptions that are made by the uplift model.

## Author Disclosure Statement

No competing financial interests exist.

## References


1. Lo VSY, Pachamanova DA. From predictive uplift modeling to prescriptive uplift analytics: A practical approach to treatment optimization while accounting for estimation risk. J Market Anal. 2015;3:79–95.
2. Lustig I, Dietrich B, Johnson C, Dziekan C. The analytics journey. Anal Mag. 2010;3:11–13.
3. Smith RE, Swinyard WR. Information response models: An integrated approach. J Mark. 1982;46:81–93.
4. Coussement K, Harrigan P, Benoit DF. Improving direct mail targeting through customer response modeling. Expert Syst Appl. 2015;42: 8403–8412.
5. Siroker D, Koomen P. A/B testing: The most powerful way to turn clicks into customers. Hoboken, NJ: John Wiley & Sons, 2013.
6. Lo VSY. The true lift model: A novel data mining approach to response modeling in database marketing. SIGKDD Explor Newsl. 2002; 4:78–86.
7. Rzepakowski P, Jaroszewicz S. Decision trees for uplift modeling. In: Proceedings of the 2010 IEEE International Conference on Data Mining. ICDM'10. Washington, DC: IEEE Computer Society, 2010, pp. 441–450.
8. Radcliffe NJ, Surry PD. Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Edinburgh, UK: Stochastic Solutions, 2011.
9. Kane K, Lo VSY, Zheng J. True-lift modeling: Comparison of methods. J Market Anal. 2014;2:218–238.
10. Siegel E. Uplift modeling: Predictive analytics can't optimize marketing decisions without it. White Paper. Prediction Impact, 2011. Available online at https://www.predictiveanalyticsworld.com/patimes/uplift-modeling-predictive-analytics-cant-optimize-marketing-decisions-without-it/ (accessed March 24, 2015).
11. Hansotia B, Rukstales B. Direct marketing for multichannel retailers: Issues, challenges and solutions. J Database Mark. 2002;9:259–266.

