JMLR: Workshop and Conference Proceedings 67:1–13, 2016

Causal Inference and Uplift Modeling
A review of the literature

Pierre Gutierrez *pierre.gutierrez@dataiku.com*
Jean-Yves Gérardy *jean-yves.gerardy@dataiku.com*

Abstract

Uplift modeling refers to the set of techniques used to model the incremental impact of an action or treatment on a customer outcome. Uplift modeling is therefore both a Causal Inference problem and a Machine Learning one. The literature on uplift is split into 3 main approaches–the Two-Model approach, the Class Transformation approach and modeling uplift directly. Unfortunately, in the absence of a common framework of causal inference and notation, it can be quite difficult to assess those three methods. In this paper, we use the Rubin (1974) model of causal inference and its modern "econometrics" notation to provide a clear comparison of the three approaches and generalize one of them. To our knowledge, this is the first paper that provides a unified review of the uplift literature. Moreover, our paper contributes to the literature by showing that, in the limit, minimizing the Mean Square Error (MSE) formula with respect to a causal effect estimator is equivalent to minimizing the MSE in which the unobserved treatment effect is replaced by a modified target variable. Finally, we hope that our paper will be of use to researchers interested in applying Machine Learning techniques to causal inference problems in a business context as well as in other fields: medicine, sociology or economics.

Keywords: Uplift Modeling, Causal Inference, Machine Learning

1. Introduction

Uplift modeling refers to the set of techniques that a company may use to estimate customer uplift, that is, the effect of an action on some customer outcome. For example, a manager at a telecommunication company could be interested in estimating the effect of sending a promotional e-mail to different customer profiles on their propensity to renew their phone plan in the next period. With that information at hand, the manager is able to efficiently target customers.

Estimating customer uplift is both a Causal Inference and a Machine Learning problem. It is a causal inference problem because one needs to estimate the difference between two outcomes that are mutually exclusive for an individual (either person i receives a promotional e-mail or does not receive it). To overcome this counter-factual nature, uplift modeling crucially relies on randomized experiments, i.e. the random assignment of customers to either receive the treatment (the treatment group) or not (the control group). Uplift modeling is also a machine learning problem as one needs to train different models and select the one that yields the most reliable uplift prediction according to some performance metrics. This requires sensible cross-validation strategies along with potential feature engineering.

The uplift modeling literature proposes three main approaches to combine this Causal Inference aspect with the Machine Learning one. This gives rise to different uplift metrics and evaluation methods that are not easily comparable. We blame this difficulty in comparison on the fact that researchers are not using a common framework and notation of causal inference.

In this paper, we present an overview of the 3 different approaches–the Two-Model approach, the Class Transformation approach and modeling uplift directly–using the Rubin (1974) model of causal inference as a common frame of reference. We also strive to adopt a methodical "Machine Learning" way of building the predictive models. Our goal in this paper is to provide a framework that unifies all the different uplift approaches so as to make their comparison and evaluation easier. Given the growing interest in using Machine Learning tools to do causal inference (especially among econometricians, see recent papers for example Athey and Imbens (2015b)) we hope that our paper will point researchers interested in that field to resources available in the uplift literature. Finally, our paper contributes to the literature by showing that, in the limit, the uplift estimator minimizing the Mean Square Error (MSE) also minimizes the MSE in which the unobserved uplift is replaced by a modified target variable.

The rest of the paper is organized as follows: Section 2 introduces the causal inference framework and notations, section 3 describes the different approaches to Uplift modeling and section 4 discourses on how to properly evaluate Uplift models. We conclude in section 5.

2. Causal Inference: Basics

We rely on the Rubin (1974) model of causal inference and use the standard notation of the econometric literature. At the core of the model are the notions of potential outcomes and causal effects. We consider a framework with N individuals indexed by i. Denoting $Y_i(1)$ person i's outcome when he receives the active treatment and $Y_i(0)$ person i's outcome when he receives the control treatment, the *causal effect*, τ_i, of the active treatment *vis-à-vis* the control treatment is given by:

$$\tau_i = Y_i(1) - Y_i(0) \tag{1}$$

Researchers are typically interested in estimating the Conditional Average Treatment Effect (CATE), that is, the expected causal effect of the active treatment for a subgroup in the population:

$$CATE : \tau(X_i) = E[Y_i(1)|X_i] - E[Y_i(0)|X_i] \tag{2}$$

Where X_i is a $L \times 1$ vector of random variables (features). Of course, we will never observe both $Y_i(1)$ and $Y_i(0)$. Letting $W_i \in 0, 1$ be a binary variable taking on value 1 if person i receives the active treatment, and 0 if person i receives the control treatment, person i's observed outcome is actually:

