To Treat, or Not to Treat: Reducing Volatility in Uplift Modeling Through Weighted Ensembles

Jannik Rößler
University of Cologne
roessler@wim.uni-koeln.de

Roman Tilly
University of Cologne
tilly@wim.uni-koeln.de

Detlef Schoder
University of Cologne
schoder@wim.uni-koeln.de

Abstract

When conducting direct marketing activities, companies strive to know whom to target with a marketing incentive to maximize the campaign effect. For example, which customer should receive churn prevention incentive to minimize overall churn rate? Uplift modeling is a promising approach to answer such a question. It allows to separate customers who would likely react positively to a treatment from those who would remain neutral or even react negatively. However, while different uplift approaches have been proposed, they usually suffer from high volatility and their performance often depends largely on data set and application context. Thus, it is difficult for practitioners and researchers to apply uplift modeling. To overcome these problems, we propose a weighted ensemble of different uplift modeling approaches to reduce volatility and improve robustness. We evaluate the novel approach against single uplift modeling approaches on multiple data sets and find that the ensemble is indeed more robust.

1. Introduction

Direct marketing refers to marketing activities that are tailored to an individual recipient, as opposed to mass marketing, which targets larger groups with the same activity [2]. A central goal in direct marketing is to optimize the overall performance of marketing activities, that is, maximizing the positive behavioral change (e.g., click-through, conversion, sales, non-churn) from all individual marketing activities and reducing their adverse effects on recipients' behavior (e.g. scattering loss, churn) [15]. This applies to digital marketing activities, such as online or email advertisements, related product offerings in e-commerce, cross- or up-selling offers, as well as to analog marketing activities, such as individual print mailings or door-to-door offerings.

Uplift modeling particularly aims at using data and predictive models to target the right individuals for a marketing activity in order to improve the overall marketing performance. More generally, uplift modeling estimates the differential effect of a treatment on a specific behavior of recipients, that is, the change in probability for a recipient to exhibit a specific behavior, that is caused by the treatment [14]. Thus, uplift modeling can identify those individuals who are likely to respond in the desired direction if targeted with a treatment (such as, a marketing activity), but unlikely to respond otherwise or even respond negatively [17].

Three general approaches have been proposed in the uplift modeling literature, namely, two-model approach, class transformation approach, and direct uplift approach [13]. While each of these approaches has been successfully demonstrated in some cases, literature disagrees about which approach performs best [7, 17]. For example, while Radcliffe and Surry [21] claim that the two-model approach is rarely working well for real-world problems, the analysis by Gubela et al. [12] slightly favors the two-model approach, though no approach is found to be generally superior over the others. Devriendt et al. [7] argue for the direct uplift approach while at the same time acknowledging its high volatility in performance across various data sets. In general, researchers agree that uplift modeling suffers from high volatility in terms of prediction performance across *different* data sets, but also across different cross-validation folds of the *same* data set [1, 7, 12]. Moreover, the performance of each approach depends on parameters such as the size of the data set, the relative sizes of treatment and control groups, or the treatment ratio [8, 12, 19]. These problems make it extremely difficult to generalize the performance of different uplift models, leading some authors to the conclusion that there is no "single method that works the best for all data sets" [17].

One common technique to reduce error in prediction performance is to use an ensemble, a combination of multiple algorithms. Ensembles are considered to solve a plethora of challenges because of their premise to

URI: https://hdl.handle.net/10125/70805
978-0-9981331-4-0

compensate the error of a single algorithm by using multiple algorithms [24]. Thus, we propose to combine the three uplift modeling approaches into one, weighted ensemble. Subsequently, this study strives to address the following research question:

RQ: Can the volatility in uplift prediction performance be reduced by using an ensemble of different uplift approaches?

The remainder is organized as follows. We first describe related work in section 2. The research methodology, including the proposed method, data and metrics are presented in section 3. Results are shown in section 4 while section 5 contains the discussion. Section 6 concludes this paper with limitations, future work, and implications.

2. Related Work

We review relevant literature with respect to, first, the approaches used in the uplift modeling context and their respective advantages and disadvantages, second, the problem of volatility of prediction performance for uplift modeling in general and third, ensemble learning.

2.1. Uplift Modeling

Despite its practical relevance for researchers and practitioners in marketing, uplift modeling has yet received rather little attention [28]. Typically, marketing campaigns are based on traditional response modeling [6], although the success of uplift modeling has been shown several times [22, 26].

Uplift modeling was introduced by Radcliffe [20] in 1999 under the term 'differential response analysis' and ever since, various terms were used instead, such as, 'uplift modeling', 'true lift modeling', or 'incremental value modeling' [7]. Uplift modeling differs from traditional response modeling in that it models the *change* in response behavior that can be *attributed to* the treatment, while traditional response modeling models the 'gross' response behavior when the treatment is applied, including response caused by the treatment as well as auto-response.

