Adapting Neural Networks for Uplift Models

Mouloud Belbahri [1 2]   **Olivier Gandouet** [1]   **Ghaith Kazma** [1]

Abstract

Uplift is a particular case of individual treatment effect modeling. Such models deal with cause-and-effect inference for a specific factor, such as a marketing intervention. In practice, these models are built on customer data who purchased products or services to improve product marketing. Uplift is estimated using either i) *conditional mean regression* or ii) *transformed outcome regression*. Most existing approaches are adaptations of classification and regression trees for the uplift case. However, in practice, these conventional approaches are prone to overfitting. Here we propose a new method using neural networks. This representation allows to jointly optimize the difference in conditional means and the transformed outcome losses. As a consequence, the model not only estimates the uplift, but also ensures consistency in predicting the outcome. We focus on fully randomized experiments, which is the case of our data. We show our proposed method improves the state-of-the-art on synthetic and real data.

1. Introduction

Uplift is a particular case of causal inference and follows the potential outcomes framework (Rubin, 1974; Holland, 1986). Uplift models are developed for randomized controlled trials (RCTs) settings where both the treatment and outcome are binary random variables. Uplift refers to modeling behavioral change that results directly from a specific treatment, such as a marketing intervention.

Mainly known in the marketing research literature, uplift models (Radcliffe & Surry, 1999; Hansotia & Rukstales, 2001; Lo, 2002; Radcliffe, 2007) provide a solution to the

[1]Analytics and Modeling - Advanced Projects, TD Insurance, Montreal, Canada [2]Department of Mathematics and Statistics, University of Montreal, Montreal, Canada. Correspondence to: Mouloud Belbahri <mouloud.belbahri@tdassurance.com>, Olivier Gandouet <olivier.gandouet@tdassurance.com>.

problem of isolating the marketing effect. Instead of modeling the positive response probability, uplift attempts to model the difference between conditional probabilities in the treatment and control groups. Uplift definition is equivalent to the one of individual treatment effect (ITE). However, uplift models performance metrics diverge from those used for ITE estimation (Radcliffe & Surry, 2011).

In practice, the predicted ITE is used to rank customers in the order that they should be targeted to maximize a future marketing campaign success. To do so, an uplift model is built using a historical list of clients to whom a business sold products and services. This is often achieved in RCTs settings. First, a pilot study is conducted. In this case, the treatment assignment (e.g. a courtesy call) is controlled, and thus the distributions of treatment and control groups are known. The model is then built on the collected data.

The customers can be segmented along two dimensions in function of the potential outcomes and the treatment, giving rise to the following groups (Kane et al., 2014) : a *persuadable* client is an individual who purchases the product only if he receives a call; a *sure thing* client is an individual who purchases the product whether he receives the call or not; a *lost cause* client is an individual who will not purchase the product regardless of whether he receives the call or not; a *do-no-disturb* client is an individual who will not purchase the product only if he receives a call.

In general, the interesting customers from a business point of view are the *persuadable* and the *do-not-disturb* clients. The persuadable clients provide incremental sales whereas the do-not-disturb individuals should be avoided because the marketing campaign would have a negative effect on them. Uplift models mainly aim to detect these groups of customers. Most existing approaches are adaptations of classification and regression trees (Radcliffe & Surry, 2011; Jaskowski & Jaroszewicz, 2012; Guelman et al., 2012). Based on several experiments on real data, and although the literature suggests that tree-based methods are state-of-the-art tools for uplift (Sołtys et al., 2015), predicting which group a customer belongs to still lacks satisfactory solutions. One problem is that uplift is estimated using either i) the difference in conditional means (i.e. an indirect estimation defined as the difference between treated and control positive response proportions), or ii) the transformed outcome

mean (Athey & Imbens, 2015). In both cases, there is a harmful compromise. Either the model focuses on learning the underlying binary classification task, or it ignores it completely. This is particularly problematic for uplift trees, as the split criteria is function of the two treatment effects. Splits are likely to be placed next to extreme values because outliers of any treatment can influence the choice of a split point. In addition, successive splits tend to group together similar extreme values, introducing more variance in the prediction of uplift (Zhao et al., 2017). As a result, the published models over-fit the training data.

In this work, we define a unified framework with the goal of improving uplift models predictions on unseen data. We take advantage of both the conditional means and the transformed outcome in order to find a representation (using neural networks) of the data that helps in reducing generalization error. The main contribution of this article is to give what is, to the best of our knowledge, the first framework which allows a balance between direct estimation of the uplift and coherent prediction of the binary outcome. Our architecture is quite simple which makes it flexible. This allows us to propose an alternative to using the transformed outcome for predicting uplift. Furthermore, it opens a door to connections with the estimation of ITE in observational data settings. Our proposed method is validated on both synthetic and a realworld data-set made available to us by a Canadian insurance company. The experimental results demonstrate its advantages over state-of-the-art uplift modeling.