$$Y_i^{obs} = W_i Y_i(1) + (1 - W_i) Y_i(0) \tag{3}$$

A popular but unfortunately wrong belief is that one can always estimate the CATE from observational data by simply computing the empirical counterpart of

$$E[Y_i^{obs}|X_i = x, W_i = 1] - E[Y_i^{obs}|X_i = x, W_i = 0] \tag{4}$$

This won't identify the CATE unless one is willing to assume that W_i is independent of $Y(1)$ and $Y(0)$ conditional on X_i. This assumption is the so-called *Unconfoundedness Assumption* or the *Conditional Independence Assumption* (CIA) found in the social sciences and medical literature. This assumption holds true when treatment assignment is random conditional on X_i.

$$CIA : \{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i|X_i \tag{5}$$

Before proceeding to uplift modeling, let us introduce additional useful notation. Let us define the propensity score, $p(X_i) = P(W_i = 1|X_i)$, i.e. the probability of treatment given X_i.

3. Uplift Modeling

Companies (typically in telecommunication or e-business sectors) are interested in uplift modeling to estimate the effect of an action on some customer outcome. For example, a gym owner might be interested in estimating the effect of sending a promotional e-mail to a customer of observed characteristics X_i on their propensity to renew their membership in the next period. In other words, uplift modeling amounts to estimating a CATE. Although companies can easily conduct randomized experiments so as to ensure that the CIA holds, the fact that we never observe the true τ_i makes it seemingly impossible to use standard supervised learning algorithms to estimate it. If we suppose for a moment that τ_i was indeed observed, we would simply split the data into a train and a test set and use one of the many available algorithms to come up with the approximation of the CATE $\hat{\tau}(X_i)$ that minimizes a loss function on the training data. We would then evaluate our model using one or more metrics (AUC, F1 score, Accuracy etc) on the test data.

The uplift literature has proposed three main approaches to estimate $\tau(X_i)$ despite the absence of the ground truth. The first one is the Two-Model approach which consists in building two predictive models, one using the treatment group data and the other using the control group data, exclusively. The second approach is referred to as the Class Variable Transformation introduced by Jaskowski and Jaroszewicz (2012) in the case of a binary outcome variable. The third one is to model uplift directly through the modification of well known classification machine learning algorithms such as decision tree (Rzepakowski and Jaroszewicz (2012), Radcliffe and Surry (2011), Athey and Imbens (2015b)), random forest (Sołtys et al. (2015), Wager and Athey (2015)) or SVM (Support Vector Machines,Zaniewicz and Jaroszewicz (2013)). We present each approach in turn.

3.1. The Two-Model Approach

The Two-Model approach has been applied in several uplift papers (Radcliffe (2007), Nassif et al. (2013)) and is often used as a baseline model. This approach was also introduced in the more recent branch of the causal inference literature that is experimenting with

modern machine learning techniques (see the Two Tree (TT) algorithm in Athey and Imbens (2015b)).

The approach consists in modeling $E[Y_i(1)|X_i]$ and $E[Y_i(0)|X_i]$ separately, using the treatment group data and the control group data, respectively. The advantage of the Two-Model approach resides in its simplicity. Because inference is done separately in the treated and control group, state-of-the-art machine learning algorithms such as Random Forest (Breiman (2001)) or XGBoost (Chen and Guestrin (2016)) can be used "as is" whether it be in a regression setting or a (multi-)classification one. Both models can achieve good prediction performance, separately. However, for uplift purposes, although the approach has been seen to perform well, some authors (Zaniewicz and Jaroszewicz (2013), Athey and Imbens (2015b)) show that it is often outperformed by other methods. One reason is that the two models focus on predicting the outcome separately and can therefore miss the "weaker" uplift signal. Radcliffe and Surry (2011) illustrate this phenomenon in a simulation study.

The subject of how we can evaluate the performance of uplift models and thus compare them will be the subject of section 4.

3.2. The Class Transformation Method

The Class Transformation method was introduced by Jaskowski and Jaroszewicz (2012) in the case of binary outcome variable ($Y_i^{obs} = \{0, 1\}$). The methods consists in creating the following target variable:

$$Z_i = Y_i^{obs} W_i + (1 - Y_i^{obs})(1 - W_i) \tag{6}$$

The new target, Z_i, is therefore equal to one in either following cases: **1)** the observation belongs to the treatment group and $Y_i^{obs} = 1$ or **2)** the observation belongs to the control group and $Y_i^{obs} = 0$. In all other cases, the target takes on value zero.