The fundamental problem with estimating the causal effect of the treatment is that usually an individual can only be either treated or not treated and, thus, the treatment-induced behavior change cannot be observed within one individual. To overcome this problem, uplift modeling uses data from randomized controlled experiments in which the population under study has been split into two different subpopulations, one which is subject to the treatment (e.g., marketing activity) and another one which is not subject to any treatment. Groups are usually referred to as treatment group and control group, respectively, and the latter serves as the baseline for response behavior against which the behavior in the treatment group is evaluated [22]. The randomized controlled experiment is not only used to calculate a treatment's overall success by means of comparison to a control group (also called the *average* treatment effect), but it can also be utilized by uplift modeling to estimate the *heterogeneous* treatment effect (HTE) for a given individual, that is, how particular individuals respond differently to the treatment [11].

Three main approaches have been proposed to estimate the individual heterogeneous treatment effect[1] [13]. First, the *two-model approach* estimates two predictive models for response behavior, one for the treatment group and another one for the control group. Subsequently, the individual uplift can be predicted by subtracting the estimate of the control group model from the estimate of the treatment group model:

$$\begin{aligned} Uplift = P(outcome|treatment) \\ - P(outcome|control) \end{aligned} \quad (1)$$

For both models, any machine learning algorithm can be used, such as Random Forest [5] or Support Vector Machines [3]. While the advantage resides in its simplicity, many authors claim that the two-model approach can be outperformed by other approaches [23, 26]. Radcliffe and Surry [21] even argue that the two-model approach "rarely works well for real-world problems" as the difference in behavior is neglected between the two populations. Further, both models can contain errors, which can be amplified when aggregated to predict uplift [7]. However, a recent benchmarking of different modeling strategies showed that the two-model approach represents a modeling strategy of first choice [12].

Second, the *class transformation approach* works by recoding the information whether a person was treated or not and whether the person responded or not [1, 16]. Generally, four different cases can be observed in the experiment [17]:

- *Control responders*: Individuals who respond without being subject to any treatment.
- *Control non-responders*: Individuals who did not respond, nor did they receive a treatment.
- *Treatment responders*: Individuals who responded while being subject to a treatment.

[1] In the remainder we will use the term uplift instead of heterogeneous treatment effect. Note that also other terms exist like conditional average treatment effect (CATE) or causal effect.

- *Treatment non-responders*: Individuals who did not respond although they received a treatment.

Using this distinction, class transformation creates a new target variable by using a mathematical operation on the treatment and response variable. For example, Jaskowski and Jaroszewicz [16] defined the new target variable $Z \in \{0,1\}$ like the following:

$$Z = \begin{cases} 1 \; if \; T = 1 \; and \; Y = 1 \\ 1 \; if \; T = 0 \; and \; Y = 0 \\ \quad 0 \; otherwise \end{cases} \qquad (2)$$

$T \in \{0,1\}$ is representing the treatment variable with $T = 1$, if the person was subject to a treatment and $T = 0$ if not, and $Y \in \{0,1\}$ is representing the response variable with $Y = 1$, if the person responded and $Y = 0$ if not. Thus, the authors decided to create a new variable Z which equals one if the individual had been treated and $Y = 1$ or if the individual had not been treated and $Y = 0$. Otherwise, Z equals zero. To understand the idea behind this transformation, we need to comprehend the advantage of uplift modeling over other predictive models, which is to differentiate the following four groups [17]:

- *Sure things*: Individuals who react in the desired way (from a company's perspective) with or without the treatment
- *Lost causes*: Individuals who do neither respond in case they are treated nor in case they are not .
- *Do-not-disturb*: Individuals who react negatively to the treatment, for example, do churn in case they are treated but do stay if not treated.
- *Persuadables*: Individuals who react in the desired way only if being subject to the treatment.

Jaskowski's and Jaroszewicz's [16] then label the control non-responders ($T = 0, Y = 0$) and treated responders ($T = 1, Y = 1$) group as positive and the control responders ($T = 0, Y = 1$) and treated non-responders ($T = 1, Y = 0$) group as negative. The intuition is that the positive group contains all persuadables, some lost causes, and some sure things, while the negative group contains all do-not-disturbs as well as the remaining lost causes and sure things. The advantage of such an approach is that due to the transformation into a single, binary target variable, any machine learning algorithm can be applied. However, Jaskowski and Jaroszewicz found that the performance of different approaches largely depends on the data set. While the two-model approach outperformed the class transformation twice, the latter achieved better results once. Similar results were found in [1]. The authors used a slightly different variation of the class transformation approach and evaluated it against other approaches, such as the previously explained two-model approach, but without a clear winner.

