# Evaluation of Uplift Models with Non-Random Assignment Bias

Mina Rafla, Nicolas Voisine, Bruno Crémilleux

# Evaluation of Uplift Models with Non-Random Assignment Bias

Mina Rafla[1,2], Nicolas Voisine[1], and Bruno Crémilleux[2]

[1] Orange Labs, 22300 Lannion, France
{mina.rafla, nicolas.voisine}@orange.com
[2] UNICAEN, ENSICAEN, CNRS - UMR GREYC, Normandie Univ
14000 Caen, France bruno.cremilleux@unicaen.fr

**Abstract.** Uplift Modeling measures the impact of an action (marketing, medical treatment) on a person's behavior. This allows the selection of the subgroup of persons for which the effect of the action will be most noteworthy. Uplift estimation is based on groups of people who have received different treatments. These groups are assumed to be equivalent. However, in practice, we observe biases between these groups. We propose in this paper a protocol to evaluate and study the impact of the Non-Random Assignment bias (NRA) on the performance of the main uplift methods. Then we present a weighting method to reduce the effect of the NRA bias. Experimental results show that our bias reduction method significantly improves the performance of uplift models under NRA bias.

**Keywords:** Uplift Modeling · Machine Learning · Non-random Assignment Bias · Treatment Effect Estimation · Causal Inference

## 1 Introduction

Uplift modeling is a predictive modeling technique that models directly the incremental impact of treatment, such as a marketing campaign or a drug, on an individual's behavior. The applications are multiple: customer relationship management, personalized medicine, advertising, political elections. Uplift models help identify groups of people likely to respond positively to treatment *only because* they received one. A major difficulty in uplift modeling is that data are only partially known: it is impossible to know for an individual whether the chosen treatment is optimal because their responses to alternative treatments cannot be observed. Several works address challenges related to the uplift modeling with single treatment [8] and multiple treatments [24]. The evaluation of uplift models is studied in [18]. State-of-art uplift modeling approaches assume that the groups of individuals are homogeneous. This means that uplift should be modeled on experimental data, i.e., data whose generation is controlled and for which there is no bias between different treatment groups. However, in practice, uplift modeling is used with observational data where bias exists. For example, an unanswered commercial call introduces a bias between treated and not

treated individuals. Similarly, it is assumed that there is no bias between data used to learn an uplift model and its deployment whereas such a bias may exist. Those biases jeopardize the practical use of uplift modeling methods [15].

This paper aims to study the Non-Random Assignment (NRA) bias, a very common bias in the context of uplift modeling. It occurs when the treatment assignment is dependent on the characteristics of individuals. We address the following research questions: what is the impact of the NRA bias on the main uplift modeling approaches? How can the bias effect be reduced? To answer the first question, we design an experimental protocol that evaluates the impact of the NRA bias on state-of-art uplift methods. Our study allows us to identify several behavioral aspects of uplift methods. Regarding the second question, we propose a weighting method to reduce the effect of the NRA bias on the performance of uplift models. Experimental results show that our bias reduction method significantly improves the performance of uplift models under NRA bias. To the best of our knowledge, this is the first work that focuses on the bias effect in uplift modeling. The remainder of this paper is organized as follows. Section 2 introduces uplift modeling definition and methods, Section 3 describes the problem setting and our experimental protocol for evaluating the impact of NRA bias. We present our bias reduction method in Section 4 then conclude in Section 5.

## 2 Uplift modeling and evaluation

### 2.1 Definition

Uplift is a notion introduced by Radcliffe and Surry [17] and defined in Rubin's causal inference models [20] as the *Individual Treatment effect (ITE)*. We now outline the notion of uplift and its modeling.

Let $X$ be a group of $N$ individuals indexed by $i : 1 \dots N$ where each individual is described by a set of variables $\mathbb{X}$. $X_i$ denotes the set of values of $\mathbb{X}$ for the individual $i$. Let $T$ be a variable indicating whether or not an individual has received a treatment. Uplift modeling is based on two groups: the individuals having received a treatment (denoted $T = 1$) and those without treatment (denoted $T = 0$). Let $Y$ be the outcome variable (for instance, the purchase or not of a product). We note $Y_i(T = 1)$ the outcome of an individual $i$ when he received a treatment and $Y_i(T = 0)$ his outcome without treatment. The uplift of an individual $i$, denoted by $\tau_i$, is defined as: $\tau_i = Y_i(T = 1) - Y_i(T = 0)$.

In practice, we will never observe both $Y_i(T = 1)$ and $Y_i(T = 0)$ for a same individual and thus $\tau_i$ cannot be calculated. However, uplift can be empirically estimated by considering two groups: a treatment group (individual with a treatment) and a control group (without treatment). The estimated uplift of an individual $i$ denoted by $\hat{\tau}_i$ is then computed by using the CATE (Conditional Average Treatment Effect)[20]:

$$CATE : \hat{\tau}_i = \mathbb{E}[Y_i(T = 1)|X_i] - \mathbb{E}[Y_i(T = 0)|X_i] \quad (1)$$

As the real value of $\tau_i$ cannot be observed, it is impossible to directly use machine learning algorithms such as regression to infer a model to predict $\tau_i$. The next section describes how uplift is modeled in the literature.

### 2.2 Uplift modeling

The uplift modeling literature and a branch of the causal inference literature have recently approached each other [6]. We sketch below the main methods in this field of research.