Under the assumption that control and treated groups are balanced across all profiles of individual (that is, $p(X_i = x) = 1/2$ for all x), (Jaskowski and Jaroszewicz (2012)) proved that:

$$\tau(X_i) = 2P(Z_i = 1|X_i) - 1 \tag{7}$$

Uplift modeling thus boils down to modeling $P(Z_i = 1|X_i)$, (*i.e.* $E[Z_i = 1|X_i]$). The Class Transformation method is popular because it tends to perform better than the Two-Model approach while still remaining simple; any off-the-shelf classifier can be used to model $E[Z_i = 1|X_i]$. However, the two assumptions (binary outcome variable and balanced dataset between control and treatments) might seem too restrictive. Fortunately, a generalization to unbalanced treatment assignment and to regression setups can be borrowed from (Athey and Imbens (2015b)) who propose to estimate the CATE by applying standard machine learning algorithms to the following transformed outcome variable, Y_i^*:

$$Y_i^* = Y_i(1)\frac{W_i}{\hat{p}(X_i)} - Y_i(0)\frac{(1 - W_i)}{(1 - \hat{p}(X_i))} \tag{8}$$

Where $\hat{p}(x)$ is a consistent estimator of the propensity score, $p(X_i)$[1]. This transformed outcome has the key property that, under the CIA, its expectation conditional on X_i is

1. When the sample size of the train set is large, any off-the-shelf ML method will work to estimate $p(X_i)$.

equal to the CATE (Angrist and Pischke (2008), Athey and Imbens (2015b)):

$$E[Y_i^*|X_i] = \tau(X_i) \tag{9}$$

This property means that any consistent estimator of $E[Y_i^*|X_i]$ is also a consistent estimator of $\tau(X_i)$.

Note that in the case with complete randomization ($p(X_i = x) = 1/2$ for all x) and binary outcome Y_i^{obs}, combining Equation 3 with Equation 8 allows us to write Equation 6 as:

$$Z_i = \frac{1}{2}Y_i^* + (1 - W_i) \tag{10}$$

and thus $2E(Z_i|X_i) = E(Y_i^*|X_i) + 1$ which is equivalent to Equation 7.

Finally, let us point out that the Class Transformation method was also used in Lai (2006). Shaar et al. (2016) also proposed a re-weighted uplift formulation by multiplying the Z_i estimated probabilities by case proportions.

$$P(Z_i = 1)\frac{1}{N}\Big[\sum_i(\mathbb{1}_{Z_i=1})\Big] - P(Z_i = 0)\frac{1}{N}\Big[\sum_i(\mathbb{1}_{Z_i=0})\Big] \tag{11}$$

The authors also introduced other class transformation approaches called "Reflective" and "Pessimistic".

3.3. Modeling Uplift Directly

The third and last approach consists in modifying existing machine learning algorithms to directly infer a treatment effect. Lo (2002) proposed a strategy based on logistic regression, Su et al. (2012) and Guelman et al. (2014) focused on k-nearest neighbors while Zaniewicz and Jaroszewicz (2013) proposed a modification of the SVM model. The most popular methods in the literature remain the tree-based ones (see Hansotia and Rukstales (2002), Radcliffe and Surry (2011), Rzepakowski and Jaroszewicz (2012) and Athey and Imbens (2015b)). Finally, Sołtys et al. (2015), Wager and Athey (2015) or Guelman et al. (2015) provided a generalization to ensemble methods. In the following section we focus on tree-based methods and discuss the principal task for tree generation: the split criterion choice[2].

Formally, in the case of a balanced randomized experiment, where the propensity score $p(X_i = x) = 1/2$ for all x, the estimator of the average treatment effect (or uplift) $\hat{\tau}$ is given by:

$$\hat{\tau} = \underbrace{\frac{\sum_i Y_i^{obs} W_i}{\sum_i W_i}}_{p} - \underbrace{\frac{\sum_i Y_i^{obs}(1 - W_i)}{\sum_i (1 - W_i)}}_{q} \tag{12}$$

This corresponds to the difference in the sample average outcome between treated and untreated observations.

The first split criterion, proposed by Hansotia and Rukstales (2002), is the difference of uplift between the two leaves:

$$\Delta = |\hat{\tau}_{Left} - \hat{\tau}_{Right}| \tag{13}$$

2. For simplicity of exposition, we focus on binary trees.

Where the subscripts $Left$ and $Right$ refer to the estimator of equation 12 computed using observations present in the left and right leaves following the split.