2. Problem Setup

To formalize the problem, let $Y \in \{0, 1\}$ be a binary response variable, $\mathbf{X} = (X_1, \ldots, X_p)$ a vector of covariates, and $T \in \{0, 1\}$ the treatment indicator variable. The binary variable T indicates if a customer is exposed to treatment ($T = 1$) or control ($T = 0$). The observed outcome Y can be described as

$$Y = TY^{(1)} + (1 - T)Y^{(0)},$$

where $Y^{(0)}$ and $Y^{(1)}$ are the potential outcomes under control and treatment respectively. Suppose n independent customer observations $\{y_i, \mathbf{x}_i, t_i\}_{i=1}^n$ are at our disposal. Although each customer i is associated with two potential outcomes, only one of them can be realized as the observed outcome y_i. For RCTs settings, potential outcomes $(Y^{(0)}, Y^{(1)})$ are independent ($\perp\!\!\!\perp$) of the treatment assignment indicator T conditioned on all pre-treatment characteristics $(X_1, \ldots, X_p)$, i.e.,

$$\left(Y^{(0)}, Y^{(1)}\right) \perp\!\!\!\perp T \mid \mathbf{X}. \tag{1}$$

This allows the unbiased estimation of the average treatment effect (ATE). Moreover, the *propensity score* $e(\mathbf{X}) = \Pr(T = 1 \mid \mathbf{X})$ is constant with respect to $\mathbf{X}$. In our data, $e(\mathbf{X}) = 1/2$. An uplift model estimates the conditional average treatment effect (CATE or ITE) in various sub-populations as a function of the possible values of the covariates $\mathbf{X}$,

$$
\begin{aligned}
u(\mathbf{X}) &= \mathbb{E}(Y^{(1)} \mid \mathbf{X}) - \mathbb{E}(Y^{(0)} \mid \mathbf{X}) \\
&= \mathbb{E}(Y \mid T = 1, \mathbf{X}) - \mathbb{E}(Y \mid T = 0, \mathbf{X}) \\
&= \Pr(Y = 1 \mid T = 1, \mathbf{X}) - \Pr(Y = 1 \mid T = 0, \mathbf{X}) \\
&= m_1(\mathbf{X}) - m_0(\mathbf{X}). \tag{2}
\end{aligned}
$$

Uplift is naturally described as the difference between the two conditional means $m_1(\mathbf{X})$ and $m_0(\mathbf{X})$. Our general framework for learning uplift is depicted in Figure 1. Before formally introducing it, we must first define some quantities of interest that will be used throughout the rest of the paper. The first interesting quantity is the transformed outcome variable Z, which was proposed by (Athey & Imbens, 2015) and is defined as

$$Z = \frac{TY}{e(\mathbf{X})} - \frac{(1 - T)Y}{1 - e(\mathbf{X})}. \tag{3}$$

This transformation has the key property that, under the complete randomization or unconfoundness assumption (Rosenbaum & Rubin, 1983), its expectation conditional on $\mathbf{X}$ is equal to $u(\mathbf{X})$. Therefore, the transformed outcome is a consistent unbiased estimate of uplift. In our case, with $e(\mathbf{X}) = 1/2$, it simplifies to

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

In the uplift context, we can have an interesting interpretation of $Z = 2$ and $Z = -2$. Take the first case. This happens when a client is called ($T = 1$) and then buys the insurance policy ($Y = 1$). Thus, for its counterfactual (i.e., $T = 0$), there are two possibilities. Either the client would have bought the policy ($Y = 1$) and in this case would belong to the *sure things* group, or else the client would not have bought the policy ($Y = 0$) in which case would in fact be a *persuadable*. Thus, in the worst case scenario, individuals with $Z = 2$ are *sure things*. The same logic applies for $Z = -2$ where a client could either be a *sure thing* or a *do-not-disturb*. For *lost cause* clients, by definition $Z = 0$. This interpretation does not cover all scenarios for $Z = 0$ but gives insights about ordering customers with $Y = 1$. Combined with the fact that Z is an unbiased estimate of

uplift, this interpretation motivates us to start building our framework with a constrained version of the transformed outcome regression.

3. General framework

We propose a *Siamese* neural network (Bromley et al., 1994) architecture called SMITE [1] which works in tandem on two different input vectors $(\mathbf{X}, \mathbf{1})$ and $(\mathbf{X}, \mathbf{0})$, and outputs the prediction of the conditional mean $m_T(\mathbf{X}) = Tm_1(\mathbf{X}) + (1 - T)m_0(\mathbf{X})$ as well as the one of the uplift $u(\mathbf{X})$. Two input vectors are fed to the two identical neural network sub-components (i.e., with shared weights $\boldsymbol{\theta}$). The inputs contain the covariates vector $\mathbf{X}$ and, for the top network sub-component, the treatment variable fixed to 1. The treatment variable is fixed to 0 for the bottom network sub-component as shown in Figure 1. The sub-components output the predicted conditional means for treated (μ_1) and for control (μ_0). The predicted conditional mean μ_T is based on the actual treatment $T \in 0, 1$, i.e., $\mu_T = T\mu_1 + (1 - T)\mu_0$. Then μ_T is compared to the outcome Y based on binary cross-entropy (BCE) loss $\mathcal{L}$. Moreover, the difference of the predicted output values μ_1 and μ_0 gives direct prediction for $u(\mathbf{X})$. This prediction is compared to the transformed outcome Z based on the mean squared error (MSE) loss $\mathcal{J}$.

Note that an alternative to using Z is easily implemented in the same framework and is explained in Section 3.2. The intuition behind the SMITE architecture is based on two factors.

1. Classification models with binary outcomes are relatively easy to develop[2]. However, these models fail to properly capture the ITE because they are not designed for this task.

2. Methods that aim to directly predict the ITE fail to generalize because they focus only on ITE and do not take into account the underlying binary classification task.

