Uplift Modeling with High Class Imbalance

Otto Nyberg OTTO.NYBERG@HELSINKI.FI
Tomasz Kuśmierczyk TOMASZ.KUSMIERCZYK@GMAIL.COM
Arto Klami ARTO.KLAMI@HELSINKI.FI
Department of Computer Science, University of Helsinki, Helsinki, Finland

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Uplift modeling refers to estimating the causal effect of a treatment on an individual observation, used for instance to identify customers worth targeting with a discount in e-commerce. We introduce a simple yet effective undersampling strategy for dealing with the prevalent problem of high class imbalance (low conversion rate) in such applications. Our strategy is agnostic to the base learners and produces a 6.5% improvement over the best published benchmark for the largest public uplift data which incidentally exhibits high class imbalance. We also introduce a new metric on calibration for uplift modeling and present a strategy to improve the calibration of the proposed method.

Keywords: uplift; heterogeneous treatment effect; class imbalance; undersampling

1. Introduction

Uplift modeling, also known as individual treatment effect (Rubin, D. B, 1974) and heterogeneous treatment effect (Athey and Imbens, 2015), is the art of modeling the causal effect of some treatment on the outcome for individual observations. This is a common task in e-commerce where various actions (ads, discounts, chat service etc.) are used to increase conversion rates (Diemert et al., 2018; Guelman et al., 2014). The task is to identify users for which the treatment effect is sufficiently large to merit treatment. This requires well-calibrated estimates of both the treatment effect and the conversion rate. The same task occurs also in medicine where a treatment should only be given to individuals that benefit from it (Jaskowski and Jaroszewicz, 2012), and in education where identifying students that benefit from intervention can help prevent dropping out (Olaya et al., 2020b).

Several practical methods for estimating uplift have been proposed including double classifiers (Radcliffe and Surry, 1999), class-variable transformations (Jaskowski and Jaroszewicz, 2012), revert-labeling (Athey and Imbens, 2015), tree-based methods (Rzepakowski and Jaroszewicz, 2010), causal forests (Athey et al., 2019), and support-vector machines (Zaniewicz and Jaroszewicz, 2013). For more exhaustive lists and reviews, see Gubela et al. (2020); Guelman et al. (2014); Gutierrez and Gérardy (2017). Despite a rich range of methods, one frequently occurring challenge is ignored by the existing work: Especially in e-commerce the problem is highly imbalanced in the sense that the rate of positive instances (e.g. conversion) is extremely low compared to the rate of negative instances (e.g. user leaving without buying) regardless of whether or not a treatment was applied. For example, in the Criteo data set (Diemert et al., 2018) the overall rate of positive instances is only around 0.23%,

with 0.24% rate for treated observations and 0.17% for untreated ones. This challenge resembles that of class imbalance in classification (see Haixiang et al. (2017) for a review) for which it is well established that explicitly accounting for the class imbalance is critical (Wallace and Dahabreh, 2014), but for uplift modeling the problem has remained unaddressed. This also explains why the Criteo dataset has remained largely neglected despite being the largest publicly available uplift dataset by two orders of magnitude.

We fill this void. We show that for highly imbalanced problems a specific form of undersampling leads to an equivalent learning problem that is easier to solve, although at the expense of calibration. We formulate an uplift modeling approach combining (a) undersampling carried out as preprocessing with (b) postprocessing calibration extended from isotonic regression (Zadrozny and Elkan, 2002). Both steps can be combined with multiple uplift modeling algorithms. We demonstrate the approach on the Criteo data, achieving a 6.5% increase in the *area under the uplift curve (AUUC) metric* (Jaskowski and Jaroszewicz, 2012) over the method (double-classifier with logistic regression) previously identified as the best performing model for this data (Diemert et al., 2018; Semenova and Temirkaeva, 2019).

We make five concrete contributions: (a) We analyze class imbalances in uplift modeling identifying it as a problem neglected in current research. (b) We present a technique based on undersampling the majority class for addressing class imbalance compatible with several methods for estimating uplift. (c) We present a calibration method for correcting a bias caused by undersampling. (d) We propose two new metrics for measuring calibration of uplift estimates.(e) We demonstrate clear improvements in uplift metrics on the Criteo data (Diemert et al., 2018) which has high class imbalance.

2. Background and Related Work

2.1. The Uplift Learning Problem

The uplift problem is general, but for clarity of presentation we map the terminology to its e-commerce application. An *observation* is an individual user, the *features* x represent information about the user (e.g. demographics, browsing history), a *treatment* $t \in \{0, 1\}$ indicates whether the user was treated with some marketing action (with *control* referring to the group of users that was not), and an *outcome* $y \in \{0, 1\}$ (i.e. class label) indicates whether the user converted (e.g. made a purchase).

The uplift problem is then the task of estimating

$$\tau(x) = p(y|x, do(t = 1)) - p(y|x, do(t = 0)) \tag{1}$$

where $\tau(x)$ is uplift for an observation with features x, and $do(t = .)$ is the do-operator introduced by Pearl (2009). That is, uplift is the difference between the probability of a positive outcome y when the observation is treated ($do(t = 1)$) and the corresponding probability when it is not ($do(t = 0)$). The task is characterized by the *fundamental problem of causal inference* (Holland, 1986); for a given observation we can know the value of y either when the treatment is applied or when it is not, but never both.