Lastly, the *direct uplift approach* relies on modified machine learning algorithms to infer uplift directly. According to the current literature, decision trees and different ensembles of decision trees are the most popular adapted algorithms [14, 21, 22, 26]. For example, Hansotia and Rukstales [14] modified the CHAID algorithm to choose in each branch the partition that results in the greatest difference in incremental response rates between the two resulting nodes. Specifically, the incremental response rate Δp is calculated for each node by calculating the difference between the response rate in the control group p^C and the response rate in the treatment group p^T. Subsequently, the difference in the incremental response rates for both nodes ($L \coloneqq left \; and \; R \coloneqq right$) is calculated by subtracting the incremental response rate of the right node from the incremental response rate of the left node:

$$\begin{aligned} \Delta(\Delta p) &= \Delta p_L - \Delta p_R \\ &= (p_L^T - p_L^C) - (p_R^T - p_R^C) \end{aligned} \qquad (3)$$

The algorithm chooses the partition in each splitting step such that $\Delta(\Delta p)$ is maximized.

An alternative splitting criterion has been proposed by Rzepakowski and Jaroszewicz [22]. It is based on information theory and uses a distribution divergence, that is, Kullback-Leibler divergence or Euclidean distance. The distribution divergence is used to build a tree in which the distributions of the target variable differ as much as possible between treatment and control groups. Thus, the goal is to maximize the gain in information of a split which is calculated by subtracting the divergence of the parent node from the conditional divergence of each child node:

$$\begin{aligned} D_{gain}(A) = \sum_a \frac{N(a)}{N} D(P^T(Y|a) \colon P^C(Y|a)) \\ - D(P^T(Y) \colon P^C(Y)) \end{aligned} \qquad (4)$$

where a refers to each of the child nodes, N refers to the total number of instances in the parent node, and $N(a)$ refers to the number of instances in the child node a. $P^T(Y)$ and $P^C(Y)$ are the outcome class distributions for treatment and control groups, respectively. The authors show that their approach is not only superior to the two-model approach but that it also outperforms the approach by Hansotia and Rukstales. The direct uplift approach was extended by Soltys et al. [26] to two different ensemble methods, namely Bagging and Random Forest. The authors achieved excellent results, exceeding the performance of other uplift modeling approaches, including the two-model approach and the direct uplift approaches mentioned above. Their approach differs from ours as they apply an ensemble only to the direct uplift approach while we use an ensemble based on every uplift approach, namely two-model, class transformation, and direct uplift. The data sets used by Rzepakowski and Jaroszewicz [22] and Soltys et al. [26] are less than ideal for uplift modeling. Since the data sets had not been collected for uplift modeling in the first place, they did not contain actual response / target variables which is why the authors arbitrarily selected one of the features as target variable in order to make the data set suitable to uplift modeling. Further, the chosen data sets were rather small or very small with number of instances ranging from 57 to 12,960 (median: 569), which is problematic for virtually any machine learning algorithm. Hansotia and Rukstales [14], on the other hand, only used a single data set to evaluate the performance of their classifier.

2.2. Volatility in Uplift Modeling

One of the most important problems in uplift modeling is the lack of suitable, publicly available data sets [22]. Without appropriate data sets, it is almost impossible for researchers to make fair comparisons between different uplift modeling approaches and, even worse, it is almost impossible to derive drivers and factors of a classifiers' performance. Especially in the uplift modeling context, researchers have shown several times that the performance of different uplift modeling approaches is highly volatile and largely depends on the data set, its size, the ratio between treatment and control size, or the response ratio [7, 8, 12, 17]. For example, Kane et al. [17] evaluated two two-model approaches and two class transformation approaches and found that the class transformation approach performed better. However, they summarized that their results were not generalizable to other data sets. Finally, they concluded that there "may not be a single method that works the best for all data sets".

Devriendt et al. [7] found similarly that various approaches perform well on some data sets but worse on others. In a more extensive evaluation, they compared ten different classifiers, including various two-model, class transformation, and direct uplift approaches on four different data sets. The direct uplift random forest worked well on three of four data sets while the two-model approach also exhibited good performance in half of the cases. The class transformation approach was the most volatile, scoring poorly three times, but then again very good once. Further, most of the approaches also exhibited large volatility across different cross-validation folds of the *same* data set. The authors conclude that "the experimental results indicate a large variability in terms of performance of the various uplift modeling approaches […], with no clear winner" [7:40] and that the results demonstrate a strong dependency on data and application.

The most recent benchmarking of various uplift modeling approaches is provided by Gubela et al. [12]. The authors evaluated eight different uplift modeling techniques, including one two-model approach and various class transformation and direct uplift approaches. Unfortunately, they did not consider the direct uplift approaches proposed by Hansotia and Rukstales [14], Rzepakowski and Jaroszewicz [22] or Soltys et al. [26]. Their experimental setup involved as much as 27 data sets from several digital marketing campaigns. The overall uplift of the data sets ranged from -2.24% to 3.60% and their size from 3,204 to 1,199,581 cases. Treatment and control response rates ranged from 2.81e-3% to 0.56% and from 1.51e-3% to 0.53%, respectively. With their comprehensive collection of diverse data sets, the authors provide a good foundation for a comparison of various uplift modeling techniques. Their results substantiate the assertion that predictions of uplift models are highly volatile and data-dependent. None of the eight proposed techniques was superior in every data set and there was substantial volatility across the different approaches.