We developed the SMITE architecture with the above observations in mind. Essentially, back-propagating both losses gradients through the shared representation C_H consists in fitting a *trade-off* between the conditional means and the transformed outcome regressions using a single model. With little fine-tuning, this representation takes advantage of both for improving uplift prediction.

---

[1] Siamese Model for Individual Treatment Effect

[2] Depending on the complexity of the data, any supervised learning algorithm used properly tends to give satisfactory results.
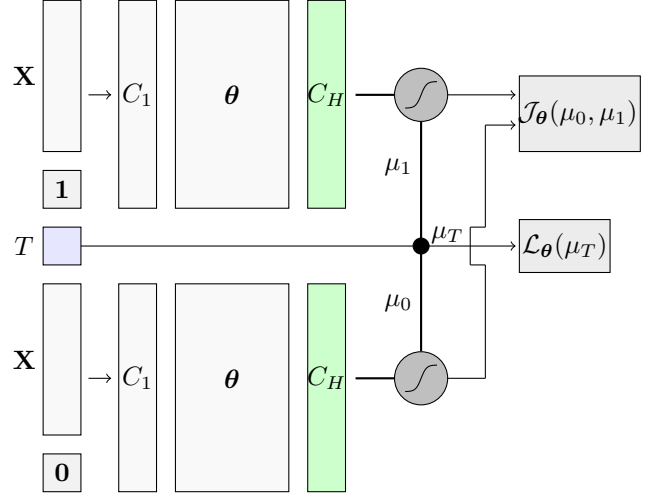

Figure 1. SMITE: Neural network architecture for uplift estimation.

3.1. The optimization problem

We want to train a model with parameters $\boldsymbol{\theta}$, by solving a constrained optimization problem. Formally, using the transformed outcome Z, we want to solve

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min} \quad \mathcal{J}\Big(\mu_0(\boldsymbol{\theta}), \mu_1(\boldsymbol{\theta}), Z\Big) \tag{4}$$

subject to,

$$\mathrm{d}\Big(\Pr(Y = 1 \mid T, \mathbf{X}), \mu_T(\boldsymbol{\theta})\Big) < \epsilon, \tag{5}$$

where $\mathrm{d}(.)$ is a distance and $\epsilon > 0$. $\mu_{\cdot}(.)$ are functions of the inputs $\mathbf{X}$ and should be denoted $\mu_{\cdot}(\mathbf{X})$. To simplify the notation, we prefer to omit this specification throughout the rest of the paper. The objective function $\mathcal{J}(.)$ is what we refer to as the *direct* uplift loss function. The rationale behind this constrained formulation is that we want a model capable of direct prediction of the uplift $u(\mathbf{X})$, i.e., $\mu_1(\hat{\boldsymbol{\theta}}) - \mu_0(\hat{\boldsymbol{\theta}})$, and coherent estimation of $m_T(\mathbf{X})$, i.e., $\mu_T(\hat{\boldsymbol{\theta}})$.

The direct uplift loss can be defined as a distance between the predicted uplift and the transformed outcome. A natural choice is the mean squared error (MSE),

$$\mathcal{J}(\boldsymbol{\theta}, Z) = \frac{1}{n} \sum_{i=1}^{n} \Big(z_i - \{\mu_{1i}(\boldsymbol{\theta}) - \mu_{0i}(\boldsymbol{\theta})\} \Big)^2. \tag{6}$$

A convenient choice of distance $\mathrm{d}(.)$ in our case is binary cross-entropy (BCE). Formally, we define $\mathcal{L}(\boldsymbol{\theta}, Y)$ with respect to the parameters $\boldsymbol{\theta}$ as

$$\mathcal{L}(\boldsymbol{\theta}, Y) = -\sum_{i=1}^{n} \Big(y_i \log\{\mu_{t_i}(\boldsymbol{\theta})\} + (1-y_i)\log\{1-\mu_{t_i}(\boldsymbol{\theta})\} \Big)$$

where $\mu_{t_i}(\boldsymbol{\theta}) = t_i \mu_{1i}(\boldsymbol{\theta}) + (1-t_i)\mu_{0i}(\boldsymbol{\theta})$ is the model's output of observation i with $T = t_i$. It is helpful to rewrite the full constrained problem (5) in his Lagrangian form

$$\hat{\boldsymbol{\theta}}_\lambda = \operatorname*{argmin}_{\boldsymbol{\theta}} \ \mathcal{J}(\boldsymbol{\theta}, Z) + \lambda \mathcal{L}(\boldsymbol{\theta}, Y)$$

where $\lambda > 0$ can be seen as a regularization constant. However, in the SMITE framework, we use the transformation $\alpha = 1/(1+\lambda)$. This transformation allows to interpret $\alpha \in [0,1]$ as a hyper-parameter which controls the trade-off between the two losses in the optimization problem. Hence,

$$\hat{\boldsymbol{\theta}}_\alpha = \operatorname*{argmin}_{\boldsymbol{\theta}} (1-\alpha)\mathcal{J}(\boldsymbol{\theta}, Z) + \alpha \mathcal{L}(\boldsymbol{\theta}, Y). \quad (7)$$

Equation (7) is the one of greatest interest. It encompasses exactly what motivates the SMITE framework, to jointly optimize both loss functions.