The problem can be solved by learning a model for $\tau(x)$ on a static data collection of triplets $\{x, y, t\}$ where treatments t have been applied for some subset of observations

using a pre-determined policy. We will assume that this policy is random. This choice allows estimating $\tau(x)$ since the groups of observations that received the treatment are independent of the control observations. This also holds for the data used in our experiments.

Uplift models are typically evaluated using uplift curves (see Figure 2 (right) for an example) and the area under the uplift curve (AUUC) (Jaskowski and Jaroszewicz, 2012) which provides an overall measure of goodness.

2.2. Methods for Modeling Uplift

As mentioned in Section 1, the uplift problem can be solved with a wide range of algorithms. We will here introduce two prominent methods used in our experiments, but note that the undersampling approach is compatible with all other methods as well. The two methods considered here are the double-classifier approach (Radcliffe and Surry, 1999) and the class-variable transformation (Lai, 2006; Jaskowski and Jaroszewicz, 2012). The double-classifier (DC) approach builds directly on Eq. (1). Since $\tau(x)$ is the difference between the two probabilities $p(y|x, do(t = 1))$ and $p(y|x, do(t = 0))$, we can separately train two classifiers, one for each of these problems. By using classifiers capable of providing probability estimates $\hat{p}(y|x, t)$ (e.g. logistic regression), we get an estimate (here called *score*) for uplift as the difference between these two quantities as

$$s^{\mathrm{DC}}(x) = \hat{p}(y|x, t = 1) - \hat{p}(y|x, t = 0).$$

Class-variable transformation (CVT) (Jaskowski and Jaroszewicz, 2012) methods define a new binary variable z so that $z = 1$ if either $t = 1$ and $y = 1$, or $t = 0$ and $y = 0$. Otherwise $z = 0$. This relabeling results in $p(z|x)$ being a weighted average with $p(z = 1|x) = p(t = 1)p(y|x, t = 1) + p(t = 0)p(y = 0|x, t = 0)$. Assuming balanced treatment and control groups ($p(t = 1) = p(t = 0) = 1/2$), this results in

$$\tau(x) = 2p(z|x) - 1. \tag{2}$$

This only requires estimating one probability $p(z|x)$ instead of estimating a difference between two quantities, and by plugging in an estimate for the class probability we get the score $s^{\mathrm{CVT}}(x) = 2\hat{p}(z = 1|x) - 1$. To satisfy $p(t = 0) = p(t = 1)$ we used undersampling. Note that this undersampling that drops excess observations in the larger group is carried out solely for validity of Eq. (2) and is not to be confused with the proposed undersampling scheme for addressing class imbalance.

2.3. Imbalances in Uplift Modeling

Imbalance has been used as a term in uplift modeling in multiple contexts. Olaya et al. (2020a) used it to describe differences in treatment and control distributions arising from a non-random treatment policy. This is more commonly referred to as confounding effects (Austin, 2011). Betlei et al. (2018), in turn, analyzed the problem arising from *imbalance* in the sample size of treatment and control groups.

Our paper deals with *class imbalance* characterized by an imbalance in the number of positive and negative observations which has so far been neglected in the literature. While Gubela et al. (2020) also dealt with *class imbalance* in uplift modeling, they focused

exclusively on regression and their method does not extend to binary outcomes that is the focus of this paper. For sufficiently large and balanced data, the classifier-based methods presented in section 2.2 work well. With severe class imbalance, however, the underlying classification problems are difficult (Bose and Chen, 2009; Weiss, 2004) and the probability estimates are poorly calibrated (Weiss and Provost, 2003).

The Criteo data has extreme class imbalance. The conversion rate in the control group is 0.174% and in the treatment group 0.240%. For DC we need to estimate class probabilities that are extremely small, and avoiding naive solutions that favor the negative class with 99.8% prior probability is hard. Class imbalance also causes problems with CVT (Eq. (2)), but the problem is not immediately apparent as the better treatment-balance masks it. The classification problem for z follows the 85:15 split in treatments and seems reasonably balanced, but the conditional distributions of the outcome y given the transformed variable z are extremely imbalanced with 99.96% of the observations with $z = 0$ corresponding to negative treatment observations ($t = 1, y = 0$) and 98.7% of the observations with $z = 1$ corresponding to negative control observations ($t = 0, y = 0$). In practice, it is more likely for high accuracy to be achieved by learning a difference between the features of treatment and control groups – either because the original treatment policy is not truly random, or simply because of overlearning – rather than actually modeling the uplift.

The above example refers specifically to the Criteo data, but the low rate of positive outcomes is prevalent in all e-commerce. Typical studies report rates in the order of $0.1-5\%$, depending on whether the outcome corresponds to views, visits, or purchases (Diemert et al., 2018; Richardson et al., 2007), but no practical methods for addressing the imbalance exists. Notably the most interesting cases - the purchases - fall in the lower end of this spectrum.

3. Class Imbalanced Uplift Modeling

Next we address the class imbalance problem in uplift modeling. We start by presenting how undersampling of the negative observations can be used in conjunction with both DC and CVT, explain a practical method for selecting the undersampling factor, present a novel calibration method that corrects for miss-calibration caused by the undersampling strategy, and finally present novel metrics for evaluating the calibration.