To summarize, there is no single uplift modeling approach which consistently outperforms the others. Existing methods are not generalizable across data sets but rather data and application dependent and suffer from high volatility. From a business perspective, this renders uplift modeling rather impractical as companies would probably have to evaluate all available approaches for every new campaign, which consumes time and resources. Hence, there is strong demand for a more robust and more widely applicable uplift modeling approach.

2.3. Ensembles

An ensemble is a well-known technique to reduce generalization error by training several different models separately and have all of them vote on the test sample. Finally, the average of all votes is taken [10]. This technique is also known as *model averaging* [10]. The premise is that the error of a single algorithm will likely be compensated by the other algorithms, reducing the overall generalization error and increasing predictive performance [24].

This improvement in performance can be explained by two reasons [24]. First, for small data sets in particular, a single algorithm is prone to predict all of the training

Table 1. Overview of approaches used in the experiment

Uplift modeling approach	Base learner	Parameters for base learner	Source
Two-Model	Random Forest	Number of estimators: 200 Max depth: 25 Max features: auto	e.g. [12]
Class Transformation			[16]
Direct Uplift			[26]
Weighted Ensemble	Direct Uplift + Two-Model + Class Transformation	see above Distribution Divergence: Euclidean Distance	

data perfectly while failing to fit unseen instances. To circumvent this disadvantage, an ensemble averages many different predictions reducing the risk of selecting a single, incorrect hypothesis. Second, a single algorithm might not be able to create the optimal hypothesis as it is outside the feature space of the algorithm. By using many algorithms, the feature space is extended and hence, the optimal hypothesis is more likely to be found

Another advantage of using an ensemble method is that class imbalances can be mitigated [24]. For example, Nikulin et al. [18] propose to train each algorithm of an ensemble on a different, balanced subset of the training data in order to cope with imbalances. This is particularly useful in an uplift modeling context because data suffers from high imbalances in class distribution [7]. Usually there are many more non-responders in a marketing campaign than responders.

The effectiveness of ensemble methods has been shown several times. For example, Fernández-Delgado et al. [9] found in an extensive evaluation using 179 classifiers and 121 data sets, that the random forest [5], a kind of ensemble approach, is among the best performing algorithms. Vafeiadis et al. [27] showed that boosting [4], another kind of ensemble method, can clearly outperform non-boosted algorithms in a customer churn prevention case.

3. Research Methodology

To overcome the problems of high volatility, data and application dependency, we propose to combine two-model, class transformation, and direct uplift approaches into a weighted ensemble. The performance of the ensemble is then evaluated against other existing approaches, using nine data sets.

3.1. Weighted Ensemble Approach

The weighted ensemble combines the predictions from three uplift base models, namely a model based on the two-model approach, a class transformation model, and a direct uplift model. Each of these base models is itself based on Random Forest as a base learner with the following hyperparameters: *Number of estimators:* 200, *max depth:* 25 and *max features*: auto. We chose Random Forest as a base learner for each approach as Gubela et al. [12] found that it is the most promising algorithm when working with uplift modeling. The hyperparameters were fixed to facility comparability between different approaches. The values were chosen after several tests to avoid overfitting.

For the two-model approach, both models use Random Forest as a base learner. The further procedure of this technique has already been described in section 2. The class transformation approach is based on Jaskowski's and Jaroszewicz's mathematical operation [16], as described in section 2. After creating a new target variable, Random Forest is used to train a model. For these two approaches, cost-sensitive learning is applied in order to cope with imbalanced data sets. Here, weights are calculated according to the relative appearance of each class in the data set. The more underrepresented a class is, the higher the weight. These weights are then embedded into the learning algorithm, which makes the model more suitable for learning from very imbalanced data [25]. The direct uplift approach is based on the work of Soltys et al. [26], who used the splitting criteria proposed by Rzepakowski and Jaroszewicz [22] but in a Random Forest rather than in a single decision tree. Euclidean distance is used as a distribution divergence. While the first two approaches were implemented using the sklearn[2] package, the direct uplift approach was implemented using the causalml[3] package.

To train the weighted ensemble prediction model, the three base models are combined according to their relative prediction performance along the following steps.

1) Base model training. Each of the base models is trained using a training data set. Stratification is applied when splitting data sets to preserve treatment-, control- as well as response-ratios.