Rzepakowski and Jaroszewicz (2012) then proposed three new criteria based on information theory of the form:

$$\Delta_{gain} = D_{after\_split}(P^T, P^C) - D_{before\_split}(P^T, P^C) \tag{14}$$

Where $D(.)$ is a divergence measure, P^T is the probability distribution of the outcome in the treated group and P^C is the probability distribution of the outcome in the control group. The criterion is thus the gain in divergence following a split. The authors proposed three divergence metrics: Kullback, Euclidean and Chi-Squared, defined as:

$$\begin{aligned} KL(P:Q) &= \sum_{k=Left,Right} p_k log\frac{p_k}{q_k} \\ E(P:Q) &= \sum_{k=Left,Right} (p_k - q_k)^2 \\ \chi^2(P:Q) &= \sum_{k=Left,Right} \frac{(p_k - q_k)^2}{q_k} \end{aligned}$$

where subscript k indicates in which leaf we compute p and q that were defined in equation 12.

In their "Causal tree model", Athey and Imbens (2015b) propose the following criterion:

$$\Delta = \frac{1}{\#children} \sum_{k=1}^{\#children} \hat{\tau}_k^2 \tag{15}$$

It is easy to see that this last criterion boils down (up to a constant) to the Euclidean one from Rzepakowski and Jaroszewicz (2012) in the case of a binary outcome and randomized experiment.

It is worth noting the "honest" approach from Athey and Imbens (2015a). Instead of using the same data to generate the tree and estimate the uplift value inside the leaves, the authors randomly split the training set into two parts, one to generate the tree and the other to evaluate the uplift inside the leaves.

Though we have focused here on binary classification, some papers generalize the approach to other cases. Rzepakowski and Jaroszewicz (2012) proposed a tree method for the case of multi-treatment. It is also important to note that while Athey and Imbens (2015b) approach reproduces the traditional uplift criteria in the case of a binary outcome it is the only one that is generalized to regression settings.

4. Evaluation

In this section, we introduce how to properly evaluate Uplift Models and derive classical metrics using our "econometrics" notations. Uplift evaluation deserves an entire section because it differs drastically from the traditional machine learning model evaluation. In

machine learning, the standard is to use cross-validation: separate the data into a training and a testing datasets; learn on the training data, predict the target on the test data and compare to the ground truth. In uplift modeling, cross validation is still a valid idea but there is no more ground truth because we can never observe the effect of being treated and not treated on a person at the same time.

To illustrate the different metrics, we use simulated data from chapter 4.4 of (Kuusisto (2015)). The data contains 10000 individuals and is split into a treated dataset of 4997 individuals and a control dataset of 5003 individuals. The target is an indicator (0/1) for churn. There is a strong negative effect since 25 percent of the dataset was simulated to have a "Sleeping Dog" behavior[3]. We used 19 categorical features and a 80/20 cross validation split.

We implemented the three approaches described earlier. For the Two-Model Approach, we used two gradient boosted trees models, whereas the Class Transformation Approach was implemented through a random forest algorithm. To test the third method, we used the random forest and Causal Conditional Inference Forests from the uplift package[4]as well as a causal tree from the causalTree R package (found in Susan Athey's github repository[5]). We did not use grid search for any of these models to get the best hyper-parameters. Our goal in this section is solely to illustrate the different methods – we are not trying to establish that one method is better than another.

4.1. Traditional Uplift Metrics

Part of the Uplift literature stumbles with the problem that it is not possible to observe both the control and the treatment outcomes for an individual, which makes it difficult to find a loss measure for each observation. As a result, most of the Uplift literature must resort to aggregated measures such as uplift bins or uplift curves. For example, Ascarza (2016) uses bins of uplifts; Rzepakowski and Jaroszewicz (2012), Sołtys et al. (2015), Jaskowski and Jaroszewicz (2012), Jaroszewicz and Rzepakowski (2014) or Nassif et al. (2013) use what they call "uplift curve" and Radcliffe (2007) and Radcliffe (2008) Qini measure. A good review of these metrics can be found in Naranjo (2012). To visualize the metrics, we show the corresponding curves in the Appendix.

A common approach to evaluate an uplift model is to first predict uplift for both treated and control observations and compute the average prediction per decile in both groups. Then, the difference between those averages is taken for each decile. This difference thus gives an idea of the uplift gain per decile. For example, as we can see in Appendix Figure 1, the Two-Model Approach gives us an uplift of 0.3 in the first decile whereas the Class Transformation Approach gives us an uplift of 0.4. Though useful, Figure 1 makes it hard to effectively compare models. For instance, the second model seems to perform better in the first decile but not on the second.