***Meta-Learners*** Meta-Learners take advantage of usual machine learning algorithms to estimate the CATE. The most classical and intuitive approach is the T-Learner (also known as the **Two-Model approach** in the uplift literature, which is the name that we use in this paper). The T-Learner is made of two independent predictive models, one on the treatment group to estimate $P(Y|X, T = 1)$ and another on the control group to estimate $P(Y|X, T = 0)$. The estimated uplift of an individual $i$ is the difference between those values for the given individual, i.e. $\hat{\tau}_i = P(Y = 1|X_i, T = 1) - P(Y = 1|X_i, T = 0)$. The advantages of this approach are the simplicity and the possibility to use any machine learning algorithm but it has also known limitations [18]. The causal inference community defines other methods such as the S-Learner which includes the variable $T$ in the features with a standard regression, the X-Learner which performs a two-step regression before the estimation of the CATE to deal with the unbalanced size of treatment groups [7], the DR-Learner [9] which combines a two-model approach and the use of the Inverse Propensity Weighting [14].

***Class-Transformation Approach*** The principle of this approach [8] is to map the uplift modeling problem to a usual supervised learning problem. The outcome variable $Y$ is transformed into a variable $Z$ as illustrated in Eq. 2. Then a machine learning algorithm is used to learn a model and to predict $P(Z|X)$. The estimated uplift of an individual $i$ is $\hat{\tau}_i = 2 \times P(Z = 1|X_i) - 1$

$$Z = \begin{cases} 1, & \text{if } T = 1 \text{ and } Y = 1 \\ 1, & \text{if } T = 0 \text{ and } Y = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Several studies [3, 8] show that this approach has a better performance than the two-model approach.

***Direct-Approaches*** These methods modify supervised learning algorithms to suit them to fit the uplift modeling problem. Then uplift is directly estimated. Examples include methods based on decision trees [22, 24], k nearest neighbors [5], logistic regression [12] or reinforcement learning [11].

### 2.3 Uplift evaluation

Real values of uplift being not observed, supervised machine learning techniques cannot be used and therefore performance measures of the supervised setting are inoperative. That is why uplift is evaluated through the ranking of the individuals according to their estimated uplift value. The intuition is that a good uplift model estimates higher uplift values to individuals in the treatment group with positive outcomes than those with negative outcomes and vice versa for the control group. The qini measure (also known as Area Under Uplift Curve [2, 16]) is based on this principle to evaluate uplift methods. It is a variant of the Gini coefficient. Qini values are in $[-1, 1]$, the higher the value, the larger the impact of the predicted optimal treatment.

## 3 Evaluation of uplift with biased data

This section presents the NRA bias and the experimental protocol that we designed to assess performance of uplift methods under this bias.

### 3.1 Problem setting

State-of-art uplift methods assume that data are unbiased and that the treatment group comes from the same distribution as the control group, which is not true for real data. In practice, there are differences between treatment and control groups, also known as Non-Random Assignment bias, a prevalent type of bias in uplift modeling. Formally, this bias occurs when $P(T = 1|X) \neq P(T = 0|X)$ (which also means $P(X|T = 1) \neq P(X|T = 0)$). Usually it is easier to collect control data and the treatment group is the most biased because it is more challenging to apply a treatment to individuals and collect the corresponding data due to ethical, political or economic constraints.

This bias problem has been studied in the literature on clinical studies where the goal is to estimate the "Average Treatment Effect" (ATE) defined as $\mathbb{E}[Y_i(T = 1) - Y_i(T = 0)]$. In order to estimate it, the "Propensity Score Matching" (PSM) [21] is used to extract balanced treatment groups on which ATE is estimated. Similarly, in the uplift literature, since uplift methods assume the homogeneity between treatment groups, PSM is used to extract an unbiased sample from a biased dataset. Uplift modeling is applied subsequently as carried in [15]. However, this procedure clearly suffers from a loss of data.

### 3.2 Designing of the experimental protocol

This section describes the experimental protocol that we designed to evaluate the behavior of uplift methods under the NRA bias. The principle, to create a NRA bias and observe its impact, is to introduce imbalances in the data regarding the initial distribution of the variables. We do this by modifying proportions of individuals in a non-random way (for example, decreasing the proportion of

specific socio-professional categories or ages till it disappears in the data). Such a protocol must satisfy several conditions to correctly evaluate the impact of NRA in order to avoid introducing a bias due to the protocol itself. (1) The chosen variables to introduce bias have to be correlated with the outcome $Y$ or $Y$ given the treatment $T$, otherwise the bias will not affect the uplift modeling. (2) In contrast, the choice of the values of the variables, according to which the proportions of individuals vary, is random. If not, the construction of the populations $E1$ and $E2$ (which will be explained below) may be biased. (3) The bias must be tunable in order to change its rate and quantify its impact on the uplift methods. (4) The created bias is only in the treatment group in order to imitate the natural phenomena as previously explained in Section 3.1. (5) The total size of each of the biased learning samples is always the same in order to avoid any variation in the performance due to different learning data sizes.