In practice, we train the SMITE models by minimizing (7) using the stochastic gradient descent (SGD) optimizer with fixed learning rate. In our experiments, the *trade-off* constant α is fine-tuned using cross-validation (see Figure 3). Similar experiments were performed for the learning rate ξ.

3.2. An alternative to the Z transform

As mentioned before, the flexibility of SMITE allows us to offer alternatives to the loss functions defined in the general framework. For example, instead of using the direct uplift loss (6), one can think of the L_1-norm loss. Furthermore, we noticed that the learning curve associated with (6) could have erratic behavior in early stages of training (see Figure 2). This led us to define an alternative to the Z transform which improves the performance of our uplift models.

Our intuition started with the following observation :

"Suppose a treatment has a significant positive (resp. negative) effect on a sub-sample of customers. Therefore, within the sample of customers that had a positive (resp. negative) response, we expect a higher proportion of treated customers."

More precisely, we denote the proportion of treated customers among those that had positive outcomes by $\pi_1(\mathbf{X})$.

When $e(\mathbf{X}) = 1/2$, which is the case of our data[3], it is easy to show that $\pi_1(\mathbf{X})$ can be written as a function of the conditional means,

$$\pi_1(\mathbf{X}) = \frac{m_1(\mathbf{X})}{m_0(\mathbf{X}) + m_1(\mathbf{X})}.$$

Similarly, the proportion of treated customers among those that had negative outcomes is

$$\pi_0(\mathbf{X}) = \frac{(1 - m_1(\mathbf{X}))}{(1 - m_0(\mathbf{X})) + (1 - m_1(\mathbf{X}))}.$$

From these definitions, we propose an alternative to the direct uplift loss. The *indirect* uplift loss has the following form

$$\mathcal{I}(\boldsymbol{\theta}, T) = -\sum_{i=1}^{n} \Big(t_i \log\{\Pi_{y_i}(\boldsymbol{\theta})\} + (1-t_i)\log\{1-\Pi_{y_i}(\boldsymbol{\theta})\} \Big),$$

where, $\Pi_{y_i}(\boldsymbol{\theta}) = y_i \Pi_{1i}(\boldsymbol{\theta}) + (1-y_i)\Pi_{0i}(\boldsymbol{\theta})$, with

$$\Pi_1(\boldsymbol{\theta}) = \frac{\mu_1(\boldsymbol{\theta})}{\mu_0(\boldsymbol{\theta}) + \mu_1(\boldsymbol{\theta})},$$

$$\Pi_0(\boldsymbol{\theta}) = \frac{(1 - \mu_1(\boldsymbol{\theta}))}{(1 - \mu_0(\boldsymbol{\theta})) + (1 - \mu_1(\boldsymbol{\theta}))}.$$

The indirect uplift loss is simply the BCE distance between a function of the predicted conditional means $\Pi(\hat{\boldsymbol{\theta}})$ and the actual treatment variable T. The associated learning curves are smoother than the ones of the direct uplift loss (see Figure 2). In Section 6, we will refer to the SMITE using $\mathcal{J}$ as SMITE Transformed Outcome (TO) and the SMITE using $\mathcal{I}$ as SMITE Indirect Estimation (IE). In this case, the optimization problem becomes

$$\hat{\boldsymbol{\theta}}_\alpha = \operatorname*{argmin}_{\boldsymbol{\theta}} (1-\alpha)\mathcal{I}(\boldsymbol{\theta}, T) + \alpha \mathcal{L}(\boldsymbol{\theta}, Y). \quad (8)$$

4. Related Work

The intuitive approach to model uplift is to build two classification models (Hansotia & Rukstales, 2001). This consists of fitting two separated conditional probability models: one for the treated individuals, and another for the untreated individuals. The uplift is estimated as the difference between these two conditional probability models. The asset of this

---

[3]Although the equations are simplified based on this assumption, the development can easily be generalized for any constant $e(\mathbf{X})$. Moreover, one can use an over-sampling procedure to recover the $e(\mathbf{X}) = 1/2$ case.
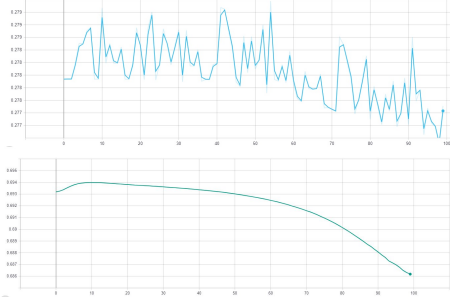
Figure 2. Uplift loss functions learning curves: direct uplift loss $\mathcal{J}$ (top) and indirect uplift loss $\mathcal{I}$ (bottom).