To have a clearer idea, we can draw cumulative decile charts like in 2(a). The leftmost bar corresponds to the uplift in the first 10 percent, the following bar corresponds to the 20 first percent and so on. A well performing model features large values in the first quantiles

3. A Sleeping Dog is an individual who churns when exposed to the active treatment but does not churn if in the control group. Here, a quarter of the population would thus react negatively to a targeted action.
4. packagehttps://cran.r-project.org/web/packages/uplift/index.html
5. https://github.com/susanathey/causalTree.git

and decreasing values for larger ones. Finally, we can look at the cumulative gain chart: we calculate the uplift times the number of individuals taken into account for each bin,

$$\Big(\frac{Y^T}{N^T} - \frac{Y^C}{N^C}\Big)(N^T + N^C) \tag{16}$$

where Y^T (respectively Y^C) and N^T (respectively N^C) are the sum of the treated (respectively control) individual outcomes and the number of treated (respectively control) observations in the bin.

Figure 2(b) shows an example in the Two-Model approach case. This is useful to marketers because they can easily see if the treatment has a global positive or negative effect and if they can expect a better gain by targeting part of the population. We can thus choose the decile that maximizes the gain as the limit of the population to be targeted.

The problem with the charts so far is that they do not provide any metric and hence cannot be used to compare models accurately. We can nonetheless easily generalize the cumulative gain chart for each observation of the test set with the following parametric uplift curve defined for each t as:

$$f(t) = (\frac{Y_t^T}{N_t^T} - \frac{Y_t^C}{N_t^C})(N_t^T + N_t^C) \tag{17}$$

where the t subscript indicates that the quantity is calculated for the first t observations, sorted by inferred uplift value.

Figure 3(a) shows an example of such curves and a random line corresponding to the global effect of the treatment. The positive slope of this random line means that treating the whole population has an overall beneficial effect. Each point on a curve corresponds to the inferred uplift gain. The higher this value, the better the model. The continuity of the uplift curves makes it possible to calculate the area under the curve as a way to evaluate and compare the different uplift models. This measure is thus similar to the well-known AUC (Area under ROC curve) in a binary classification setup. In our case, the Two-Model approach seems to be consistently better than the other methods. Finally, the bell shape of the curves shows the strong positive and negative effect present in the dataset. In contrast, if these effects were absent, the curves would be closer to the random line.

In the literature, uplift curves are often defined by the difference between two lifts calculated on the treated and control datasets. This might not be ideal because there is no guarantee that the highest scoring examples in the treatment and control groups are similar. However, it is said to work well in practice and, in the case of randomized and balanced experiments, the two methods converge. We prefer our formula because it is closer to the Qini original definition. The Qini curve is introduced in Radcliffe (2007) as the parametric curve with the following equation:

$$g(t) = Y_t^T - \frac{Y_t^C N_t^T}{N_t^C} \tag{18}$$

The authors define the Qini coefficient to be the area under the Qini curve. Examples of such curves are given in 3(b). There is an obvious parallel with the uplift curve since

$$f(t) = \frac{g(t)(N_t^T + N_t^C)}{N_t^T} \tag{19}$$

In balanced cases, the curves will almost be proportional to a factor of two, as we can see in figure 3.

Radcliffe (2007) also introduced a similar formula for regression problems. The formula is the same except that the count of positive examples in treated and control groups is replaced by the continuous value of the target.

4.2. Metric Based on Y^*

In section 3.2, we introduced Y_i^*, a transformation of the target variable. It is thus natural to wonder if a loss function (for example the Mean Squared Error (MSE)) between our estimator ($\hat{\tau}$) and the real Y_i^* can be used in a cross validation scheme. In this section we discuss the pertinence of using:

$$MSE(Y_i^*, \hat{\tau}) = \sum_i^n \frac{1}{n}(Y_i^* - \hat{\tau}_i)^2 \tag{20}$$

as an approximation for

$$MSE(\tau_i, \hat{\tau}_i) = \sum_i^n \frac{1}{n}(\tau_i - \hat{\tau}_i)^2 \tag{21}$$

Though equation 21 can be calculated for simulated data where we know the true causal effect, τ_i, it is impossible to derive from observational data, as noted in Athey and Imbens (2015a). Because the MSE is impossible to calculate, the authors introduce an estimator that can only be used in a decision tree setting. The estimation approach is mandatory for uplift evaluation. The advantage of using our metric is that it does not depend on the chosen machine learning model. Note that in the previous subsection, the curves are also based on a local estimation of τ.