More precisely, as shown in Fig. 1, two populations $E1$ and $E2$ are created. This is done by choosing a set of variables $V$ and dividing its values into two groups, $C1$ and $C2$, such that the number of individuals defined by the values of $C1$ is equivalent to the number of individuals defined by $C2$. Let $E1$ (resp. $E2$) be the population whose variables correspond to $C1$ (resp. $C2$) and whose sizes are $N1$ and $N2$ respectively. We use a 10-fold cross-validation. In the first training sample, $E1$ and $E2$ have an equal size (i.e. $N1 = N2$), it is considered unbiased and gives a reference value of the qini. The NRA bias is gradually introduced in the treatment group by increasing the size of $E1$ and decreasing the size of $E2$ while preserving the total size of the treatment group. We identify the bias rate of a sample by the variable $b$ where $b = (N1 - N2) \times 100/N$. $b$ goes from $b = 0$ in the unbiased situation to $b = 100$ the most biased situation according to the NRA bias. An uplift model is then learned on each biased sample defined by $b$. All models are then tested on the same test sample and evaluated using the qini. The evolution of the qini according to $b$ allows studying the behavior of an uplift method towards the NRA bias.

### 3.3 Experiments

We apply our protocol to several real and synthetic datasets using the main uplift approaches [3].

***Datasets*** We use four datasets from politics and marketing fields as well as four synthetic datasets (cf. Table 1). For all the datasets, the outcome is binary.

1. Criteo [3]: a usual marketing dataset for uplift modeling.
2. Hillstrom[4]: a classical dataset for uplift modeling. It is made up of two treatment groups and a control group. We only use the group of people who received an advertising campaign via mail for women's products as the treatment group.

[3] For a reproducible purpose, codes and experiment results are available in the supplementary material [19]

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

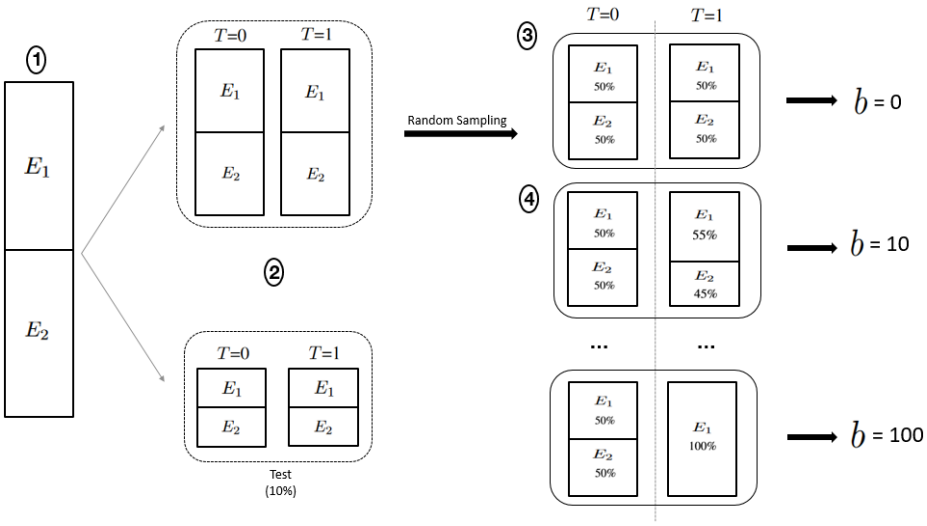


**Fig. 1.** Biased samples generation procedure: (1) Variable(s) $V$ is chosen to create E1 and E2. (2) Creating training and test sets with 10-fold cross validation. (3) Random sampling of treatment and control groups. (4) The sizes of the treatment and control groups are always the same throughout the biasing process.

3. Gerber [4]: a policy-relevant dataset used to study the effect of social pressure on voter turnout.
4. Retail Hero[5]: a dataset of the X5 sales group, the treatment is the action to send SMS to encourage consumers to increase their purchases.
5. Megafon[6]: a synthetic dataset created for uplift modeling. It is generated by telecom companies in order to reproduce the situations encountered by these companies.
6. Zenodo[7]: a synthetic dataset containing trigonometric patterns specifically designed for uplift modeling. We used a subset of 20,000 rows of data (data identified by the variable trial_id = 1 and trial_id = 2).
7. Synth1 and Synth2: two synthetic datasets that we have built as a 2D grid of size 10x10 in which each cell corresponds to a particular uplift drawn at random. Synth1 is a dataset with a high ATE value and Synth2 has a low response rate.

***Uplift methods*** We test 13 uplift methods: two-model approach (2M); class-transformation approach (CT), each with Xgboost and logistic regression (LR); DR-Learner (DR); X-Learner and S-Learner, each with Xgboost and linear regression (LinR). Direct-approaches based on decision trees are tested as well: KL, ED [22] and CTS [24].

[5] `https://ods.ai/competitions/x5-retailhero-uplift-modeling/data`
[6] `https://ods.ai/tracks/df21-megafon/competitions/megafon-df21-comp/data`
[7] `https://zenodo.org/record/3653141\#.YUCYEufgoW8`

**Table 1.** Dataset characteristics.
- Datasets have a balanced size of treatment and control groups.
- Independence between treatment and control groups is measured using the C2ST test [13]. A p-value smaller than 0.05 means the null hypothesis is rejected (i.e. treatment independence).
- *Value after re-balancing the dataset using PSM [21]

| Datasets | #Rows | #Variables | Response Ratio | ATE | Treatment Independence |
|---|---|---|---|---|---|
| Criteo | 50000 | 13 | 0.16 | 0.08 | 0.1 |
| Hillstrom | 42693 | 8 | 0.129 | 0.04 | 0.33 |
| Gerber | 76419 | 10 | 0.34 | 0.06 | 0.43 |
| RetailHero | 200039 | 11 | 0.619 | 0.033 | 0.7 |
| Megafon | 600000 | 36 | 0.2 | 0.04 | 0.4375* |
| Synthetic Zenodo | 20000 | 16 | 0.3 | 0.109 | 0.22 |
| Synth1 | 40000 | 2 | 0.32 | 0.241 | 0.197 |
| Synth2 | 40000 | 2 | 0.007 | 0.00125 | 0.33 |