technique is its simplicity. However, it does not perform well in practice (Radcliffe & Surry, 2011). Both models focus on predicting only a one-class probability and for each model, information about the other treatment is never provided to the learning algorithm. Using a unique classification model corrects this drawback. Some traditional uplift estimation methods use the treatment assignment as a feature, adding explicit interaction terms between each covariate and the treatment indicator, and then train classification models to estimate the uplift (Lo, 2002). The parameters of the interaction terms measure the additional effect of each covariate because of the treatment. Several k-nearest neighbours (Cover & Hart, 1967) based methods are also adopted for uplift estimation (Crump et al., 2008; Alemi et al., 2009; Su et al., 2012). The main idea is to estimate the uplift for an observation based on its neighbourhood containing at least one treated and one control observations. However, these methods quickly become computationally expensive for large datasets. Furthermore, adaptive tree-based methods such as random forests outperform the nearest neighbours approach, especially in the presence of irrelevant covariates (Wager & Athey, 2018). Therefore, most active research in uplift modeling is in the direction of classification and regression trees (Breiman et al., 1984). State-of-the-art proposed methods view the forests as an adaptive neighborhood metric, and estimate the treatment effect at the leaf node (Su et al., 2009; Chipman et al., 2010; Guelman et al., 2012). In (Radcliffe & Surry, 1999; 2011) or (Hansotia & Rukstales, 2001; 2002), modified split criteria that suited the uplift purpose were studied. The criteria used for choosing each split during the growth of the uplift trees is based on maximization of the difference in uplifts between the two child nodes. However, in practice, these approaches are prone to overfitting. The splitting criteria introduced are based solely on maximizing differences in uplifts. Therefore, splits are likely to be placed next to extreme values because outliers of any treatment can influence the choice of a split point. In addition, successive splits tend to group together similar extreme values, introducing more variance in the prediction of uplift (Zhao et al., 2017).

Recently, representation learning approaches have been proposed for estimating individual treatment effect based on observational data. An approach in the literature emphasizes learning a covariate representation that has a balanced distribution across treatment and outcome (Johansson et al., 2016). Some similar methods try to minimize the distribution difference between treated and control groups in the embedding space (Shalit et al., 2017; Yao et al., 2018). Another approach is based on propensity score matching (Rosenbaum & Rubin, 1983) with neural networks (Alaa et al., 2017; Schwab et al., 2018) and (Shi et al., 2019) use regularized neural networks to estimate the average treatment effect (ATE).

5. Uplift performance metrics

The *ground truth* is not observed in the case of individual treatment effect estimation. In order to evaluate uplift models, we refer to two known metrics in the uplift literature, the *Qini coefficient* and the *Kendall's uplift rank correlation*. The computation of both statistics requires the construction of the *Qini curve* (Radcliffe, 2007). These statistics are used as goodness-of-fit measures for uplift models.

For a given model, let $\hat{u}_{(1)} \geq \hat{u}_{(2)} \geq ... \geq \hat{u}_{(n)}$ be the sorted predicted uplifts. Let $\phi \in [0, 1]$ be a given proportion and let $N_\phi = \{i : \hat{u}_i \geq \hat{u}_{(\lceil \phi n \rceil)}\} \subset \{1, \ldots, n\}$ be the subset of observations with the $\phi n \times 100\%$ highest predicted uplifts $\hat{u}_i$ (here $\lceil s \rceil$ denotes the smallest integer larger or equal to $s \in \mathbb{R}$). The *Qini curve* is defined as a function f of the fraction of population targeted ϕ, where

$$f(\phi) = \frac{1}{n_t} \left(\sum_{i \in N_\phi} y_i t_i - \sum_{i \in N_\phi} y_i(1 - t_i) / \sum_{i \in N_\phi} (1 - t_i) \right),$$

where $n_t = \sum_{i=1}^{n} t_i$ is the number of treated customers, with $f(0) = 0$ and $f(1)$ is the average treatment effect (ATE). In other words, $f(\phi)$ represents the incremental number of positive responses for a fraction ϕ or targeted customers relative to the total number of targeted customers in the sample. The Qini helps visualize the performance of an uplift model, a bit like the ROC curve of a binary classification model. Following the same logic, a straight line between the points $(0, 0)$ and $(1, f(1))$ defines a benchmark to compare the performance of the model to a strategy that would randomly target individuals. This means that if a proportion ϕ of the population is targeted, we expect to observe an uplift of $\phi f(1)$. The *Qini coefficient* is defined as

$$q = \int_0^1 Q(\phi) \, d\phi = \int_0^1 \{f(\phi) - \phi \, f(1)\} \, d\phi,$$

where $Q(\phi) = f(\phi) - \phi \, f(1)$. This area can be numerically approximated using a Riemann method such as the trapezoid

rule formula: the domain of $\phi \in [0, 1]$ is partitioned into K bins, or $K + 1$ grid points $0 = \phi_1 < \phi_2 < ... < \phi_{K+1} = 1$, to approximate q by

$$\hat{q} = \frac{1}{2} \sum_{k=1}^{K} (\phi_{k+1} - \phi_k)\{Q(\phi_{k+1}) + Q(\phi_k)\}. \quad (9)$$

It is important to note, unlike the area under the ROC curve, $\hat{q}$ could be negative. This would simply mean that a strategy following the model in question does worse than random targeting. The second goodness-of-fit measure of an uplift model is defined by the similarity between the theoretical uplift percentiles (given by the model's predictions) and empirical percentiles (observed in the data) based on the model. This can be approximated by the *Kendall's uplift rank correlation* (Belbahri et al., 2019) defined as

$$\hat{\rho} = \frac{2}{K(K-1)} \sum_{i<j} \text{sign}(\bar{\hat{u}}_i - \bar{\hat{u}}_j) \, \text{sign}(\bar{u}_i - \bar{u}_j), \quad (10)$$

where $\bar{\hat{u}}_k$ is the average predicted uplift in bin k, $k \in 1, ..., K$, and $\bar{u}_k$ is the observed uplift in the same bin.

6. Experiments

We conduct a simulation study to examine the performance SMITE uplift models. We compare its performance with existing uplift methods using three types of datasets: i) a real-world dataset from a marketing campaign; ii) boot-strapped versions of the real-world dataset with synthetic outcomes generated from tree-based models, and iii) synthetic datasets generated from parametric models.