3.1. Undersampling

There is a wide range of approaches for coping with high class imbalance in classification (see Kaur et al. (2019)). One of these, namely undersampling, is a classic method carried out by reducing the number of observations in the largest class(es), motivating us to extend the idea for uplift modeling. Compared to other methods, such as oversampling small classes or re-weighting the observations (Haixiang et al., 2017), it has the advantage of reducing the computational cost and it often performs well in tasks with extreme imbalance and large sample size, such as click-stream modeling in e-commerce (McMahan et al., 2013).

Classifiers trained on undersampled data are not calibrated for the original data, but instead the probability estimates depend on the undersampling rate and true probability in a non-linear manner. Since uplift corresponds to the difference between two class probabilities, undersampling strategies for uplift modeling need to account for this. We propose to do this by applying undersampling in a stratified manner as explained next.

Our undersampling scheme drops observations so that $N_t = k * \tilde{N}_t$ where N_t is the number of observations before undersampling, $\tilde{N}_t$ is the number of observations after, and k is an undersampling factor to be determined. Our undersampling scheme only drops observations from the majority class (here assumed to be the negative class). This is done separately for treated and control observations with the same k. Thus the number of negative observations $\tilde{N}_{neg,t}$ after undersampling is

$$\tilde{N}_{neg,t} = \frac{1/k - \bar{p}(y|t)}{1 - \bar{p}(y|t)} \cdot N_{neg,t}. \tag{3}$$

Here $\frac{1/k-\bar{p}(y|t)}{1-\bar{p}(y|t)}$ is a constant that describes how large proportion of the negative observations are kept, whereas for the positive observations we always keep all. Note that this expression takes different values for $t = 0$ and $t = 1$. As a consequence, the average conversion rate in the undersampled data equals $k * \bar{p}(y|t)$ where $\bar{p}(y|t)$ is the average conversion rate in the original data. This linear dependence is a very desirable property, but strictly speaking it holds only for the averages, not for more general expressions involving x.

If we denote the true outcome probability by $p(y|x,t)$ and the probability after undersampling by $\tilde{p}(y|x,t)$, then in any local neighborhood of x we have

$$\tilde{p}(y|x,t) \approx \frac{p(y|x,t)}{p(y|x,t) + \frac{(1/k-\bar{p}(y|t))}{1-\bar{p}(y|t)} \cdot (1 - p(y|x,t))}. \tag{4}$$

For small $p(y|x,t)$ the denominator is dominated by the second term and $1 - p(y|x,t)$ is approximately 1. Consequently for small $\bar{p}(y|t)$ the second term becomes approximately $\frac{1}{k}$. With these assumptions the linear dependence extends to the more general case so that

$$\tilde{p}(y|x,t=0) \approx k \cdot p(y|x,t=0) \text{ and}$$
$$\tilde{p}(y|x,t=1) \approx k \cdot p(y|x,t=1).$$

Since both probabilities are multiplied by the factor k, we get an unbiased uplift estimate by normalizing $\tau(x) \approx \frac{1}{k}\tilde{\tau}(x)$, where $\tilde{\tau}(x)$ is the uplift as estimated from the undersampled data. Uplift can be estimated as usual but the result has to be divided by the undersampling factor k. This approximation does not hold for high conversion probabilities, but this is irrelevant as the procedure is designed for addressing cases with very low conversion probabilities. One particularly desirable property of this is that in any local neighborhood where the difference $p(y|x,t=1) - p(y|x,t=0) = 0$, the difference will be zero also after undersampling.

Figure 1 illustrates the method from a different perspective by plotting iso-uplift contours. Each line plots how the uplift τ^* estimated from undersampled data should change for some fixed true uplift τ as function of the conversion rate $p(y|do(t=0))$. Here we used $k = 250$ and $p(t=1) = 0.002$ $\forall t \in \{0,1\}$ that roughly matches the values for the Criteo data.

We observe that the signs are correctly retained. For modeling uplift we would also need to retain the ranks of individual observations, but due to the non-linear transformations this is not guaranteed. The true uplift for all points along the blue curve is 0.2% and observations on this curve should ideally all be ranked higher than observations along the orange curve (0.1%). Due to the curves not being constant, an observation on the orange curve may