***Implementation details*** For each dataset (except Synth1 and Synth2) and for each uplift method, the experimental protocol is applied twice with different contents of $V$: once with the variable the most correlated with $Y$ and once with the variable the most correlated with $Y$ given the treatment group ($T = 1$). For Synth1 and Synth2, $V$ contains the two variables of these datasets. Moreover, given a set $V$, the experiment is repeated twice in order to provide different splittings of $C1$ and $C2$.

### 3.4 Results

***Qini variability according to*** $b$ Fig. 2 illustrates the results (due to space constraints, it is not possible to give all the results). We observe that the NRA bias strongly affects the performance of uplift models[8] (the higher the bias rate, the more significant the decrease of the qini). To provide a global view of the results, we compute for each dataset and each uplift method the *Average Qini*, i.e., the average of qini values according to the bias rates going from $b = 0$ to $b = 100$ (cf. Table 2).

***Overall ranking*** To better compare the methods according to their resistance to NRA bias, Fig. 3 shows the average rank obtained by each method based on the Average Qini (all divisions of $V$ are taken into account).

The results of these experiments provide the following messages: (i) the most resistant models to the NRA bias are the ED and X-Learner_LinR, DR_LinR, two-model approach with the logistic regression: the qini strongly decays only when the bias rate is high; (ii) the models where the qini gently degrades as the bias rate increases are tree based methods (KL, and CTS) and (iii) the models

[8] When comparison with state of the art is possible, the achieved qini values without bias ($b = 0$) are those usually found in the literature [3].

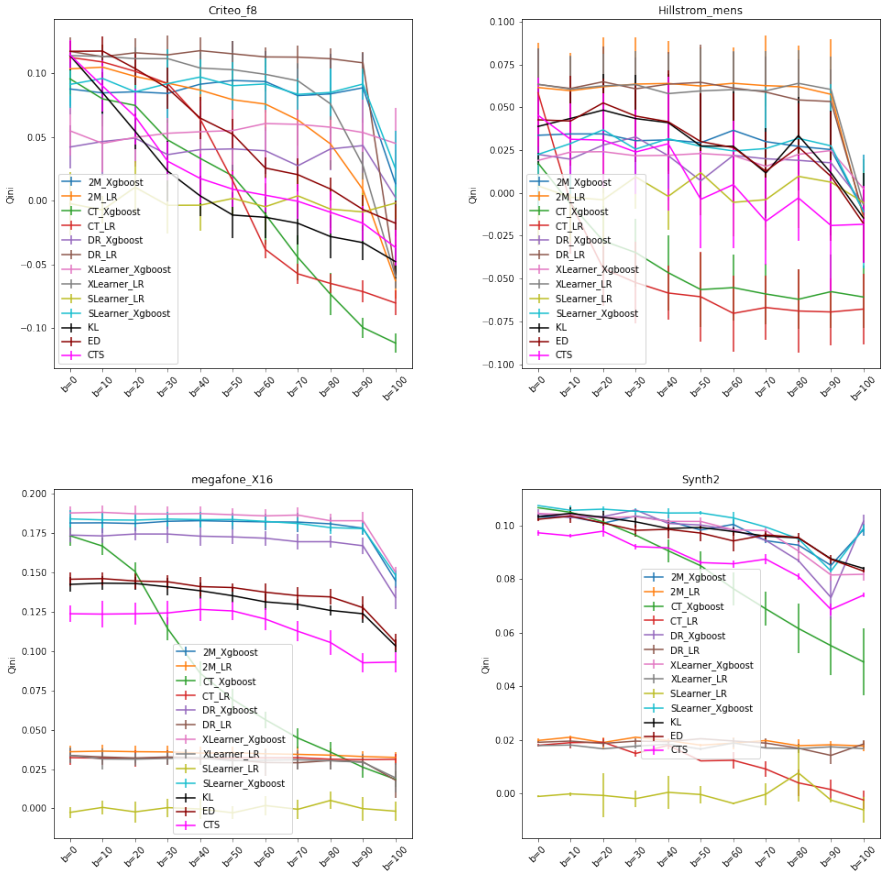


**Fig. 2.** Qini values of uplift methods according to NRA bias rates in the Criteo dataset with the 'f8' variable (top left), Hillstrom dataset with the 'mens' variable (top right), Megafon dataset with the 'X16' variable (bottom left) and Synth2 dataset with its both variables (bottom right). A method name is followed by the learning algorithm used with it.

strongly affected by the bias even with low bias rates are the class-transformation based methods and the S-Learner_LinR.