The goal is to show that we can substitute Y_i^* for the unobserved τ_i in the MSE metrics to evaluate the performance of our model. Note that by adding and subsequently subtracting Y_i^* in equation 21, we can write:

$$\begin{aligned} MSE &= \sum_i^n \frac{1}{n}(\tau_i - Y_i^* + Y_i^* - \hat{\tau}_i)^2 \\ &= \sum_i^n \frac{1}{n}[(\tau_i - Y_i^*)^2 + 2((\tau_i - Y_i^*)(Y_i^* - \hat{\tau}_i)) + (Y_i^* - \hat{\tau}_i)^2] \\ &\longrightarrow_p E[(\tau_i - Y_i^*)^2] + 2E[(\tau_i - Y_i^*)(Y_i^* - \hat{\tau}_i)] + E[(Y_i^* - \hat{\tau}_i)^2] \end{aligned} \tag{22}$$

To minimize the MSE, we can ignore the first term in 22 because it does not depend on $\hat{\tau}_i$. In the second term, notice how $\hat{\tau}_i \perp\!\!\!\perp \tau_i | X_i$ and $\hat{\tau}_i \perp\!\!\!\perp Y_i^* | X_i$ because the estimator is a

function of X_i. As a result:

$$\begin{aligned}
E[(\tau_i - Y_i^*)(Y_i^* - \hat{\tau}_i)] &= E[\tau_i Y_i^* - \tau_i\hat{\tau}_i - (Y_i^*)^2 + Y_i^*\hat{\tau}_i] \\
&= E[E[\tau_i Y_i^* - \tau_i\hat{\tau}_i - (Y_i^*)^2 + Y_i^*\hat{\tau}_i|X_i]] \\
&= E[E[\tau_i Y_i^*|X_i]] - E[E[\tau_i\hat{\tau}_i|X_i]] - E[E[(Y_i^*)^2|X_i]] + E[E[Y_i^*\hat{\tau}_i|X_i]] \\
&= E[E[\tau_i Y_i^*|X_i]] - \underbrace{E[\tau(X_i)E[\hat{\tau}_i|X_i]]}_{\tau_i \perp\!\!\!\perp \hat{\tau}_i|X_i} - E[E[(Y_i^*)^2|X_i]] \\
&+ \underbrace{E[\tau(X_i)E[\hat{\tau}_i|X_i]]}_{\hat{\tau}_i \perp\!\!\!\perp Y_i^*|X_i \text{ and } E[Y_i^*|X_i]=\tau(X_i)} \\
&= E[E[\tau_i Y_i^*|X_i]] - E[E[(Y_i^*)^2|X_i]] \qquad (23)
\end{aligned}$$

As 23 does not depend on our estimator, we see that in the limit, minimizing $MSE(Y_i^*, \hat{\tau})$ amounts to minimizing $MSE(\tau_i, \hat{\tau}_i)$, which is what we wanted to show.

5. Conclusion

This paper presents an overview of the uplift literature using a common causal inference framework. The uplift literature proposes three different approaches to estimate causal effects. The first one is the Two-Model method consisting in training two separated models: one on the treatment group and one on the control group. The uplift of a test observation is then computed as the difference between its prediction in the two models. Although the Two-Model is easy to implement, the approach can be surpassed by the Class-Transformation approach which aims at modeling a transformed outcome variable whose conditional expectation is equal to the true uplift. Traditionally, this second method has been relying on the assumption of complete treatment randomization. However, a generalization to the unbalanced case is straightforward. The third approach amounts to modifying existing Machine Learning models to fit the uplift framework. In this paper, we restricted our attention to tree-based methods and presented the different split criteria from the literature.

As for model evaluation, we saw that in the absence of the true uplift, no loss can easily be computed to evaluate the performance of a model. One approach consists in sorting treated and untreated test observations in ascending order of predicted uplift, separately. Both groups are then binned into deciles and the model performance is evaluated through the pairwise difference in the uplift average per decile. A variation to the pairwise decile comparison is to look at the cumulative difference throughout deciles. These two techniques are useful to gain general sense of how a model is performing, but they remain visual methods. A more precise evaluation method, which is actually a generalization of the cumulative decile comparison one, is the uplift curve. Test observations are sorted in ascending order of predicted uplift. The uplift curve is defined as a parametric function of the number of observations selected that returns the difference in the average predicted uplift between the treatment and control groups. The uplift curve typically features a bell shape and the area under this curve serves as a performance metrics (just like a traditional AUC). Finally, this paper contributes to the literature by proposing a method that uses a transformed outcome Y_i^* directly in an MSE equation. We indeed prove that, in the limit, minimizing the MSE

formula that uses Y_i^* in place of the true treatment effect also minimizes the MSE equation that uses the true treatment effect.

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Eva Ascarza. Retention futility: Targeting high risk customers might be ineffective. *Available at SSRN*, 2016.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *arXiv preprint arXiv:1504.01132*, 2015a.