Ensemble methods are amongst the most powerful for the uplift case. A method that has proven to be very effective in estimating ITE is one based on generalized random forests (Athey et al., 2019). This method uses honest estimation, i.e., it does not use the same information for the partition of the covariates space and for the estimation of the treatment effects. This has the effect of reducing overfitting. Another candidate that we consider in our experiments is also based on random forests that were designed for uplift (Guelman et al., 2012; Jaskowski & Jaroszewicz, 2012; Sołtys et al., 2015). For this method, we consider two different split criteria, one based on Kullback-Leibler divergence and one based on Euclidean distance.

6.1. Fine tuning SMITE

The backbone of the SMITE architecture is a 6-layers feed forward MLP with the following input neurons sizes: $\{200, 200, 300, 100, 50, 10\}$. The first two layers are linearly connected while the rest are connected through a leaky ReLu activation function. Both the depth and the width of this network were explored using a random grid search
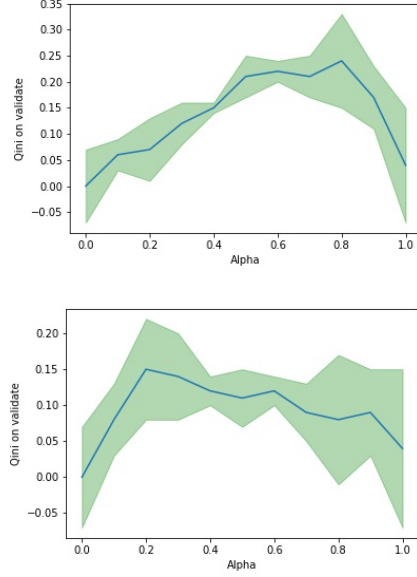

Figure 3. Fine-tuning of the trade-off constant α on the Canadian insurer dataset. The top curve corresponds to the (IE) loss and the bottom one to the (TO) loss.

on data generated with the framework described in Section 6.4.2. Two hyper-parameters are fine-tuned for each set of experiments: the learning rate ξ and the trade-off constant α. To find an adequate candidate for α, we follow the following protocol: the dataset D is split into two subsets (D_T, D_H) containing respectively 70% and 30% of D. D_H remains unseen during any of our fine-tuning or model selection process. We then repeat k-times a random separation of D_T into two datasets denoted $(D_{\text{train}}^i, D_{\text{valid}}^i)$ with sizes 60% and 40% of D_T.

At a fixed learning rate of $\xi = 0.03$, we fit the network on D_{train}^i for different values of α, from 0 to 1 with 0.1 increments. We measure the average Qini coefficient $\hat{q}$ (9) obtained on $\{D_{valid}^i\}_{i \in [1,k]}$ and we estimated a 95% confidence interval (CI) for $\hat{q}$. The selected α corresponds to the highest average as long as the 95% CI lower bound remains higher than 0. The resultant of this protocol is illustrated in Figure 3. Once α is fixed, we follow then the same process for the learning rate ξ.

6.2. Data

A data set with $n = 50,000$ customers and $p = 40$ co-variates from an insurer is at our disposal to evaluate the performance of our methods. This company was interested in designing strategies to maximize its conversion rate. An experimental acquisition campaign was implemented for 6 months, for which half of the potential clients were randomly allocated into a treatment group and the other half

into a control group. Potential clients under the treatment group were contacted using a different (and more costly) strategy. The goal of the analysis was to be able to select who among the clients portfolio should be contacted in that way going forward in order to maximize the benefits. The observed difference in sales rates between the treated group and the control shows some evidence of a positive impact of the treatment (ATE = 0.55%).

We also evaluate the performance of SMITE using synthetic data generated using two types of mechanisms. First, we leverage the real-world data to generate semi-synthetic datasets where the outcome is generated from a known non-parametric model. We first fit an uplift random forest (RF) to a random sample S of the real-world dataset. Then, we sample a bootstrap replicate of S for which we generate a synthetic outcome variable based on the RF conditional means. This method is similar to the one used in (Belbahri et al., 2019). We repeat this process several times. Each bootstrapped data size is $n = 10,000$ observations and $p = 40$ covariates. These simulations allow us to evaluate if SMITE can recover the right representation for a model generated from a tree-based model. This also makes comparisons to random forest methods fair.

The third simulation study is based on fully synthetic datasets generated from a parametric model inspired by the scenarios proposed by (Tian et al., 2014). In this case, we generate datasets of sizes $n = 20,000$ observations and $p = 100$ covariates.

6.3. Benchmarks and implementation

For random forests (RF), we use the implementation from the **R** libraries **grf** (Tibshirani et al., 2019) and **uplift** (Guelman, 2014) which were developed by the authors of the methods respectively. Implementations of these causal forests have several useful parameters that make it easy to fine tune them. In particular, **grf** provides cross-validation procedures for hyper-parameters tuning which we use to get the best possible candidates to compare our results to. We also use the **R** library **tools4uplift** (Belbahri et al., 2020) for computing the model's goodness-of-fit measures. SMITE is implemented in **Python**, so we reimplemented the Qini (9) and Kendall (10) coefficients. In total, we consider 4 different random forests to compare to SMITE: causal forest, honest causal forest (H), RF with Kullback-Leibler (KL) and RF with Euclidean distance (ED) split criteria.