[2] https://scikit-learn.org/

[3] https://github.com/uber/causalml

Table 2. Overview of data sets used in the experiment

Id	Name	Number of features	Number of samples (treatment/control)	Treatment Response Rate	Control Response Rate	Source
1	Hillstrom	10	64,000 (42,694/21,306)	0.167	0.106	[4]
2	Hillstrom/Women	10	42,693 (21,387/21,306)	0.151	0.106	[4]
3	Starbucks	10	126,184 (63,112/63,072)	0.017	0.007	[5]
4	Customer Acquisition	286	9,974 (6,193/3,781)	0.111	0.065	Private
5	Churn Prevention	44	10,097 (6,684/3,413)	0.662	0.673	Private
6	Churn Prevention/A	44	6,754 (3,341/3,413)	0.673	0.673	Private
7	Churn Prevention/B	44	6,756 (3,343/3,413)	0.650	0.673	Private
8	Criteo	14	25,309,482 (21,408,827/3,900,655)	0.002	0.002	[8]
9	Criteo Resampled	14	7,797,062 (3,896,407/3,900,655)	0.002	0.002	[8]

2) Calculating qini coefficients. Qini coefficients q are computed based on a validation set for all base models m_i to measure their performance, respectively. The qini coefficient is a common performance metric in uplift modeling [7, 12]. It is defined as the ratio of the area under the actual qini curve and the diagonal, corresponding to random targeting, to the area under the optimal qini curve and the diagonal. This value ranges from -1.0 to 1.0. The qini curve is the cumulative difference in response rate between treatment and control group. It is calculated on a per-segment basis in descending order. The optimal qini curve ranks treatment responders first, treatment non-responders second, control responders third and treatment responders fourth.

3) Weighting. Qini coefficients are used to calculate the weights as follows:

$$w_i = \frac{q_i}{\max\limits_{i}(q_i)} \tag{5}$$

Thus, the weight of each base model corresponds to its performance relative to the best base model. The best base model with the highest qini coefficient receives a weight of 1, while the other base models receive smaller weights. The weights are then normalized with min-max normalization to avoid negative values.

4) Score normalization. The individual scores predicted by each base model m_i for a given case X are normalized to [0,1] using min-max normalization on the test data T:

$$m_i(X)' = \frac{m_i(X) - \min\limits_{X\in T}\big(m_i(X)\big)}{\max\limits_{X\in T}\big(m_i(X)\big) - \min\limits_{X\in T}\big(m_i(X)\big)} \tag{6}$$

This is necessary as each base model returns different ranges of scores and, thus, combining the scores without normalization would be biased towards the base model with the highest range in scores.

5) Ensemble model. Lastly, the ensemble model E is established as a weighted combination of the base models as follows:

$$E(X) = \sum_{i} m_i(X)' * w_i \tag{7}$$

3.2. Evaluation

Nine real-world data sets were used to evaluate model performances (see Table 2). Some of the data sets are publicly available (see given references for more details regarding these data sets) and some of the data sets were obtained from companies.

The Hillstrom data set is an email marketing campaign from MineThatData[2]. Like other researchers [7, 17], we considered online visits as the response and we distinguished between two data sets: one that contained both treatments (men's and women's merchandise) and another that only contained the treatment featuring women's merchandise.

The Starbucks data set was provided by Udacity as part of their Data Scientist Nanodegree. It was made public in a blogpost [3].

We obtained private data sets from a marketing campaign to acquire new customers in the retail industry as well as from a churn prevention campaign from a company with fixed term contracts and auto renewal. The churn prevention campaign includes two different treatments (A and B), for which we also created separate data sets. Both campaigns contained continuous and categorical variables covering socio-demographic information and campaign details. Additionally, the

[4] https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html

[5] https://github.com/joshxinjie/Data_Scientist_Nanodegree/tree/master/starbucks_portfolio_exercise

churn prevention campaign included some consumer behavior data. None of the campaigns contained customer relationship information such as a customer life-cycle value.

The Criteo data set was made available by Diemert et al. [8]. As the treatment-control-ratio is around 5:1 we also created another, more balanced data set by resampling the Criteo data set down to a treatment-control-ratio of about 1:1.

To assess the performance of the four approaches (i.e., two-model, class transformation, direct uplift, ensemble), models were trained/validated and tested using 10-fold cross-validation for data sets one through seven. First, the data set was split using 10-fold cross-validation into 90% training/validation and 10% test data. The training/validation set was again split into training (80%) and validation (20%) data sets. Thus, 72%, 18% and 10% of the data set were used for training, validation and test, respectively. See Figure 1 for an illustration. For the Criteo Resampled data set we used 6-fold cross-validation as it is large enough to avoid taking into account more folds. For the (full) Criteo data set we omitted cross-validation because it was more than three times larger than the Criteo Resampled data set. Instead we used a single stratified split into 64% training, 16% validation and 20% test data.