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050:5, 2015b.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016.

Leo Guelman, Montserrat Guillen, Ana M Pérez-Marín, et al. Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study. Technical report, 2014.

Leo Guelman, Montserrat Guillén, and Ana M Pérez-Marín. Uplift random forests. *Cybernetics and Systems*, 46(3-4):230–248, 2015.

Behram Hansotia and Brad Rukstales. Incremental value modeling. *Journal of Interactive Marketing*, 16(3):35, 2002.

Szymon Jaroszewicz and Piotr Rzepakowski. Uplift modeling with survival data. 2014.

Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In *ICML Workshop on Clinical Data Analysis*, 2012.

Finn C Kuusisto. *Machine Learning for Medical Decision Support and Individualized Treatment Assignment*. PhD thesis, University of Porto, 2015.

Lily Yi-Ting Lai. *Influential marketing: A new direct marketing strategy addressing the existence of voluntary buyers*. PhD thesis, Citeseer, 2006.

Victor SY Lo. The true lift model: a novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter*, 4(2):78–86, 2002.

Oscar Mesalles Naranjo. Testing a new metric for uplift models. 2012.

Houssam Nassif, Finn Kuusisto, Elizabeth S Burnside, and Jude W Shavlik. Uplift modeling with roc: An srl case study. In *ILP (Late Breaking Papers)*, pages 40–45, 2013.

Nicholas J Radcliffe and Patrick D Surry. Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, 2011.

NJ Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Market J Direct Market Assoc Anal Council*, 1:14–21, 2007.

NJ Radcliffe. Hillstroms minethatdata email analytics challenge: An approach using uplift modelling. *Stochastic Solutions Limited*, 1:1–19, 2008.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012.

Atef Shaar, Talel Abdessalem, and Olivier Segard. Pessimistic uplift modeling. *arXiv preprint arXiv:1603.09738*, 2016.

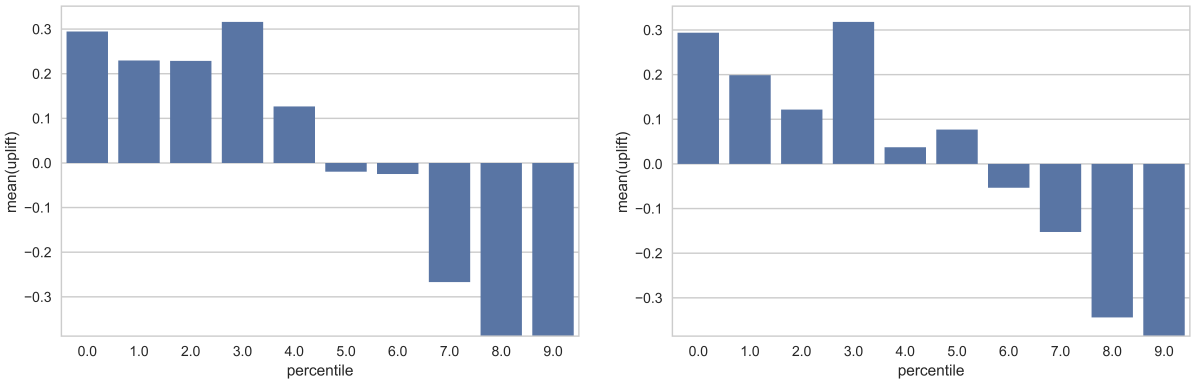
Michał Sołtys, Szymon Jaroszewicz, and Piotr Rzepakowski. Ensemble methods for uplift modeling. *Data mining and knowledge discovery*, 29(6):1531–1559, 2015.

Xiaogang Su, Joseph Kang, Juanjuan Fan, Richard A Levine, and Xin Yan. Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research*, 13(Oct):2955–2994, 2012.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*, 2015.

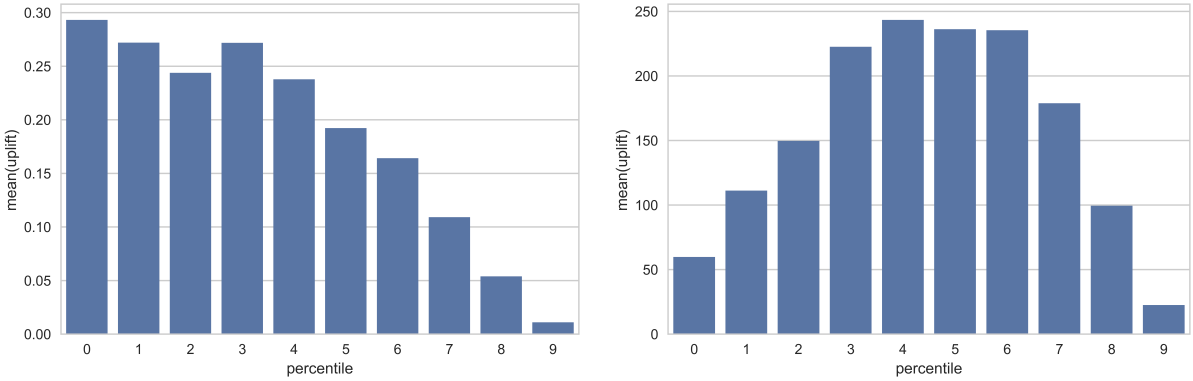
Lukasz Zaniewicz and Szymon Jaroszewicz. Support vector machines for uplift modeling. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 131–138. IEEE, 2013.