***Methods comparison with statistical tests*** We study now the significance of the results regarding the comparison of the uplift methods (cf. Table 2) by using a statistical test. Following the study [1], we choose the Friedman test with the post hoc test of Nemenyi to compare the performance (average qini) of more than two methods across several datasets. Fig. 4 depicts the results with a heatmap. The null hypothesis states that there is no significant difference

**Table 2.** Average Qini (multiplied by 100) and its variance (shown in brackets) across datasets and uplift methods (in bold, the best value for each dataset). A dataset name is followed by the names of the $V$ variables used to generate the NRA bias (due to space constraints, the results are given for a single splitting of the $V$ values).

| | TwoModel | | ClassTransformation | | DR | | XLearner | | SLearner | | Trees | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Xgboost | LR | Xgboost | LR | Xgboost | LinR | Xgboost | LinR | Xgboost | LinR | KL | ED | CTS |
| Criteo_f2 | 6.6(1.7) | 7.2(1.6) | 0.2(1.9) | 1.9(1.2) | 4.4(2.8) | **9.9(0.9)** | 5.5(2.6) | 8.5(0.8) | 8.0(1.9) | -0.2(1.9) | 0.6(1.4) | 4.9(1.3) | 2.1(1.5) |
| Criteo_f8 | 8.1(2.6) | 6.3(2.0) | 0.1(1.7) | 1.7(1.0) | 3.7(2.3) | **9.8(1.0)** | 5.4(2.6) | 8.1(1.1) | 8.4(1.9) | -0.2(1.7) | 1.2(1.6) | 5.2(1.2) | 2.4(1.6) |
| Gerber_p2002 | -2.4(2.0) | 1.1(1.1) | -2.1(1.5) | -0.4(1.2) | -2.0(1.9) | 0.8(1.1) | -2.3(1.9) | **1.4(1.1)** | -2.0(2.0) | 0.1(0.9) | -1.5(1.8) | -0.9(1.5) | -0.1(1.7) |
| Gerber_p2004 | -2.1(2.0) | 0.8(1.1) | -1.8(1.7) | -1.2(1.3) | -2.1(1.9) | 0.7(1.1) | -2.1(1.8) | **1.2(1.3)** | -1.8(2.0) | 0.0(1.1) | -1.7(1.8) | -1.5(1.9) | -0.6(1.9) |
| Hillstrom_mens | 2.7(2.1) | **5.5(2.6)** | -4.1(2.0) | -4.6(2.2) | 1.9(2.4) | 5.4(2.1) | 2.0(2.6) | **5.5(2.2)** | 2.5(2.7) | 0.2(2.4) | 2.8(2.6) | 2.9(2.5) | 1.0(2.8) |
| Hillstrom_newbie | 2.8(2.2) | **6.2(2.7)** | 0.1(2.1) | 2.4(1.9) | 1.0(2.4) | 5.9(2.0) | 2.1(2.3) | 6.0(2.0) | 3.3(2.2) | -0.1(2.4) | 4.2(2.2) | 4.3(2.5) | 4.3(2.5) |
| Megafone_X16 | 17.8(0.5) | 3.5(0.4) | 8.6(0.6) | 3.2(0.4) | 16.9(0.5) | 3.0(0.5) | **18.3(0.4)** | 3.0(0.6) | 17.9(0.4) | -0.0(0.6) | 13.2(0.5) | 13.7(0.5) | 11.6(0.7) |
| Megafone_X21 | 18.2(0.4) | 3.5(0.4) | 12.0(0.4) | 2.4(0.5) | 17.4(0.5) | 3.0(0.4) | **18.8(0.4)** | 3.1(0.4) | 18.4(0.4) | -0.0(0.6) | 13.9(0.5) | 14.0(0.6) | 10.7(0.8) |
| Synth1 | 7.0(0.9) | 0.9(1.6) | 1.7(0.9) | -2.9(1.3) | 9.7(1.5) | -0.4(1.5) | **12.6(1.6)** | -1.6(2.0) | 12.2(1.2) | 0.6(1.6) | 9.7(1.2) | 8.8(1.6) | 8.7(1.2) |
| Synth2 | 9.8(0.1) | 1.9(0.1) | 8.1(0.5) | 1.1(0.2) | 9.7(0.2) | 1.9(0.1) | 9.7(0.2) | 1.8(0.1) | **10.1(0.1)** | -0.1(0.4) | 9.7(0.1) | 9.6(0.2) | 8.7(0.1) |
| retailHero_age | 0.7(0.4) | 1.2(0.3) | 0.3(0.4) | 0.8(0.4) | 0.5(0.4) | **1.3(0.4)** | 0.5(0.3) | 1.2(0.3) | 0.9(0.3) | -0.0(0.3) | 0.8(0.3) | 0.9(0.3) | 0.9(0.4) |
| retailHero_trNum | 0.8(0.4) | 1.2(0.3) | 0.4(0.3) | 1.1(0.4) | 0.4(0.4) | **1.3(0.4)** | 0.5(0.4) | 1.2(0.4) | 0.9(0.4) | -0.0(0.4) | 0.7(0.4) | 0.7(0.4) | 0.6(0.4) |
| zenodoSynth_X10 | 9.7(1.8) | 12.6(1.9) | 7.0(2.2) | 12.1(1.5) | 7.8(1.9) | 12.2(1.9) | 9.4(1.7) | 12.1(1.7) | 11.5(2.0) | 0.0(2.5) | 12.8(1.9) | **13.0(1.9)** | 10.6(2.6) |
| zenodoSynth_X31 | 9.8(2.4) | 12.2(2.0) | 6.6(2.0) | 12.0(1.9) | 7.7(2.1) | 12.3(1.9) | 9.7(2.2) | 12.4(1.7) | 11.7(2.2) | 0.1(1.9) | 12.7(1.9) | **13.2(2.0)** | 10.2(2.2) |

in performance according to the average qini between two methods across the datasets. With a value of $p$ (p-value) smaller than 0.05, the null hypothesis is rejected (in green in Fig. 4). Fig. 4 and Fig. 3 confirm that the S-Learner and the class-transformation based approaches are the least resistant towards the NRA bias.