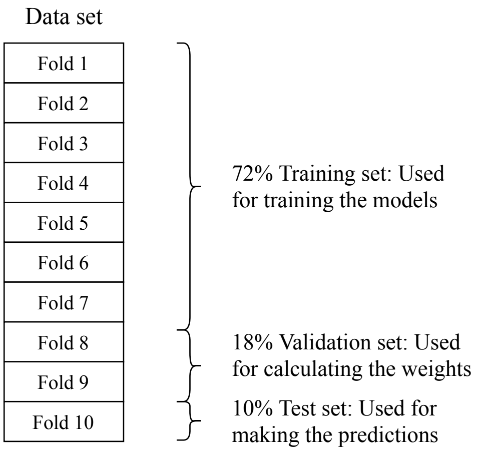
Data set
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Fold 6
Fold 7
Fold 8
Fold 9
Fold 10
72% Training set: Used for training the models
18% Validation set: Used for calculating the weights
10% Test set: Used for making the predictions

Figure 1. Illustration of training, validation and test split

For each model, predictions were computed for the test data sets and qini coefficients were calculated. To measure volatility of the different models, two metrics were used. First, to measure the volatility of approaches across *different* data sets, we calculated the average qini coefficient of all cross-validation folds for each data set and approach - except for Criteo for which no cross-validation had been conducted and hence only one qini was available. Next, the (average) qini coefficients of all approaches for each data set were normalized using min-max normalization to obtain relative values that are comparable across data sets.

Second, to measure volatility for different splits of the *same* data set, the standard deviation of qini coefficients of the cross-validation folds for each data set and approach were calculated. Next, standard deviations of all methods for each data set were normalized using min-max normalization to obtain relative values that are comparable across data sets.

Finally, to compare different approaches, we calculated a single number metric, which was the average for each approach across all data sets, for both, average qini coefficient and standard deviation of qini coefficient.

4. Results

4.1. Volatility across different data sets

According to the average value across all data sets, the class transformation approach performed worst with a value of 0.241 (difference to best approach: 0.646), followed by the two-model approach with a score of 0.501 (0.386). The direct uplift approach performed slightly better with a score of 0.633 (0.254) and the weighted ensemble approach performed best with a mean of 0.887. The following in-depth analysis confirms these findings.

The class transformation approach performed worst, as its performance was lowest and second lowest four times each. Only on the Churn Prevention/B data set it achieved a relative average qini coefficient of 0.931; slightly worse than the ensemble approach with a difference of only 0.069.

The performance of the two-model approach was remarkably volatile. It produced the best average qini coefficient on three data sets (Churn Prevention/A, Criteo, Criteo Resampled), but also the worst three on other data sets (Churn Prevention/B, Hillstrom, Customer Acquisition). It took third place once (Hillstrom/Women) and second place once (Churn Prevention).

A similar picture was observed for the direct uplift approach. While it achieved the best average qini coefficient four times, it scored worst twice, and second worst twice. Once the direct approach took second place with an average qini coefficient of 0.839 (Criteo).

The newly proposed weighted ensemble approach had remarkably low volatility across different data sets. It had the highest average qini coefficient on two data sets (Churn Prevention, Churn Prevention/B) and the second highest on all remaining data sets. Further, it was only slightly inferior to the best approach on four data sets: Hillstrom/Women (average qini coefficient: 0.917

Table 3. Average qini coefficient for each approach across different data sets

Approach	Hillstrom	Hillstrom/Women	Starbucks	Churn Prevention	Churn Prevention/A	Churn Prevention/B	Customer Acquisition	Criteo Resampled	Criteo	Average
Ensemble	0.836	0.917	0.981	1.000	0.730	1.000	0.610	0.912	0.996	0.887
Direct uplift	1.000	1.000	1.000	0.000	0.000	0.154	1.000	0.705	0.839	0.633
Two-model	0.000	0.174	0.600	0.731	1.000	0.000	0.000	1.000	1.000	0.501
Class transformation	0.128	0.000	0.000	0.082	0.634	0.931	0.392	0.000	0.000	0.241

Note: Higher values are better first second

Table 4. Standard deviation of qini coefficient for each approach and data set

Approach	Hillstrom	Hillstrom/Women	Starbucks	Churn Prevention	Churn Prevention/A	Churn Prevention/B	Customer Acquisition	Criteo Resampled	Criteo	Average
Ensemble	0.000	0.017	0.439	0.000	0.000	0.156	0.301	0.298	*n/a*	0.151
Direct uplift	0.767	0.000	0.356	0.972	0.315	0.502	0.000	0.000	*n/a*	0.364
Two-model	0.115	0.222	0.000	1.000	0.024	1.000	1.000	1.000	*n/a*	0.545
Class transformation	1.000	1.000	1.000	0.177	1.000	0.000	0.843	0.007	*n/a*	0.628

Note: Lower values are better first second

/ difference: 0.083), Starbucks (0.981 / 0.019), Criteo (0.996 / 0.004), and Criteo Resampled (0.912 / 0.088).

The results for volatility of all approaches across different data sets are summarized in Table 3. The higher the average qini coefficient, the better the approach

4.2. Volatility across different folds

The results for volatility of the approaches across different cross-validation folds of the same data set give a similar picture. According to the average value across all data sets, the class transformation approach performed worst with a value of 0.628 (difference to best approach: 0.477), followed by the two-model approach with a score of 0.545 (0.394). The direct uplift approach performed slightly better with a score of 0.364 (0.213) and the weighted ensemble approach performed best with a mean of 0.151.The following in-depth analysis confirms these findings.