Appendix A. Charts



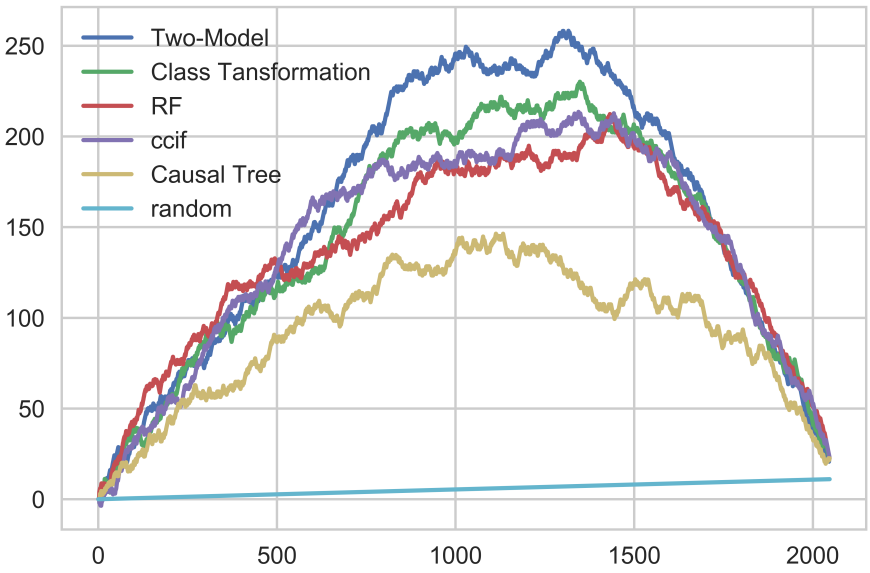
(*a*) Two-Model approach (*b*) Class Transformation approach

Figure 1: Uplift decile charts for the Two-Model approach (a) and the Class Transformation approach (b).

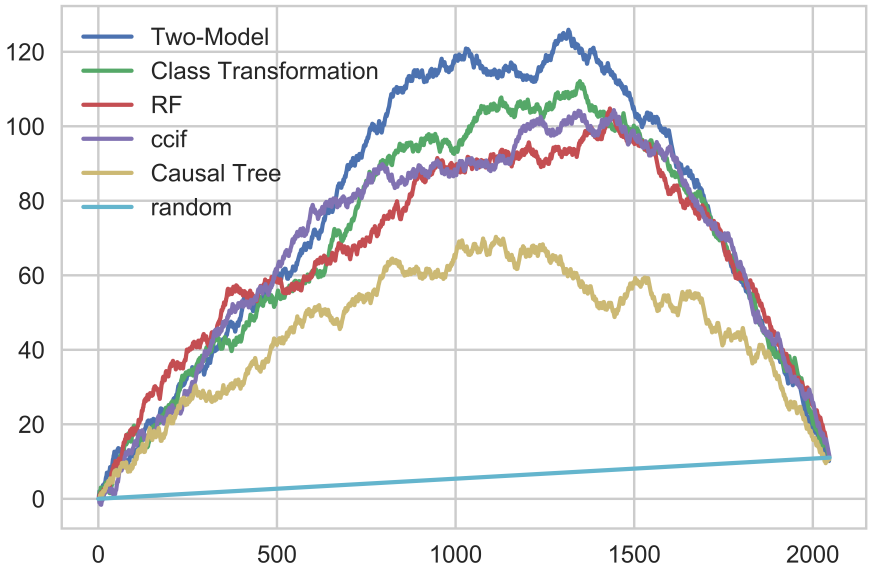


(*a*) Cumulative uplift (*b*) Cumulative gain

Figure 2: Cumulative uplift and gain for the Two-Model approach.



(a) Uplift curves



(b) Qini curves

Figure 3: Uplift curves and Qini curves applied to several uplift approaches.