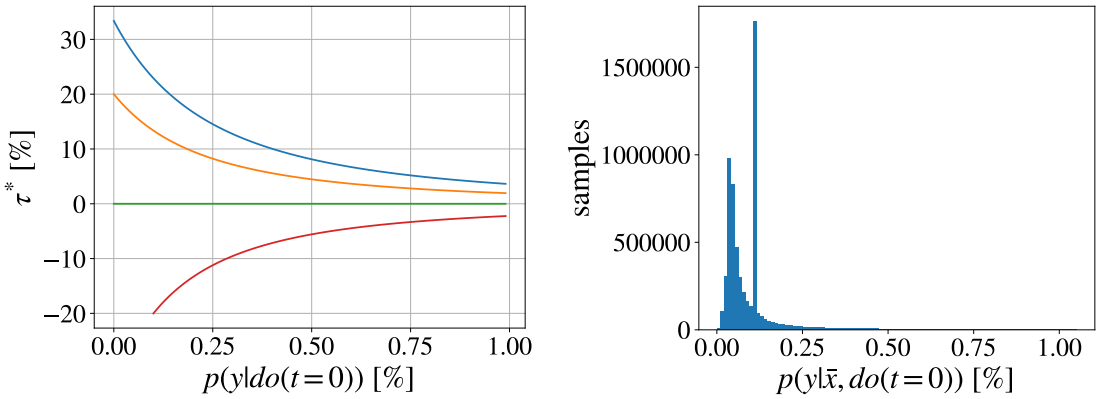

Figure 1: Illustration of the effect of undersampling on uplift estimates. (*Left:*) Iso-uplift contours as function of base conversion probability $p(y|do(t=0))$ where τ^* refers to the uplift that *should* be present in data after undersampling. The contours in order from top down correspond to 0.2%, 0.1%, 0.0%, and −0.1% of uplift. (*Right:*) The histogram showing 97.5% of all predicted conversion rates for the Criteo data in the control group using logistic regression. The majority of the observations fall between 0.0% and 0.25%. Note that the scales of the horizontal axes in the plots are from 0% to 1%.

be ranked above one on the blue curve if it has considerably lower conversion probability. The observations with the smallest conversion probabilities $p(y|do(t=0))$ thus have higher estimated uplift (for positive uplifts) and may outrank observations with higher true uplift and high conversion probability. If the difference in conversion probabilities is small this can only happen for small differences in uplift. The conversion probabilities generally tend to be very close. For the Criteo data almost all conversion probabilities fall between 0.0% and 0.25% and hence such violations are rare in practice. The iso-uplift contours in this region are not constant, but the range is narrow enough for the the linearity assumption to hold well. The high peak in the histogram is caused by a large number of users with identical features, a property of the Criteo data.

To further justify the above approach, we clarify why some intuitively appealing alternatives would not work. For DC one could use classical undersampling separately for treatment and control models, followed by separate calibration of both before estimating the uplift. This would require selecting two undersampling rates and running a calibration algorithm twice, and has the risk of accumulation of calibration error in computing the difference of two separately calibrated estimates. More importantly, this approach would not be compatible with any other than DC-based models. Another intuitive alternative would be to drop majority observations in treatment and control groups at the same rate. Without further calibration this would introduce a new source of bias. More specifically, the group with lower conversion rate would overall get too high probability estimates relative to the estimates for the other group. The difference between these two would be systemati-

cally biased and one consequence is that the sign of the uplift estimates after undersampling would not necessarily equal the true sign. In contrast, our stratified approach always retains the correct sign. While it could be possible to device a calibration algorithm to correct for this potentially severe bias, our proposed approach has the advantage of directly providing nearly unbiased estimates for low overall conversion rates.

3.2. Choice of Undersampling Factor

The approach has one key parameter, the undersampling factor k, which needs to be determined. To improve the balance we must have $k \geq 1$ ($k = 1$ indicates no undersampling) but we also have a natural upper bound since we cannot continue undersampling once all of the negative observations in either group have been exhausted. This implies $k < \min\left(\frac{1}{\bar{p}(y|t=0)}, \frac{1}{\bar{p}(y|t=1)}\right)$. We use validation to select the best k in this range, using 2/3 of the development observations for training the classifier for a set of valid k and evaluating the performance (AUUC) on the remaining 1/3. Calibration is then performed on the validation data, using the scores $s(x)$ obtained with the optimal k.

3.3. Calibration

A classifier is well-calibrated if the class probabilities accurately represent the true probabilities, evaluated for a group of observations by comparing the average predictions to the empirical class rate (Naeini and Cooper, 2016). We extend this to uplift modeling calling the model well-calibrated if the average uplift estimate for a group of observations is close to the difference in empirical conversion rates between treated and untreated observations.

To correct for the imperfections caused by the linearity assumption, we calibrate our uplift estimates. While there are many calibration methods available for classification (see Section 4 in Guo et al. (2017) for a recent review), we chose to extend isotonic regression (Zadrozny and Elkan, 2002) to uplift modeling by building on revert-labeling (Athey and Imbens, 2015). Revert-labeling is a relabeling scheme that converts the uplift problem into a regression problem using

$$r_i = \frac{y_i(t_i - p(t=1))}{p(t=1)(1-p(t=1))}$$

where $p(t = 1)$ denotes the probability of a randomly selected observation belonging to the treatment group.[1] Modeling $f(s) = r$ directly results in calibrated estimates so that $f(s) \approx \tau(s)$; see Athey and Imbens (2015) for derivations and details. We use isotonic regression to model this.

The full calibration method is described in Algorithm 1. Note that other non-parametric calibration methods than isotonic regression could be used with only minor adjustements.

3.4. Measuring Calibration

The uplift performance metric of AUUC is insensitive to calibration as it is based on ranking alone. Hence, we need separate metrics for evaluating calibration. We propose two new metrics extending the calibration metrics of *expected calibration error* and *maximum*

1. While revert-labeling is closely related to the class-variable transformation, the former converts the learning problem into a regression problem whereas the latter converts it to a classification problem.

calibration error proposed by Naeini et al. (2015) to evaluate calibration of classification methods. We call the metrics *expected uplift calibration error* (EUCE) and *maximum uplift calibration error* (MUCE). Both are based on binning the ranked observations into m bins of $C = \frac{N}{m}$ observations per bin so that the observations with the smallest scores are in the first bin etc. For each bin i we estimate the uplift directly as the difference between the empirical rates of positive outcomes for the treatment and control observations as follows:

$$b_i = \frac{\sum_{j=(i-1)C}^{iC} y_{j,t=1}}{N_{i,t=1}} - \frac{\sum_{j=(i-1)C}^{iC} y_{j,t=0}}{N_{i,t=0}}.$$