12. Jaśkowski M, Jaroszewicz S. Uplift modeling for clinical trial data. In: ICML Workshop on Clinical Data Analysis. Edinburgh, UK: ICML Workshop on Machine Learning for Clinical Data Analysis, 2012.
13. Guelman L, Guillen M, Perez-Marin AM. Random forests for uplift modeling: An insurance customer retention case. In: Engemann KJ, Gil-Lafuente AM, Merigo J (Eds.): Modeling and Simulation in Engineering, Economics and Management. Vol. 115. Lecture Notes in Business Information Processing. Germany: Springer Berlin Heidelberg, 2012, pp. 123–133.
14. Porter D. Pinpointing the persuadables: Convincing the right voters to support Barack Obama. 2012. Available online at www.predictive analyticsworld.com/patimes/video-dan-porter-clip/ (last accessed April 12, 2016).
15. Issenberg S. How President Obama's campaign used big data to rally individual voters. Technol Rev. 2013;116:38–49.
16. Zhao Y, Fang X, Simchi-Levi D. Uplift modeling with multiple treatments and general response types. In: Proceedings of the 2017 SIAM International Conference on Data Mining. Houston, TX: SIAM. 2017, pp. 588–596.
17. Vittal S. Optimal Targeting through Uplift Modeling: Generating higher demand and increasing customer retention while reducing marketing costs. White Paper. Portrait Software, 2006. Available online at http://www.pbinsight.com/assets_microsite/resources/files/telenor-cs.pdf (accessed March 14, 2018).
18. Radcliffe NJ. Identifying who can be saved and who will be driven away by retention activity. White Paper. Stochastic Solutions Limited, 2007. Available online at http://stochasticsolutions.com/pdf/SavedAndDrivenAway.pdf (accessed March 24, 2015).
19. Radcliffe N, Surry P. Differential response analysis: Modeling true responses by isolating the effect of a single action. In: Credit Scoring and Credit Control IV Conference, Edinburgh, UK, 1999.
20. Larsen K. Net models. In: M2009—12th Annual SAS Data Mining Conference. Las Vegas, NV, 2009.
21. Hansotia B, Rukstales B. Incremental value modeling. J Interact Mark. 2002;16:35–46.
22. Lai LYT, Simon Fraser University (Canada). Influential marketing: A new direct marketing strategy addressing the existence of voluntary buyers. Canadian theses on microfiche. Simon Fraser University (Canada), 2006. Available online at https://books.google.be/books?id=5EvSuAAACAAJ (accessed June 1, 2015).
23. Shaar A, Abdessalem T, Segard O. Pessimistic uplift modeling. Available online at: http://arxiv.org/abs/1603.09738 (accessed April 12, 2016).
24. Guelman L, Guillen M, Pérez-Marín AM. Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study. Working Papers 2014-06. Universitat de Barcelona, UB Riskcenter, 2014. Available online at http://ideas.repec.org/p/bak/wpaper/201406.html (accessed November 4, 2016).
25. Pechyony D, Jones R, Li X. A joint optimization of incrementality and revenue to satisfy both advertiser and publisher. In: Proceedings of the 22nd International Conference on World Wide Web. www'13 Companion. Rio de Janeiro, Brazil: ACM, 2013, pp. 123–124.
26. Cao Y, Data A, Xu C. Untangle customers incrementality using uplift modeling with a case study on direct marketing. In: Midwest SAS Users Group Conference Proceedings (MWSUG 2017), St. Louis, Missouri, 2017, BF03.
27. Verbeke W, Baesens B, Bravo C. Profit-driven business analytics: A practitioner's guide to transforming big data into added value. Hoboken, NJ: John Wiley & Sons, 2017.
28. Chickering DM, Heckerman D. A decision theoretic approach to targeted advertising. In: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence. UAI'00. Stanford, CA: Morgan Kaufmann Publishers, Inc., 2000, pp. 82–88.
29. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. Commun Stat Theory Methods. 1994; 23:2379–2412.
30. Vansteelandt S, Goetghebeur E. Causal inference with generalized structural mean models. J R Stat Soc Series B Stat Methodol. 2003;65: 817–835.
31. Robins J, Rotnitzky A. Estimation of treatment effects in randomised trials with noncompliance and a dichotomous outcome using structural mean models. Biometrika. 2004;91:763–783.
32. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Boca Raton, FL: Chapman & Hall/CRC, 1984.
33. Quinlan JR. C4.5: Programs for machine learning. San Francisco, CA: Morgan Kaufmann Publishers, Inc., 1993.
34. Kass GV. An exploratory technique for investigating large quantities of categorical data. Appl Stat. 1980;29:119–127.
35. Rzepakowski P, Jaroszewicz S. Decision trees for uplift modeling with single and multiple treatments. Knowl Inf Syst. 2012;32:303–327.
36. Michel R, Schnakenburg I, von Martens T. Effective customer selection for marketing campaigns based on net scores. J Res Interact Mark. 2017; 11:2–15.
37. Soltys M, Jaroszewicz S, Rzepakowski P. Ensemble methods for uplift modeling. Data Min Knowl Discov. 2014;29:1–29.
38. Guelman L, Guillén M, Pérez-Marín AM. A decision support framework to implement optimal personalized marketing interventions. Decis Support Sys. 2015;72, 24–32.
39. Breiman L. Arcing classifier (with discussion and a rejoinder by the author). Ann Stat. 1998;26:801–849.
40. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
41. Breiman L. Bagging predictors. Mach Learn, 1996;24:123–140.
42. Zhao Y, Fang X, Simchi-Levi D. A practically competitive and provably consistent algorithm for uplift modeling. In: 2017 IEEE International Conference on Data Mining (ICDM). New Orleans, LA: IEEE, 2017, pp. 1171–1176.
43. Baesens B. Analytics in a big data world: The essential guide to data science and its applications. Hoboken, NJ: John Wiley & Sons, 2014.
44. Holland PW, Glymour C, Granger C. Statistics and causal inference. ETS Res Rep Ser 1985;1985:i–72.
45. Radcliffe NJ. Using control groups to target on predicted lift: Building and assessing uplift models. Direct Market J Direct Market Assoc Anal Council. 2007;1:14–21.
46. Crone SF, Lessmann S, Stahlbock R. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. Eur J Oper Res. 2006;173:781–800.
47. Verbeke W, Dejaegera K, Martens D, et al. New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. Eur J Oper Res. 2012;218:211–229.
48. Verbraken T, Verbeke W, Baesens B. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Trans Knowl Data Eng. 2013;25:961–973.
49. Lo VSY, Kane K, Zheng J. Identifying individuals who are truly impacted by treatment: Introduction to most recent advances in uplift modeling. In: Presented at the Bentley Analytics Virtual Symposium. 2014. Available online at https://www.researchgate.net/profile/Victor_Lo3/publication/270217235_Identifying_Individuals_Who_Are_Truly_Impacted_by_Treatment_Introduction_to_Recent_Advances_in_Uplift_Modeling/links/54a2dbbf0cf257a63604da2a/Identifying-Individuals-Who-Are-Truly-Impacted-by-Treatment-Introduction-to-Recent-Advances-in-Uplift-Modeling.pdf (accessed December 15, 2015).

**Cite this article as:** 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, DOI: 10.1089/big.2017.0104.

### Abbreviations Used

AUUC = area under uplift curve
CART = classification and regression trees
CHAID = chi-square automatic interaction detection
CN = control nonresponders
CR = control responders
CTS = context treatment selection
IV = information value
NIV = net information value
NWOE = net weights of evidence
TN = treatment nonresponders
TR = treatment responders
WOE = weights of evidence