6.4. Experimental results

For all models we follow the same learning process and performance metrics computations.

Table 1. Results on Canadian insurance dataset. Reported performance metrics are averaged over 30 runs.

Model	$\hat{q} \pm$ 2S.E.	$\hat{\rho} \pm$ 2S.E.
Training Set ($n = 21,000$, $p = 40$)		
Causal Forest	5.894 ± 0.056	0.96 ± 0.01
Causal Forest (H)	2.690 ± 0.162	0.97 ± 0.01
Random Forest (ED)	1.245 ± 0.041	0.91 ± 0.07
Random Forest (KL)	1.179 ± 0.040	0.91 ± 0.025
SMITE (TO)	0.285 ± 0.030	0.56 ± 0.06
SMITE (IE)	0.322 ± 0.034	0.64 ± 0.04
Validation Set ($n = 14,000$, $p = 40$)		
Causal Forest	0.004 ± 0.035	0.01 ± 0.11
Causal Forest (H)	0.035 ± 0.063	0.02 ± 0.10
Random Forest (ED)	-0.079 ± 0.04	-0.18 ± 0.08
Random Forest (KL)	-0.077 ± 0.05	-0.09 ± 0.11
SMITE (TO)	0.182 ± 0.023	0.39 ± 0.07
SMITE (IE)	$\mathbf{0.261} \pm 0.037$	$\mathbf{0.48} \pm 0.07$
Out-of-sample ($n = 15,000$, $p = 40$)		
Causal Forest	0.051 ± 0.039	0.06 ± 0.11
Causal Forest (H)	0.056 ± 0.055	0.06 ± 0.11
Random Forest (ED)	0.080 ± 0.026	0.05 ± 0.08
Random Forest (KL)	-0.002 ± 0.03	-0.06 ± 0.09
SMITE (TO)	0.167 ± 0.028	0.33 ± 0.05
SMITE (IE)	$\mathbf{0.244} \pm 0.030$	$\mathbf{0.51} \pm 0.08$

6.4.1. REAL DATA

For the real-world dataset, a sample of 30% is dedicated to out-of-sample performance measures. We do not score these customers until all models are properly fine-tuned. The remaining 70% is used for fine-tuning. We repeat the experiment 30 times. We randomly split the subset into 60% for training and 40% for validation. We select the model which performs best on the validation set. Using that model, we score the out-of-sample customers and compute the performance metrics. Results are presented in Table 1.

We observe that the random forests tend to overfit. Indeed, based on the out-of-sample results from the hold out set, the models seem to be close to random targeting. This might be due to the small (in magnitude) overall impact of the initiative. On the other hand, we observe a good adequacy of SMITE for both TO and IE. Overall, the results of SMITE (IE) appear to outperform all other methods. It is also interesting to note that the distribution of the ITE obtained by the model SMITE (IE) are in line with what was logically expected (see Figure 4: giving what appears to be a better service only slightly increases the propensity of buying.

6.4.2. SIMULATED DATA

For both simulated datasets, we do not consider any out-of-sample subset. In Monte Carlo settings, we are able to

Table 2. Results on bootstrap semi-synthetic data. Reported performance metrics are averaged over 30 runs.

Model	$\hat{q} \pm 2\text{S.E.}$	$\hat{\rho} \pm 2\text{S.E.}$
Training Set ($n = 7,000$, $p = 40$)		
Causal Forest	2.75 ± 0.06	0.94 ± 0.01
Causal Forest (H)	0.78 ± 0.09	0.66 ± 0.06
Random Forest (ED)	1.78 ± 0.05	0.90 ± 0.02
Random Forest (KL)	1.47 ± 0.04	0.84 ± 0.03
SMITE (TO)	0.53 ± 0.12	0.56 ± 0.09
SMITE (IE)	0.73 ± 0.08	0.74 ± 0.02
Validation Set ($n = 3,000$, $p = 40$)		
Causal Forest	0.07 ± 0.10	0.05 ± 0.09
Causal Forest (H)	0.12 ± 0.12	0.08 ± 0.11
Random Forest (ED)	0.02 ± 0.09	0.01 ± 0.09
Random Forest (KL)	0.02 ± 0.10	0.04 ± 0.09
SMITE (TO)	0.15 ± 0.07	0.27 ± 0.07
SMITE (IE)	$\mathbf{0.28} \pm 0.05$	$\mathbf{0.34} \pm 0.06$

Table 3. Results on synthetic data. Reported performance metrics are averaged over 30 runs.

Model	$\hat{q} \pm 2\text{S.E.}$	$\hat{\rho} \pm 2\text{S.E.}$
Training Set ($n = 14,000$, $p = 100$)		
Causal Forest	8.90 ± 0.10	0.99 ± 0.00
Causal Forest (H)	4.06 ± 0.10	0.97 ± 0.01
Random Forest (ED)	4.74 ± 0.09	0.96 ± 0.01
Random Forest (KL)	4.57 ± 0.10	0.96 ± 0.01
SMITE (TO)	3.98 ± 0.40	0.93 ± 0.02
SMITE (IE)	3.82 ± 0.12	0.93 ± 0.01
Validation Set ($n = 6,000$, $p = 100$)		
Causal Forest	2.09 ± 0.17	0.76 ± 0.06
Causal Forest (H)	2.05 ± 0.13	0.75 ± 0.04
Random Forest (ED)	1.89 ± 0.14	0.73 ± 0.04
Random Forest (KL)	1.68 ± 0.10	0.72 ± 0.05
SMITE (TO)	2.31 ± 0.22	0.83 ± 0.03
SMITE (IE)	$\mathbf{2.80} \pm 0.12$	$\mathbf{0.89} \pm 0.02$