The class transformation approach performed worst as the standard variation was the highest on four data sets and second-highest in another data set (Customer Acquisition with 0.843). On three data sets (Churn Prevention/B, Churn Prevention, Criteo Resampled) did class transformation produce the lowest or second lowest standard deviation of qini coefficients, respectively.

The two-model approach did slightly better than class transformation as its standard deviation was among the lowest on three data sets (Churn Prevention/A, Hillstrom, and Starbucks). However, it performed poorly on the Churn Prevention, Churn Prevention/B, Customer Acquisition and Criteo Resampled data sets.

The direct uplift approach yielded the lowest standard variation on three data sets (Hillstrom/Women, Customer Acquisition, Criteo Resampled). However, among the remaining data sets, it performed mediocre or poorly.

The most robust approach was again the weighted ensemble. It achieved the lowest standard deviation on three data sets (Churn Prevention, Churn Prevention/A, Hillstrom), like no other approach, and second-lowest on another three (Churn Prevention/B, Hillstrom/Women, Customer Acquisition). Further, on Hillstrom/Women, it was only marginally worse than the best approach (difference: 0.017). The approach was inferior to two other approaches only on the Starbucks and the Criteo Resampled data set. Measurements were not available for Criteo because no cross-validation had been conducted on this data set (see 3.2).

The results for volatility for all approaches across different cross-validation folds of the same data set are summarized in Table 4. The lower the standard deviation of the qini coefficient, the better the approach.

5. Discussion

We have shown that our weighted ensemble approach can reduce the volatility in uplift prediction performance. Across different data sets as well as across different cross-validation folds the performance of the proposed approach was more robust than other single approaches.

Ensembles are considered to solve a plethora of challenges because of their ability to reduce the generalization error. Our study does not only support this finding but also reveals the successful application of ensembles in the uplift modeling context. Further, we showed that single approaches such as two-model, class transformation and direct approach suffer from volatility supporting the findings of other researchers.

By reducing the generalization error, the predictions became far more robust such that the results were more reliable and applicable.

6. Conclusion and Future Work

Previous studies on uplift modeling have shown high volatility of different approaches across different real-world data sets and even across different cross-validation folds of the same data set. Existing approaches have been found to be highly data and application dependent and cannot be generalized well.

In this study, we proposed a weighted ensemble approach that combines two-model, class transformation, and direct uplift approaches to tackle these issues. We evaluated the weighted ensemble approach against existing approaches on nine real-world data sets.

The results have shown that the ensemble approach is far more most robust across different data sets as well as cross-validation folds of the same data set. Therefore, we conclude that our ensemble approach provides a promising solution to cope with high volatility in the uplift modeling context.

Nevertheless, our study is not without limitations.

First, the evaluation of an uplift modeling approach is generally difficult [7]. Although many researchers fall back on qini coefficient and qini curve as evaluation metrics, they both entail some flaws. For example, it is questionable what an optimal qini curve looks like as an individual can only be either treated or not treated and thus, we do not observe the treatment-induced behavior change. Further, such metrics are not intuitively interpretable making it difficult to derive implications for business decisions.

Second, although we used nine data sets, it can still be questioned whether our results can be generalized to other data sets. As already mentioned by other researchers, the number of publicly available uplift modeling data sets is low [22]. Further research should evaluate the ensemble approach on other data sets, such as the diverse collection by Gubela et al. [12].

Despite these limitations, our results suggest a first way to cope with the high volatility. One crucial step for future research is to evaluate other ensembles with different combinations of methods and base learners. Instead of using the class transformation approach proposed by Jaskowski and Jaroszewicz [16], one could use the approach introduced by Athey and Imbens [1]. The same applies to the direct uplift approach. One could use the technique proposed by Hansotia and Rukstales [14] rather than by Soltys et al. [26]. Further, as Gubela et al. [12] show, different base learners and their hyperparameters also play an important role in the performance of uplift modeling approaches. Thus, instead of using Random Forest as a base learner, the performance could be analyzed using Support Vector Machines, Linear Regression, or similar. Other ways of reducing volatility are also highly encouraged, for example, through the use of feature engineering.

Further, there is need for more publicly available reference / benchmark data sets for uplift modeling, not only for research on reducing volatility, but to improve uplift modeling research in general. Besides using real-world data sets, it might also be worthwhile to further investigate ways to generate synthetic data sets as mentioned by Radcliffe and Surry [21].

Marketing practitioners and analysts in charge of marketing campaigns can be informed through our study about the role of ensembles in uplift modeling. Our results suggest that ensembles help reduce the volatility which is highly useful in an uplift modeling context as approaches seem to be data and application dependent. A weighted ensemble could leverage marketing data such that the effectiveness of advertisements can be improved. Costs for sending advertisements to unlikely buyers can be reduced and response rates improved due to more accurate targeting. Using the proposed robust ensemble approach, uplift modeling can be applied more broadly in marketing practice and the success and revenue of marketing campaigns can be increased.