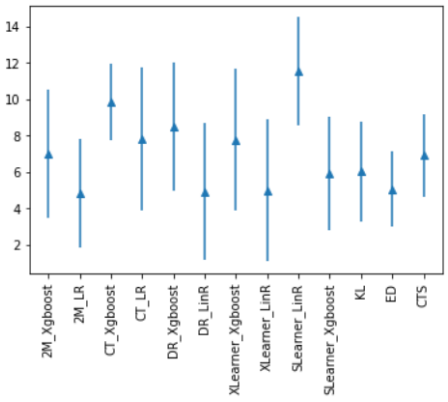


**Fig. 3.** Overall ranking for the different uplift approaches.

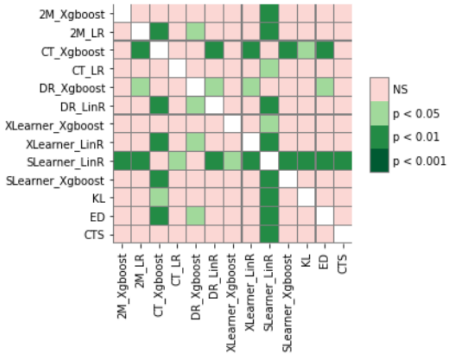


**Fig. 4.** Heat map to visualize the comparison between uplift methods. A value of $p$ smaller than 0.05 means that the null hypothesis is rejected.

## 4 Method to reduce the NRA bias impact

This section presents our weighthing method to reduce the effect on the NRA bias on the uplift modeling. Our method is inspired from the *Domain Adaptation* literature where samples of a source dataset are weighted according to their

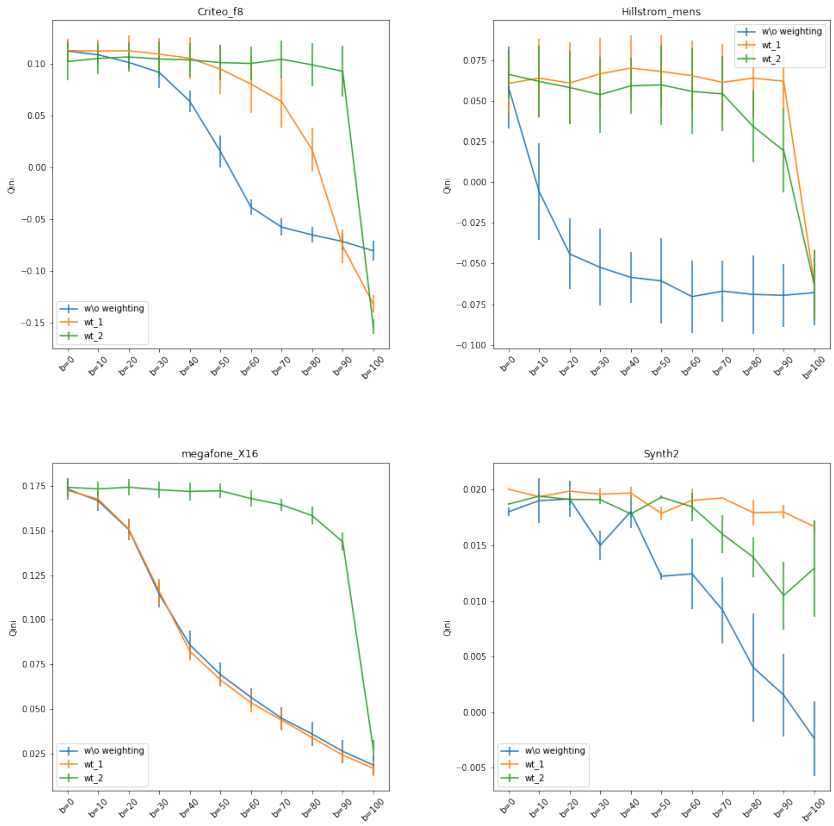


**Fig. 5.** Qini values by class-transformation based methods according to different NRA bias rates with and without reweighting. Top-left: class-transformation approach with Xgboost on Criteo dataset and 'f8' variable. Top-right: class-transformation approach with logistic regression on Hillstrom dataset and 'mens' variable. Bottom-left: class-transformation approach with Xgboost on Megafon dataset with X16 variable. Bottom-right: class-transformation approach with logistic regression on Synth2 dataset with its both variables.

importance to a target dataset [10]. The principle of our method is to weight individuals of the treatment group according to their weight in the control group to make the biased population (the treatment group) similar to the unbiased one (the control group). Our weighting technique is based on the propensity score which is the probability for an individual of being treated ($T = 1$) given his vector of observed variables $X_i$ i.e. $P(T = 1|X_i)$. In observational studies, the propensity scores are not known but they can be learned from the data using a regression algorithm. Our method weights each individual $i$ of the treatment

**Table 3.** Average Qini (multiplied by 100) and its variance (shown in brackets) with the class-transformation based methods (in bold, the best value for each dataset). Dataset name is followed by the names of the $V$ variables used to generate the NRA bias (for space constraints, the results are given for a single splitting of the $V$ values). MAE takes into account all splittings of $V$ into $C1$ and $C2$, as explained previously.