By denoting the mean uplift estimates of the model for each bin as $u_i = \frac{1}{C}\sum_{j=(i-1)C}^{iC} \tau(x_j)$, the expected uplift calibration error (EUCE) and the maximum uplift calibration error (MUCE) can then be defined as

$$EUCE = \sum_{i=1}^{m} \frac{1}{m}|b_i - u_i|, \text{ and } MUCE = \max_i |b_i - u_i|.$$

We use $m = 100$ bins, which is larger than $m = 10$ used by Naeini et al. (2015) for evaluating classification calibration, since we work with a considerably larger data set. However, the metrics are general and can be used with any m.

4. Experiments

The proposed technique is compatible with any uplift modeling method. We demonstrate it with DC and CVT explained in Section 2.2 using two choices for the base classifier for estimating the class probabilities: logistic regression (LR) and random forest (RF) as implemented in scikit-learn (Pedregosa et al., 2011). We denote these models as DC-LR, DC-RF, CVT-LR and CVT-RF, and evaluate them with and without undersampling. DC-LR has been identified as the best performing uplift model for the Criteo data in two separate studies (Diemert et al., 2018; Semenova and Temirkaeva, 2019), whereas RF was chosen because it produces well-calibrated probabilities (Niculescu-Mizil and Caruana, 2005). Methods based on causal trees and forests were excluded from the comparison due to their poor performance in earlier evaluations (Belbahri et al., 2020; Gubela et al., 2020; Munting, 2020) and excessively high memory consumption.

We ran the experiments on the Criteo data (Diemert et al., 2018). It is the largest publicly available uplift data with 25 million observations. It has two class labels, 'visit' and 'conversion,' and 12 anonymized features. We split the data so that randomly selected 75% of the observations are used as development set, and the remaining 25% are used as test set. We evaluate the uplift using AUUC and use the newly proposed EUCE and MUCE for measuring the calibration.

The main experiment evaluates the effect of undersampling in tasks with high class imbalance, and therefore we use the conversion-label with staggering $99.8 : 0.2$ ratio of negative to positive observations. The other experiment investigates the method over a range of different imbalance rates. We do this on semi-synthetic data obtained by undersampling the positive observations of the visit-label of the data. For visits the real positive

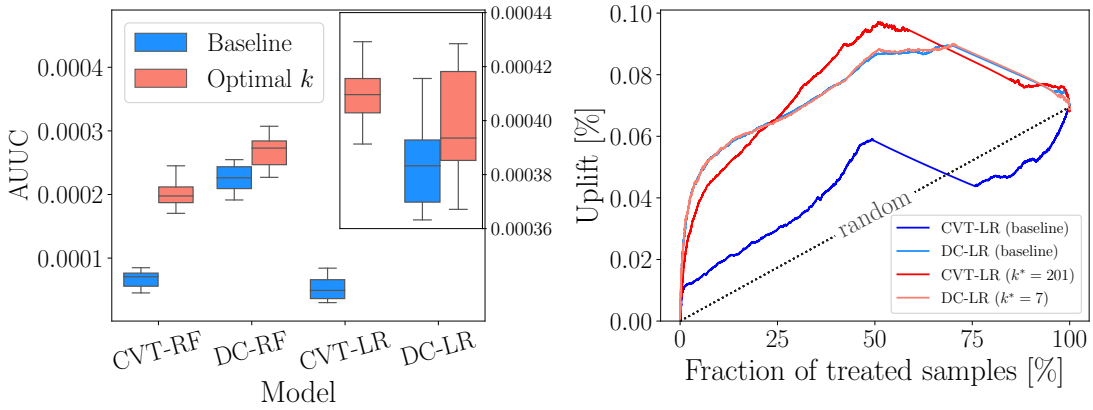

Figure 2: Results on the conversion-label of Criteo data. (*Left:*) Test set AUUC and variability over 10 different random data splits. Undersampling improves the performance for every model, but dramatically for CVT-based methods. CVT-LR with undersampling outperforms all other methods (paired t-test; $p < 0.01$). Note that the three rightmost items use a different scale for clarity.
(*Right:*) Uplift as function of the fraction of observations being treated. The vertical axis measures the increase in conversion rate as *percentage points*. The peak of CVT-LR ($k = 201$) around 50% treatment rate is just above 0.09%, and hence the conversion rate at this treatment rate is approximately $0.17\% + 0.09\% = 0.26\%$.

rate is 4.1%, and to study the transition we create five additional data sets with rates $0.1\%, 0.2\%, 0.5\%, 1\%$ and 2% by randomly dropping positive observations.[2]

5. Results

Table 1 reports the results for the main experiment providing all evaluation metrics for four models trained in three ways: (a) trained on the full data as is conventionally done (baseline), (b) using the proposed undersampling approach (uncalibrated), and (c) using the proposed undersampling approach together with calibration (calibrated). The results are averages over ten different random splits into development and test sets so that the optimal k was chosen without accessing the test data. For all models, undersampling consistently improves AUUC and calibration as measured by EUCE and MUCE. The calibration method further improves calibration in three of the cases, but at the expense of a drop in AUUC.