7. References

[1] Athey, S., and G.W. Imbens, "Machine Learning Methods for Estimating Heterogeneous Causal Effects", 2015, pp. 26.

[2] Bose, I., and X. Chen, "Quantitative models for direct marketing: A review from systems perspective", *European Journal of Operational Research 195*(1), 2009, pp. 1–16.

[3] Boser, B.E., I.M. Guyon, and V.N. Vapnik, "A training algorithm for optimal margin classifiers", *Proceedings of the fifth annual workshop on Computational learning theory*, Association for Computing Machinery (1992), 144–152.

[4] Breiman, L., "Bagging Predictors", *Machine Learning 24*(2), 1996, pp. 123–140.

[5] Breiman, L., "Random Forests", *Machine Learning 45*(1), 2001, pp. 5–32.

[6] Coussement, K., P. Harrigan, and D.F. Benoit, "Improving direct mail targeting through customer response modeling", *Expert Systems with Applications 42*(22), 2015, pp. 8403–8412.

[7] Devriendt, F., D. Moldovan, and W. Verbeke, "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), 2018, pp. 13–41.

[8] Diemert, E., A. Betlei, C. Renaudin, and M.-R. Amini, "A Large Scale Benchmark for Uplift Modeling", 2018, pp. 6.

[9] Fernandez-Delgado, M., E. Cernadas, S. Barro, and D. Amorim, "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems?", *Journal of Machine Learning Research*, 2014.

[10] Goodfellow, I., Y. Bengio, and A. Courville, *Deep learning*, The MIT Press, Cambridge, Massachusetts, 2016.

[11] Grimmer, J., S. Messing, and S.J. Westwood, "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods", *Political Analysis 25*(4), 2017, pp. 413–434.

[12] Gubela, R., A. Bequé, S. Lessmann, and F. Gebert, "Conversion Uplift in E-Commerce: A Systematic Benchmark of Modeling Strategies", *International Journal of Information Technology & Decision Making 18*(03), 2019, pp. 747–791.

[13] Gutierrez, P., and J.-Y. Gérardy, "Causal Inference and Uplift Modelling: A Review of the Literature", *International Conference on Predictive Applications and APIs*, (2017), 1–13.

[14] Hansotia, B., and B. Rukstales, "Incremental value modeling", *Journal of Interactive Marketing; Philadelphia 16*(3), 2002, pp. 35–46.

[15] Huang, E.Y., and C. Tsui, "Assessing customer retention in B2C electronic commerce: an empirical study", *Journal of Marketing Analytics 4*(4), 2016, pp. 172–185.

[16] Jaskowski, M., and S. Jaroszewicz, "Uplift modeling for clinical trial data", *ICML 2012 Workshop on Clinical Data Analysis*, (2012).

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

[18] Nikulin, V., G.J. McLachlan, and S.K. Ng, "Ensemble Approach for the Classification of Imbalanced Data", *AI 2009: Advances in Artificial Intelligence*, Springer (2009), 291–300.

[19] Oechsle, F., and D. Schönleber, "Towards More Robust Uplift Modeling for Churn Prevention in the Presence of Negatively Correlated Estimation Errors", (2020).

[20] Radcliffe, N.J., and P.D. Surry, "Differential Response Analysis: Modeling True Response by Isolating the Effect of a Single Action", *Proceedings of Credit Scoring and Credit Control VI*, (1999).

[21] Radcliffe, N.J., and P.D. Surry, *Real-World Uplift Modelling with Significance-Based Uplift Trees*, 2011.

[22] Rzepakowski, P., and S. Jaroszewicz, "Decision Trees for Uplift Modeling", *2010 IEEE International Conference on Data Mining*, IEEE (2010), 441–450.

[23] Rzepakowski, P., and S. Jaroszewicz, Szymon, "Uplift Modeling in Direct Marketing", *Journal of Telecommunications and Information Technology nr 2*, 2012, pp. 43–50.

[24] Sagi, O., and L. Rokach, "Ensemble learning: A survey", *WIREs Data Mining and Knowledge Discovery 8*(4), 2018, pp. e1249.

[25] Shultz, T.R., S.E. Fahlman, S. Craw, et al., "Cost-Sensitive Learning", In C. Sammut and G.I. Webb, eds., *Encyclopedia of Machine Learning*. Springer US, Boston, MA, 2011, 231–235.

[26] Sołtys, M., S. Jaroszewicz, and P. Rzepakowski, "Ensemble methods for uplift modeling", *Data Mining and Knowledge Discovery 29*(6), 2015, pp. 1531–1559.

[27] Vafeiadis, T., K.I. Diamantaras, G. Sarigiannidis, and K.Ch. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction", *Simulation Modelling Practice and Theory 55*, 2015, pp. 1–9.

[28] Zaniewicz, Ł., and S. Jaroszewicz, "L p -Support vector machines for uplift modeling", *Knowledge and Information Systems 53*(1), 2017, pp. 269–296.