| | Class-Transformation with LR | | | | Class-Transformation with Xgboost | | | |
|---|---|---|---|---|---|---|---|---|
| | Ref. qini | w/o weights | wt_1 | wt_2 | Ref. qini | w/o weights | wt_1 | wt_2 |
| Criteo_f2 | 11.1(0.9) | 1.9(1.2) | 6.1(1.5) | **8.2(2.0)** | 9.1(2.6) | 0.2(1.9) | 2.6(1.8) | **4.8(1.9)** |
| Criteo_f8 | 11.2(1.0) | 1.7(1.0) | 5.5(1.8) | **7.9(1.7)** | 9.6(1.2) | 0.1(1.7) | 3.4(1.5) | **5.0(1.8)** |
| Gerber_p2002 | 0.8(1.6) | -0.4(1.2) | **0.9(1.1)** | 0.5(1.2) | -1.9(2.0) | -2.1(1.5) | **-1.6(1.9)** | -2.3(1.8) |
| Gerber_p2004 | 1.1(1.4) | -1.2(1.3) | **0.9(1.3)** | 0.4(1.1) | -1.6(2.1) | -1.8(1.7) | **-1.7(1.8)** | -2.3(1.9) |
| Hillstrom_mens | 5.9(2.5) | -4.6(2.2) | **5.3(2.2)** | 4.2(2.2) | 1.7(2.1) | -4.1(2.0) | -0.2(2.7) | **0.5(2.4)** |
| Hillstrom_newbie | 6.3(1.7) | 2.4(1.9) | **5.6(2.0)** | 5.2(2.1) | 1.7(1.9) | 0.1(2.1) | 1.3(2.0) | **1.4(2.1)** |
| Megafone_X16 | 3.2(0.5) | **3.2(0.4)** | 3.1(0.4) | **3.2(0.4)** | 17.3(0.6) | 8.6(0.6) | 8.4(0.5) | **15.5(0.5)** |
| Megafone_X21 | 3.2(0.4) | 2.4(0.5) | **3.1(0.4)** | 3.0(0.5) | 17.2(0.5) | 12.0(0.4) | 12.0(0.4) | **16.0(0.5)** |
| Synth1 | -0.2(3.4) | -2.9(1.3) | -1.0(1.8) | **-0.8(0.9)** | 2.5(2.4) | 1.7(0.9) | 2.5(0.7) | **8.9(2.9)** |
| Synth2 | 1.8(0.0) | 1.1(0.2) | **1.9(0.1)** | 1.7(0.1) | 10.7(0.0) | 8.1(0.5) | 8.3(0.5) | **9.7(0.2)** |
| retailHero_age | 1.2(0.4) | 0.8(0.4) | **1.3(0.4)** | 1.2(0.3) | 0.6(0.4) | 0.3(0.4) | 0.3(0.4) | **0.6(0.4)** |
| retailHero_trNum | 1.2(0.3) | 1.1(0.4) | **1.2(0.4)** | **1.2(0.4)** | 0.7(0.4) | 0.4(0.3) | 0.4(0.3) | **0.6(0.3)** |
| zenodoSynth_X10 | 12.3(1.3) | **12.1(1.5)** | 11.9(1.7) | 9.8(1.8) | 8.0(3.1) | 7.0(2.2) | **7.4(2.0)** | 6.5(2.1) |
| zenodoSynth_X31 | 11.7(2.3) | 12.0(1.9) | **12.1(1.7)** | 9.9(2.0) | 6.9(1.9) | 6.6(2.0) | **7.2(2.5)** | 6.5(2.2) |
| **MAE** | 0 | 2.367 | 0.978 | 1.053 | 0 | 2.803 | 1.953 | 1.592 |

group by $w(X_i)$ s.t.:

$$w(X_i) = P(T = 0|X_i)/P(T = 1|X_i) \tag{3}$$

We estimate the probabilities of Eq. 3 by using logistic regression and xgboost. Then the uplift method integrates the weights to amplify the role of the under-represented individuals in the treatment group and estimate $\hat{\tau}_i$. We named wt_1 (resp. wt_2) the use of the logistic regression (resp. xgboost) in the weighting method.

We evaluate our weighting method with the two-model and the class-transformation approaches since these approaches use traditional machine learning algorithms where weights can be given directly at each line (individual). The direct-approaches cannot take into account weights, so we do not use them. Results show a large enhancement in the performance with the class-transformation methods (cf. Fig. 5) and a slight improvement with the two-model approach (the full set of results can be found in the supplementary material [19]). Table 3 details the results with the class-transformation based methods. "Ref. qini" denotes the *reference qini*, that is the qini value of a method without bias (i.e. $b = 0$) and without weighting. The Mean Absolute Error (where MAE $= \frac{1}{n}\sum_{j=1}^{n} |Ref.qini_j - AverageQini_j|$) indicates the gap between the qini obtained by an uplift method and the reference qini. The smaller the gap is, the better the weighting. The gap is much smaller with our weighting methods especially with the logistic regression (LR) than without weighting. Best average qini values are also achieved with weighting except on zenodoSynth_X10.