Figure 2 presents AUUC representing also the variability over the ten folds. As seen already in Table 1, CVT-LR with undersampling performs best and to our knowledge reaches new state-of-the-art for this task. DC-LR with undersampling is almost as good. The effect of undersampling is best seen in CVT that benefits dramatically, but also for DC we see clear gains, especially with the RF classifier. We also show the uplift curves for CVT-LR and DC-LR with and without undersampling. CVT without undersampling fails

2. Code for reproducing the experiments is available at https://github.com/Trinli/ACML-2021-uplift.

Table 1: Average performance on test set of the proposed method (with and without calibration) vs. the baseline without undersampling. For all model variants, undersampling improves the primary metric of AUUC. Calibration helps in terms of the calibration metrics (note that EUCE and MUCE values are in $[0, 2]$), but reduces AUUC especially for DC-based methods.

MODEL	**AUUC** $\uparrow$	**EUCE** $\downarrow$	**MUCE** $\downarrow$
DC-LR (baseline)	0.00038	0.0006	0.0293
DC-LR (uncalibrated)	0.00040	**0.0004**	0.0074
DC-LR (calibrated)	0.00031	0.0006	0.0153
CVT-LR (baseline)	0.00005	0.6890	0.8618
CVT-LR (uncalibrated)	**0.00041**	0.0008	0.0122
CVT-LR (calibrated)	0.00039	0.0006	0.0057
DC-RF (baseline)	0.00023	0.0020	0.0743
DC-RF (uncalibrated)	0.00027	0.0007	0.0121
DC-RF (calibrated)	0.00019	0.0005	0.0069
CVT-RF (baseline)	0.00007	0.7536	1.0003
CVT-RF (uncalibrated)	0.00020	0.0016	0.0052
CVT-RF (calibrated)	0.00020	0.0005	**0.0019**

Table 2: AUUC for baseline (no undersampling) and optimal undersampling for the conversion rate experiments. The last column $\tilde{p}(y)$ is the resulting positive rate in the training set after optimal undersampling. The conversion rate $p(y)$ in the synthetic data was varied from 0.001 to 0.041 by dropping positive observations (visit-label) from the Criteo dataset.

$\mathbf{p(y)}$	**Baseline**	**Optimal**	**Optimal** $\tilde{\mathbf{p}}(\mathbf{y})$
0.001	-0.00003	0.00016	0.300
0.002	-0.00002	0.00037	0.408
0.005	0.00024	0.00089	0.300
0.010	0.00134	0.00181	0.380
0.020	0.00374	0.00374	0.020
0.041	0.00768	0.00768	0.041

catastrophically, but with undersampling it outperforms all other models both in AUUC and peak conversion rate.

Figure 3 (left) shows that the method is robust to the choice of the undersampling factor, and that very aggressive undersampling can be performed to reduce computation time. AUUC is nearly identical for any k between 100 and 300, and the results over different folds are highly consistent. The plots show the behavior for CVT-LR, but the results are similar for other choices (omitted due to lack of space).

Table 2 shows the best-performing CVT-LR method with and without undersampling on the semi-synthetic data sets with varying class imbalance, so that AUUC is scaled by

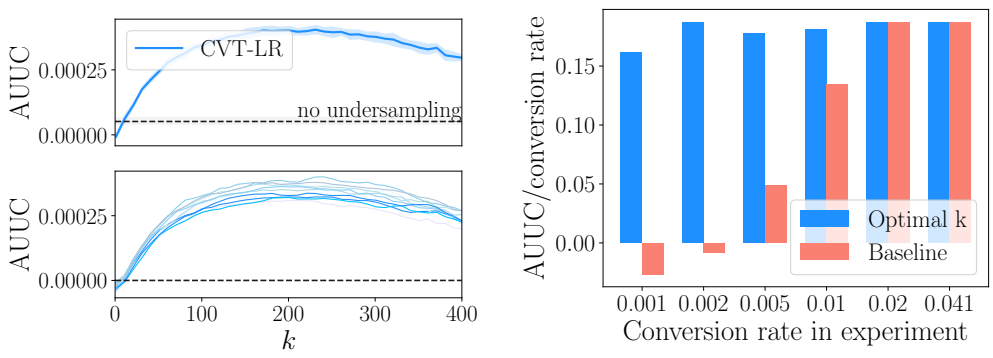

Figure 3: (*Left:*) Average (top) and individual run (bottom) test AUUC as function of k, with consistent performance over a wide range of values.
(*Right:*) CVT-LR performance with and without undersampling for different true conversion rates. AUUC is here normalized by the conversion rate to make the results comparable. Undersampling has clear benefits for all rates below 2%, and for low rates the baseline with no undersampling is worse than random.

the inverse conversion rate to normalize for differences in best achievable uplift for better comparison. This is also visualized in Figure 3 (right). The benefit of undersampling increases when the imbalance grows. For rates above 2% uplift modeling works well even without accounting for the imbalance. For the smallest rates standard CVT-LR is worse than random, but undersampling fully mitigates this.