generate as much data as we want. Therefore, we first fine-tuned the methods on a single synthetic dataset of same size as the one we used in the repeated simulations presented. We generate $n = 10,000$ observations with $p = 40$ covariates for the bootstrapped data. For the fully synthetic datasets, we report results from experiments with $n = 10,000$ observations and $p = 100$ covariates. We split into training and validation and compare the results averaged over 30 runs for the models that performed best on the validation set. Results are described in Table 2 for the bootstrap datasets and Table 3 for the synthetic datasets. For the fully synthetic data, we see that Causal Forest (H) and SMITE (TO) had similar performance on the validation set while SMITE (IE) significantly outperformed all the other methods. For the bootstrap data for which the outcomes were generated based on random forest model, it is interesting to see that SMITE (TO) and Causal Forest (H) give similar results. Moreover, once again, the SMITE (IE) model outperforms all the other models.

7. Conclusion

In this paper we present a meaningful and intuitive neural networks architecture for the problem of uplift modeling. Our framework, which we call SMITE is flexible enough to jointly optimize both the conditional mean regression and the transformed outcome regression. As a result, the SMITE framework makes it easier to predict uplift and reduce generalization error. This flexibility also allowed us to easily introduce an alternative to the MSE loss function. This alternative indirect estimation of uplift complements our unified framework. We apply our methods to both real-world and synthetic data and we compare with what is, to the best of our knowledge, state-of-the-art methods for

uplift. Our results show that the SMITE models significantly outperform random forests designed for ITE and uplift in all of our scenarios. For future work, we want to study the theoretical properties of the SMITE optimization problem as well as including more causal inference related problems such as multiple treatments and the use of observational data.
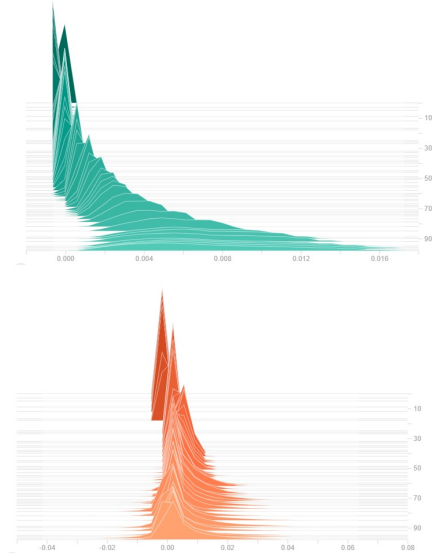

Figure 4. Examples of predicted uplift distributions on real data. We show evolution as the number of epochs increases. (top): For SMITE, individual uplifts are pushed away from zero and cover a small range within the interval $[0.000; 0.016]$ as training goes on. (bottom): A difference in conditional means neural network regression. The range covered grows as the number of epochs increases and stays more or less centered at 0. The final interval is about $[-0.02, 0.03]$. That behaviour is associated with a model that overfits .

References

Alaa, A. M., Weisz, M., and Van Der Schaar, M. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017.

Alemi, F., Erdman, H., Griva, I., and Evans, C. H. Improved statistical methods are needed to advance personalized medicine. *The open translational medicine journal*, 1:16, 2009.

Athey, S. and Imbens, G. W. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5): 1–26, 2015.

Athey, S., Tibshirani, J., Wager, S., et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Belbahri, M., Murua, A., Gandouet, O., and Partovi Nia, V. Qini-based uplift regression. *arXiv preprint arXiv:1911.12474*, 2019.

Belbahri, M., Gandouet, O., Murua, A., and Nia, V. P. *tools4uplift: Tools for Uplift Modeling*, 2020. URL https://R-Forge.R-project.org/projects/tools4uplift/. R package version 1.0.0/r54.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees*. CRC press, 1984.

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pp. 737–744, 1994.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

Guelman, L. *uplift: Uplift Modeling*, 2014. URL https://CRAN.R-project.org/package=uplift. R package version 0.3.5.

Guelman, L., Guillén, M., and Pérez-Marín, A. M. Random forests for uplift modeling: an insurance customer retention case. In *Modeling and Simulation in Engineering, Economics and Management*, pp. 123–133. Springer, 2012.

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

Hansotia, B. J. and Rukstales, B. Direct marketing for multichannel retailers: Issues, challenges and solutions. *Journal of Database Marketing and Customer Strategy Management*, 9(3):259–266, 2001.

Holland, P. W. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

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

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029, 2016.

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

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

Radcliffe, N. 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.

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

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

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

Schwab, P., Linhardt, L., and Karlen, W. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3076–3085. JMLR. org, 2017.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pp. 2503–2513, 2019.

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

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

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

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Tibshirani, J., Athey, S., and Wager, S. *grf: Generalized Random Forests*, 2019. URL https://CRAN.R-project.org/package=grf. R package version 0.10.4.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pp. 2633–2643, 2018.

Zhao, Y., Fang, X., and Simchi-Levi, D. A practically competitive and provably consistent algorithm for uplift modeling. In *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 1171–1176. IEEE, 2017.