***Statistical Test*** Following the study [1], we use Wilcoxon test [23] to determine if our weighting method significantly improves the performance of the uplift

**Table 4.** p-values obtained with the Wilcoxon test when comparing uplift methods w/o and with weighting.

| Methods | p-value | Methods | p-value |
|---|---|---|---|
| CT_LR w/o weights vs CT_LR with wt_1 | **0.0014** | 2M_LR w/o weights vs 2M_LR with wt_1 | 0.985 |
| CT_LR w/o weights vs CT_LR with wt_2 | 0.106 | 2M_LR w/o weights vs 2M_LR with wt_2 | 0.986 |
| CT_Xgboost w/o weights vs CT_Xgboost with wt_1 | 0.142 | 2M_Xgboost w/o weights vs 2M_Xgboost with wt_1 | 0.356 |
| CT_Xgboost w/o weights vs CT_Xgboost with wt_2 | **0.02** | 2M_Xgboost w/o weights vs 2M_Xgboost with wt_2 | 0.68 |

methods. This test is used to compare two methods on several datasets. As we perform two tests (on wt_1 and wt_2 methods), in order to control the familly-wise error rate due to multiple tests, the Bonferroni correction is applied and therefore the null hypothesis is rejected when the p-value is smaller than 0.025. Table 4 asserts that our weighting technique improves significantly the class-transformation based methods while there is no significant improvement with the two-model based methods.

***Discussion*** The weak impact of the weighting method on the two-model approach methods can be explained. The NRA bias does not change in the treatment group the distribution of the outcome $Y$ given populations $E_1$ and $E_2$ (cf. Section 3.2). The probability estimations $P(Y|T = 1, X)$ and $P(Y|T = 0, X)$ are then slightly affected, and the performances with and without weighting are similar. This is different with the class-transformation methods which directly estimate $Z$ based on the assumption that the treatment and control groups are equivalent. However, this assumption no longer holds with the NRA bias. Then weighting the treatment group improves the estimation of $Z$ and thus the uplift.

## 5 Conclusion

In this paper, we have studied the effect of the NRA bias when modeling uplift methods. To the best of our knowledge, this is the first work that focuses on the study of bias effect on current uplift models. We have designed an experimental protocol that allows, by varying the bias rate, to study the impact of the NRA bias on uplift methods and to identify classes of behavior for these methods. Inspired by the literature on domain adaptation, we have proposed a method to reduce the effect of the NRA bias by weighting the individuals in the treatment group. Experimental results on eight datasets show that our method significantly improves the uplift estimation performances for the class-transformation based methods.

This work opens several perspectives. As the weighting method reduces the effect of the NRA bias with the class transformation methods, it seems promising to design new methods of this family. On the other hand, it will be fruitful to study other types of bias, such as (i) the deployment bias, which occurs when uplift models are applied to different populations (Covariate Shift situation) or when the behavior of individuals changes with time (Concept Drift situation); (ii) the non-response bias which is a real challenge for uplift modeling with observational data.

## References


1. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
2. Devriendt, F., Guns, T., Verbeke, W.: Learning to rank for uplift modeling (2020)
3. Diemert, E.: A large scale benchmark for uplift modeling (2018)
4. Gerber, A.S., Green, D.P., Larimer, C.W.: Social pressure and voter turnout: Evidence from a large-scale field experiment. American Political Science Review (2008)
5. Guelman, L.A.: Optimal personalized treatment learning models with insurance applications (2015)
6. Gutierrez, P., Gérardy, J.Y.: Causal inference and uplift modelling: A review of the literature. In: PAPIs (2016)
7. Jacob, D.: Cate meets ml – the conditional average treatment effect and machine learning (2021)
8. Jaskowski, M., Jaroszewicz, S.: Uplift modeling for clinical trial data (2012)
9. Kennedy, E.H.: Optimal doubly robust estimation of heterogeneous causal effects (2020)
10. Kouw, W.M., Loog, M.: An introduction to domain adaptation and transfer learning (2019)
11. Li, C., Yan, X., Deng, X., Qi, Y., Chu, W., Song, L., Qiao, J., He, J., Xiong, J.: Reinforcement learning for uplift modeling (2019)
12. Lo, V., Pachamanova: From predictive uplift modeling to prescriptive uplift analytics: A practical approach to treatment optimization while accounting for estimation risk. Journal of Marketing Analytics (2015)
13. Lopez-Paz, D., Oquab, M.: Revisiting classifier two-sample tests (2018)
14. Lunceford, J.K., Davidian, M.: Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine **23 19**, 2937–60 (2004)
15. Olaya, D., Coussement, K., Verbeke, W.: A survey and benchmarking study of multitreatment uplift modeling. Data Min. Knowl. Discov. **34**(2), 273–308 (2020)
16. Radcliffe, N.: Using control groups to target on predicted lift: Building and assessing uplift model. Direct Marketing Analytics Journal pp. 14–21 (2007)
17. Radcliffe, N.J., Surry, P.D.: Differential response analysis: Modeling true responses by isolating the effect of a single action (1999)
18. Radcliffe, N.J., Surry, P.D.: Real-world uplift modelling with significance-based uplift trees (2012)
19. Rafla, M., Voisine, N., Crémilleux, B.: Supplementary material, `https://github.com/MinaWagdi/UpliftEvaluation_NRA/`
20. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **66**, 688–701 (1974)
21. Rubin, D.B.: Using propensity scores to help design observational studies: Application to the tobacco litigation. Health Services and Outcomes Research Methodology **2**(3), 169–188 (Dec 2001)
22. Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling with single and multiple treatments. Knowledge and Information Systems **32**, 303–327 (2011)
23. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945), `http://www.jstor.org/stable/3001968`
24. 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. pp. 588–596. SIAM (2017b)