6. Discussion

The AUUC values in Table 1 can be contrasted with earlier results. The DC-LR baseline outperforming other baselines is consistent with both Diemert et al. (2018) and Semenova and Temirkaeva (2019). The former compared DC and CVT with LR as base-learner, whereas the latter considered also RF, XGBoost SVM, and uplift random forests (Guelman et al., 2015). DC-LR was identified as the best method in both studies. Using undersampling we outperform that method by 6.5% with CVT-LR and by 3.9% with DC-LR. As an interesting sidenote, this is in contrast to the results by Rudaś and Jaroszewicz (2018) that showed a DC model should theoretically perform better than a CVT model, although their results applies only to regression.

The new calibration metrics reveal clear differences in calibration, providing another view of how CVT-based methods perform poorly without undersampling and verifying that the proposed calibration method works. For DC, however, calibration negatively impacts the main evaluation metric of AUUC. We currently recommend calibrating CVT-based methods, but leaving DC-based methods uncalibrated.

Finally, Figure 3 reveals an essential observation: for true conversion rates over 2%, conventional uplift models work well, but fail catastrophically for rates below 1%, at least with this data set. Based on figure 3 it seems the reason for this is not that the linearity

assumption is violated with conversion rates above 2%, but that the algorithm has identified most if not all of the uplift identifiable in the data. Typically the rates are smaller for the outcomes that are more important for businesses (e.g. 'purchase' vs. 'visit') and hence undersampling is particularly valuable for the use cases with the highest value. For higher rates one can safely proceed without undersampling.

We demonstrated improved performance on both labels of the Criteo data that can be seen as two separate tasks, but nevertheless a limitation of our work is that the results are shown only for one data source. This is because there are no other suitable sources. While virtually all companies in e-commerce work with this kind of data, they are not publishing their data due to regulations or privacy concerns. For instance, Athey and Imbens (2015) used search engine data and Zhao et al. (2017) used airline reservation data that both would have been large enough but the data sets remain proprietary. The few publicly available data sets, in turn, are too small or otherwise unrealistic. In particular, the second and third largest public data (the Voter data (Gerber et al., 2008) and the Hillstrom data (Radcliffe and Surry, 1999)) do not have high class imbalance and they are also too small for reliably carrying out a semi-synthetic experiment where the class imbalance is artificially increased as we did with the visit-label of the Criteo data. For both we would only have around 200 positive observations in the treatment group for class rate of 1% and around 20 positive observations for rate 0.1%. To cope with the lack of data some authors have used synthetic data (Kuusisto et al., 2014), or resorted to artificially creating an uplift data set by redefining some feature as the label for an uplift task (Zaniewicz and Jaroszewicz, 2013). Such data does not reflect any realistic setting and would not provide additional value.

7. Conclusion

We showed how uplift models can be significantly improved in cases with high class imbalance (e.g. low conversion rates), a property that holds almost universally in e-commerce. The approach is intuitive and easy to use in conjunction with current uplift methods only requiring a simple pre-processing step of stratified undersampling and post-processing to correct for potential calibration errors. Even though undersampling has been studied extensively in classification, we are the first to apply it to uplift modeling and presented all of the necessary technical details.

We used undersampling for two prominent uplift models on the largest available benchmark data and observed that the improvement depends on the method and the amount of class imbalance. For some cases (CVT-based methods, high class imbalance) undersampling is the technique that improves an uplift model from useless to great, whereas in other cases (DC-LR, data with less imbalance) there is virtually no improvement. We have not yet encountered a case where using the proposed strategy would hurt. Hence, we consider it safe to adopt the strategy as part of a standard pipeline for uplift modeling; the only drawback is the computational cost of determining k. Even though additional computation is spent, the final model is trained on less data due to aggressive undersampling and is very fast. For the optimal CVT-LR model we only used 0.4% of the data for training and still reached a new state-of-the-art result.

The most common application of uplift modeling is currently in e-commerce, and we wrote the paper using terminology and benchmark data from that field. However, we

would like to encourage researchers to explore applications also beyond e-commerce, and consequently included references to examples of uses in medicine and education.

Acknowledgments

This work was supported by Business Finland (project MINERAL) and the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI).

References

S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *arXiv*, 2015.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *Annals of Statistics*, 47 (2):1179–1203, 2019.

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, 2011.

M. Belbahri, O. Gandouet, and G. Kazma. Adapting Neural Networks for Uplift Models. In *Proc. of the 37th International Conference on Machine Learning*, 2020.

A. Betlei, E. Diemert, and M. Amini. Uplift Prediction with Dependent Feature Representation in Imbalanced Treatment and Control Conditions. In *Lecture Notes in Computer Science, vol 11305.*, volume V, pages 47–55. Springer, Cham, 2018.

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

E. Diemert, A. Betlei, C. Renaudin, and A. Massih-Reza. A Large Scale Benchmark for Uplift Modeling. *Proc. of the AdKDD and TargetAd Workshop*, 2018.

A. S. Gerber, D. P. Green, and C. W. Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1):33–48, 2008.

R. M. Gubela, S. Lessmann, and S. Jaroszewicz. Response transformation and profit decomposition for revenue uplift modeling. *European Journal of Operational Research*, 283 (2):647–661, 2020.

L. Guelman, M. Guillén, and A. M Pérez-Marín. Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study. *UB Riskcenter Working Paper Series*, 6:1–33, 2014.

L. Guelman, M. Guillén, and A. M. Pérez-Marín. Uplift Random Forests. *Cybernetics and Systems*, 46(3-4):230–248, 2015.

C. Guo, G. Pleiss, Y. Sun, and K.Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Input: N training observations with score s_i, label y_i, and group t_i.

Apply revert-label transformation:
$N_t = \sum_{i=1}^{N}(\mathbb{I}(t_i = 1))$
$N_c = \sum_{i=1}^{N}(\mathbb{I}(t_i = 0))$
$p_t = N_t/(N_t + N_c)$
for i *in* $1 : N$ **do**
  $r_i \leftarrow \frac{y_i(t_i - p_t)}{p_t(1-p_t)}$
end
Sort observations ascending by score s
Assign every observation to a separate bin stored as list l
Merge bins with identical scores
Apply pairwise-adjacent violators algorithm below:
$i = 1$
while $i <$ *number of bins in list* l **do**
  if $i < 1$ **then**
    $i \leftarrow 1$
  end
  if $l_i.u >= l_{i+1}.u$ **then**
    $l_i.s_{min} \leftarrow l_i.s_{min}$
    $l_i.s_{max} \leftarrow l_{i+1}.s_{max}$
    $l_i.r \leftarrow l_i.r \cup l_{i+1}.r$
    drop l_{i+1}
    $l_i.u \leftarrow \text{mean}(l_i.r)$
    $i \leftarrow i - 1$
  end
  else
    $i \leftarrow i + 1$
  end
end
Output: List l of bins, each represented by its extent $[l_i.s_{min}, l_i.s_{max}]$ and uplift estimate $l_i.u$.

Algorithm 1: Calibration algorithm for uplift modeling combining the revert-label approach and isotonic regression. Observations are grouped according to their scores into bins (the number of which is determined during the algorithm), stored as list l. Each bin is characterized by its boundaries $[l_i.s_{min}, l_i.s_{max}]$ and the uplift estimate $l_i.u$.

P. Gutierrez and J. Gérardy. Causal Inference and Uplift Modelling: A Review of the Literature. *Proc. of The 3rd International Conference on Predictive Applications and APIs*, 67:1–13, 2017.

G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):968–970, 1986.

M. Jaskowski and S. Jaroszewicz. Uplift modeling for clinical trial data. *ICML Workshop on Clinical Data Analysis*, 2012.

Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, 52(4), 2019.

F. Kuusisto, V. S. Costa, H. Nassif, E. Burnside, D. Page, and J. Shavlik. Support vector machines for differential prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer, 2014.

L. Yi-Ting Lai. *Influential Marketing: A New Direct Marketing Strategy Addressing the Existence of Voluntary Buyers*. PhD thesis, University of British Columbia, 2006.

H. B. McMahan, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, ..., and E. Davydov. Ad click prediction. *Proc. of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13*, page 1222, 2013.

M. Munting. An explorative study towards the feasibility of uplift modeling within a direct marketing setting and a web-based setting. Master's thesis, University of Twente, 2020.

M. P. Naeini and G. F. Cooper. Binary Classifier Calibration using an Ensemble of Near Isotonic Regression Models. In *2016 IEEE 16th International Conference on Data Mining*, pages 360–369, 2016.

M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proc. of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015:2901–2907, 2015.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proc. of the 22nd international conference on Machine learning ICML 05*, number 1999, pages 625–632, 2005.

D. Olaya, K. Coussement, and W. Verbeke. A survey and benchmarking study of multi-treatment uplift modeling. *Data Mining and Knowledge Discovery*, 34(2):273–308, 2020a.

D. Olaya, J. Vásquez, S. Maldonado, J. Miranda, and W. Verbeke. Uplift Modeling for preventing student dropout in higher education. *Decision Support Systems*, 134(May): 113320, 2020b.

J. Pearl. *Causality*. Cambridge University Press, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, ..., and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

N. J. Radcliffe and P. D. Surry. Differential response analysis: Modelling true response by isolating the effect of a single action. *Credit Scoring and Credit Control VI*, 1999.

M. Richardson, R. Ragno, and E. Dominowska. Predicting Clicks: Estimating the Click-Through Rate for New Ads. In *Proc. of the 16th international conference on World Wide Web*, pages 521–529, 2007.

Rubin, D. B. Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

K. Rudaś and S. Jaroszewicz. Linear regression for uplift modeling. *Data Mining and Knowledge Discovery*, pages 1–31, 2018.

P. Rzepakowski and S. Jaroszewicz. Decision trees for uplift modeling. *Proc. - IEEE International Conference on Data Mining, ICDM*, pages 441–450, 2010.

D. Semenova and M. Temirkaeva. The comparison of methods for individual treatment effect detection? In *CEUR Workshop Proc.*, volume 2479, pages 46–56, 2019.

B. C. Wallace and I. J. Dahabreh. Improving class probability estimates for imbalanced data. *Knowledge and Information Systems*, 41(1):33–52, 2014.

G. M. Weiss. Mining with Rarity: A Unifying Framework. *SIGKDD Explorations*, 6(1): 7–19, 2004.

G. M. Weiss and F. Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. *Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 02*, pages 694–699, 2002.

L. Zaniewicz and S. Jaroszewicz. Support vector machines for uplift modeling. In *IEEE 13th International Conference on Data Mining Workshops*, pages 131–138. IEEE, 2013.

Y. Zhao, X. Fang, and D. Simchi-Levi. Uplift modeling with multiple treatments and general response types. *Proc. of the 17th SIAM International Conference on Data Mining, SDM 2017*, pages 588–596, 2017.